Available online at www.sciencedirect.com
ScienceDirect
Appl. Comput. Harmon. Anal. 25 (2008) 226–239
Applied and Computational Harmonic Analysis
www.elsevier.com/locate/acha

Non-linear independent component analysis with diffusion maps

Amit Singer *, Ronald R. Coifman

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

Received 14 August 2007; revised 2 November 2007; accepted 5 November 2007

Available online 17 November 2007

Communicated by Charles K. Chui

Abstract

We introduce intrinsic, non-linearly invariant, parameterizations of empirical data, generated by a non-linear transformation of independent variables. This is achieved through anisotropic diffusion kernels on observable data manifolds that approximate a Laplacian on the inaccessible independent variable domain. The key idea is a symmetrized second-order approximation of the unknown distances in the independent variable domain, using the metric distortion induced by the Jacobian of the unknown mapping from variables to data. This distortion is estimated using local principal component analysis. Thus, the non-linear independent component analysis problem is solved whenever the generation of the data enables the estimation of the Jacobian. In particular, we obtain the non-linear independent components of stochastic Itô processes and indicate other possible applications.

1. Introduction

In data analysis and signal processing it is often the case that the observable variables are unknown functions of only a few underlying independent parameters. Those unknown functions map the low-dimensional parameter space into a subset of a high-dimensional manifold in the observable space. The challenge is to recover those independent quantities and to recover the low-dimensional intrinsic manifold.

In recent years there has been much progress in the development of new ideas and methods for dimensionality reduction, that is, parameterizing, or embedding, high-dimensional data in a low-dimensional space [1–4]. The embedding, however, is not unique. Indeed, any invertible mapping of the parameter space defines a legitimate embedding of the data. In other words, the embedded coordinates may be some complicated non-linear functions of the original parameters. On the other hand, the nature or the physics of a typical data problem usually suggests that there are unique parameters which have real physical meaning. For example, it is desired to parameterize a data set of images of a fixed object taken at different camera placements and lighting settings by those features, rather than by mixtures of features that are not as meaningful.

* Corresponding author.
E-mail addresses: amit.singer@yale.edu (A. Singer), coifman-ronald@yale.edu (R.R. Coifman).

1063-5203/$ – see front matter
doi:10.1016/j.acha.2007.11.001

This leads to the following question: how to find good intrinsic coordinates? Here we define intrinsic coordinates as a parametrization of the data which is invariant (modulo one-dimensional reparametrization in each intrinsic parameter) under non-linear changes of variables in the data.

In this paper we answer this question under different assumptions for the generation of the data. Such assumptions are needed, for otherwise there is no way to favor one embedding over the other. Common to all of our assumptions is that they enable the computation of the local Jacobian based metric distortion of the transformation that map the parameter space into the observable space. We use the local Jacobian to construct an anisotropic diffusion operator on the observable manifold which is equivalent to an isotropic diffusion operator on the parameter space. In other words, we compute the Laplacian on the parameter manifold, whereas classical diffusion maps compute the Laplacian on the observable manifold. Our reliance on the empirical metric distortion of the Jacobian assures us through the chain rule that the parametrization obtained is non-linearly invariant.

In particular, if the parameters are statistically independent then the Laplacian is separable and its eigenfunctions give the independent components. This extends our previous work [5] for solving the linear independent component analysis (ICA) problem to the non-linear regime. This is demonstrated here by solving the non-linear ICA problem of stochastic Itô processes. The existence and uniqueness of the non-linear ICA problem was studied in [6], where it was shown that there are many possible solutions to the problem. Therefore, it is necessary to impose some assumptions on either the generation of the data or to restrict the class of non-linear functions, in order to obtain a unique solution. For example, articulation from acoustics is recovered in [7] by assuming that the unobserved signals of articulator motion are band-pass. Here we assume that the data is generated by stochastic Itô processes. Non-linear ICA, other assumptions on data generation and mixing functions, and factorizing probability density functions are reviewed in [8,9].

We estimate the Jacobian from the local point cloud covariance obtained by short time stochastic simulations ('short bursts'). The local covariance estimation is performed by principal component analysis (PCA). The non-linear ICA algorithm combines many local PCAs with one global eigenmap computation. In that sense, it serves as a natural generalization of the spectral linear ICA algorithm [5], where a single global PCA is followed by an eigenmap computation. The diffusion map puts together the many locally linear patches into one global atlas.

The assumption of statistical independence can be relaxed by a weaker assumption of time scale separation for the different processes. Using the anisotropic diffusion we detect the slow manifold in multi-scaled coupled dynamical systems such as chemical reactions. This important application is discussed in much further detail in [10].

The anisotropic diffusion embedding presented here can also be viewed as a method to uniformizing both the density and geometric variability of the data. In general, the high-dimensional data points are distributed over the observable manifold according to some non-uniform density and the discrete graph Laplacian does not approximate the Laplace–Beltrami operator of the manifold. Instead, it approximates the backward Fokker–Planck operator with a potential term derived from the non-uniform density. It was shown in [11–13] that the Laplace–Beltrami operator can be recovered by a proper normalization of the kernel matrix, thus separating the density from the geometry. On the other hand, the anisotropic kernel approximates the Laplacian on the parametric manifold instead of the Laplacian on the observable manifold. We point out this difference by examining the different embeddings of a two-dimensional data set. For one-dimensional manifolds (i.e., curves) the Jacobian is merely a scalar and the anisotropic diffusion kernel is identical to the self-tuning spectral clustering algorithm [14]. We mention the problem of tomography from unknown random projections [15,16] as one possible application of Laplacians on curves. Uniformization of the data is also possible when the non-linear map is known to be conformal [6]. Conformal functions map small circles into approximately small circles, because their Jacobian is isotropic and is fully determined by the density of data points. Those examples serve as yet another interpretation to the self tuning clustering algorithm [14].

As a final application we mention inverse problems. For example, consider a sequence of tubes that model the vocal tract and produce acoustic signals [17]. The tubes are controlled by only a few parameters. It is easy to generate acoustic signals for any given set of parameters, but the inverse problem of determining the exact parameters that correspond to a given signal is perhaps more difficult. Generating local bursts of signals by locally perturbing the parameters would give the local Jacobian of the mapping. The real parameters of a newly observed signal would then be obtained by the anisotropic diffusion kernel.

The construction of the anisotropic kernel is summarized in Eq. (20) that defines the $N \times N$ weight matrix W in terms of the local Jacobian J

$$W_{ij} = \exp\left\{-\frac{\|J^{-1}(y^{(i)})(y^{(j)} - y^{(i)})\|^2 + \|J^{-1}(y^{(j)})(y^{(j)} - y^{(i)})\|^2}{4\varepsilon}\right\}, \tag{1}$$

where $y^{(i)}$ $(i = 1, \ldots, N)$ are the observable data points.

The organization of the paper is as follows. In Section 2 we formulate the non-linear ICA problem of stochastic Itô processes. In Section 3 we derive a symmetric second-order approximation for the distance metric using the covariance matrix. Section 4 gives the construction of the anisotropic diffusion kernel, based on the symmetric distance approximation. We prove that the discrete kernel-based operator converges to the Fokker–Planck operator in the limit of large sample size, and explain how it enables the uncoupling of the Itô processes. A numerical example that illustrates the performance of the algorithm is given in Section 5. Finally, in Section 6 we extend the anisotropic diffusion kernel to a more general setting of inversion from an observable data space to an inaccessible underlying parameter space.

2. Non-linear ICA of Itô processes

Consider a model in which the data points are non-linear functions of independent stochastic Itô processes. The goal is to recover the non-linear functions and the parameters of the dynamics. Specifically, the independent processes are given by

$$dx^i = a^i\big(x^i\big)\,dt + b^i\big(x^i\big)\,dw^i, \quad i = 1, \ldots, n, \tag{2}$$

where a^i and b^i are unknown drift and noise coefficients, and $\dot{w}^i$ are independent δ-correlated white noises (w^i are Brownian motions).

The n-dimensional process $x = (x^1, x^2, \ldots, x^n)$ is inaccessible. Instead, we observe its non-linear mapping $y = f(x)$ given by

$$y^j = f^j\big(x^1, x^2, \ldots, x^m\big), \quad j = 1, \ldots, m, \tag{3}$$

with $m \geqslant n$. The corresponding non-linear ICA problem is to recover the independent components x^i, the unknown non-linear functions f^j, and the unknown coefficients a^i and b^i of the dynamics. Since the functions f^j are arbitrary, we can only hope to recover each of the original independent components x^i up to some one-to-one mapping $\tilde{x}^i = g^i(x^i)$, that is, to find a scaled version of the components.

The processes y^j satisfy the stochastic dynamics given by Itô's lemma

$$dy^j = \sum_{i=1}^{n}\left(\frac{1}{2}\big(b^i\big)^2 f^j_{ii} + a^i f^j_i\right)dt + \sum_{i=1}^{n} b^i f^j_i\,dw^i, \quad j = 1, \ldots, m, \tag{4}$$

where lower indices correspond to partial derivatives, e.g., $f^j_i = \frac{\partial f^j}{\partial x^i}$.

The accessible $m \times m$ covariance matrix C of the observable processes is

$$C^{jk} \equiv \mathrm{Cov}\big(y^j, y^k\big) = \sum_{i=1}^{n}\big(b^i\big)^2 f^j_i f^k_i, \quad j, k = 1, \ldots, m, \tag{5}$$

which can also be written in terms of the $m \times n$ Jacobian matrix J

$$J_{ij} = f^i_j, \quad i = 1, \ldots, m,\ j = 1, \ldots, n, \tag{6}$$

as

$$C = J B^2 J^T, \tag{7}$$

where B is an $n \times n$ diagonal matrix with $B_{ii} = b_i$.

As f^j are any non-linear functions, we may first apply any one-dimensional scaling transformation $\tilde{x}^i = g^i(x^i)$ to each of the original inaccessible variables. In particular, we choose the scaling functions g^i in order to eliminate the b_i dependence in (2), such that the scaled variables $\tilde{x}^i$ would satisfy the SDE

$$d\tilde{x}^i = \tilde{a}^i\big(\tilde{x}^i\big)\,dt + 1\,dw^i. \tag{8}$$

Applying Itô's lemma to (2), we see that each g^i needs to satisfy

$$\frac{dg^i(x^i)}{dx^i} = \frac{1}{b^i(x^i)},$$

for (8) to hold. In other words, we may assume B to be the identity matrix

$$B = I. \tag{9}$$

This is similar to the linear ICA problem $X = AS$ where the sources S_i are assumed to have unit variance ($\mathbb{E}S_i^2 = 1$) by rescaling the coefficients of the mixing matrix A [18]. Hereafter we drop the tilde to facilitate notation. Equations (7) and (9) are combined to give

$$C = JJ^T. \tag{10}$$

Therefore, the covariance matrix C is a semi-definite positive matrix of rank n whose eigen-decomposition recovers the Jacobian matrix J up to an orthogonal transformation O

$$C = JJ^T = JOO^TJ^T = (JO)(JO)^T. \tag{11}$$

The accessible drift terms of (4) are (recall $b_i = 1$)

$$\frac{1}{2}\Delta f^j + a \cdot \nabla f^j, \quad j = 1, \ldots, n, \tag{12}$$

where $a = (a_1, a_2, \ldots, a_n)$ is the drift vector. However, we will not make use of this information, because the drift vector a is still unknown.

3. Euclidean distances and the Jacobian

The idea is to approximate the Euclidean distance between unobservable data points $x^{(j)}$ in the original space X using our knowledge of JJ^T and the data points $y^{(j)} = f(x^{(j)})$ in the observable space Y. Once such an approximation is obtained, the eigenfunctions of a suitable kernel would reveal the independent coordinates [5].

Let x, ξ be two points in the original space X and $y = f(x)$, $\eta = f(\xi)$ their mapping to the observable space Y. Let $g : Y \to X$ be the inverse mapping of $f : X \to Y$, that is, $g(f(x)) = x$ and $f(g(y)) = y$, $\forall x \in X$, $\forall y \in Y$.

Expanding the functions $x = g(y)$ in a Taylor series at the point y gives

$$\xi^i = x^i + \sum_j g^i_j(y)\big(\eta^j - y^j\big) + \frac{1}{2}\sum_{kl} g^i_{kl}(y)\big(\eta^k - y^k\big)\big(\eta^l - y^l\big) + O\big(\|\eta - y\|^3\big). \tag{13}$$

Therefore,

$$\begin{aligned}\|\xi - x\|^2 &= \sum_i \big(\xi^i - x^i\big)^2 \\ &= \sum_{ijk} g^i_j(y) g^i_k(y)\big(\eta^j - y^j\big)\big(\eta^k - y^k\big) + \frac{1}{2}\sum_{ijkl} g^i_j(y) g^i_{kl}(y)\big(\eta^j - y^j\big)\big(\eta^k - y^k\big)\big(\eta^l - y^l\big) \\ &\quad + O\big(\|\eta - y\|^4\big).\end{aligned} \tag{14}$$

A similar expansion at the point η yields

$$\begin{aligned}\|\xi - x\|^2 &= \sum_{ijk} g^i_j(\eta) g^i_k(\eta)\big(\eta^j - y^j\big)\big(\eta^k - y^k\big) - \frac{1}{2}\sum_{ijkl} g^i_j(\eta) g^i_{kl}(\eta)\big(\eta^j - y^j\big)\big(\eta^k - y^k\big)\big(\eta^l - y^l\big) \\ &\quad + O\big(\|\eta - y\|^4\big).\end{aligned} \tag{15}$$

Averaging (14) and (15) produces

$$\|\xi - x\|^2 = \frac{1}{2}(\eta - y)^T\big[\big(JJ^T\big)^{-1}(y) + \big(JJ^T\big)^{-1}(\eta)\big](\eta - y) + O\big(\|\eta - y\|^4\big), \tag{16}$$

because the Jacobian of the inverse g is the inverse of the Jacobian J, and

$$g_j^i(\eta)g_{kl}^i(\eta) - g_j^i(y)g_{kl}^i(y) = O\big(\|\eta - y\|\big).$$

Thus, (16) is a second-order symmetric approximation for the Euclidean distance in the original space X. The averaging of the first-order approximations (14) and (15) yields a second-order approximation (16), much like the trapezoidal rule in numerical integration.

Although the Jacobian is known only up to an orthogonal transformation, the computation of the approximation (16) requires only the knowledge of JJ^T, which is available through (10). Note that (16) may also be written as follows

$$\frac{\|J^{-1}(\eta)(\eta - y)\|^2 + \|J^{-1}(y)(\eta - y)\|^2}{2} = \|\xi - x\|^2 + O\big(\|\xi - x\|^4\big). \tag{17}$$

Other second-order symmetric approximations are also possible. For example, the mid-point Jacobian $J((y+\eta)/2)$ satisfies

$$\big\|J^{-1}\big((y+\eta)/2\big)(\eta - y)\big\|^2 = \|\xi - x\|^2 + O\big(\|\xi - x\|^4\big). \tag{18}$$

However, this approximation requires the knowledge of the Jacobian at the mid point for every pair of data points. A knowledge which is usually not available in practice. This makes the approximation (16) much more attractive in applications.

First-order non-symmetric approximations of the Euclidean distance are easily obtained, e.g.,

$$\big\|J^{-1}(y)(\eta - y)\big\|^2 = \|\xi - x\|^2 + O\big(\|\xi - x\|^3\big). \tag{19}$$

However, such approximations are not of great interest, because their limiting continuous operator has an additional drift term that depends on the specific mapping f. Therefore, their limiting operator would not be separable.

4. Convergence of the anisotropic kernel to the Fokker–Planck operator

Now that a suitable approximation of the Euclidean distance in X was established in (17), we construct the symmetric data affinity matrix

$$W_{ij} = \exp\left\{-\frac{\|J^{-1}(y^{(i)})(y^{(j)} - y^{(i)})\|^2 + \|J^{-1}(y^{(j)})(y^{(j)} - y^{(i)})\|^2}{4\varepsilon}\right\}. \tag{20}$$

The weight matrix W corresponds to an anisotropic kernel on the observable space Y, where the anisotropy is measured by the covariance matrix. We analyze the application of W to some function $q_Y : Y \to \mathbb{R}$ by considering the continuous limit. In the limit of number of data points $N \to \infty$ the discrete operator converges to the integral operator

$$\frac{1}{N}\sum_{j=1}^{N} W_{ij} q_Y\big(y^{(j)}\big) \to \int_Y \exp\left\{-\frac{\|J^{-1}(y^{(i)})(y - y^{(i)})\|^2 + \|J^{-1}(y)(y - y^{(i)})\|^2}{4\varepsilon}\right\} q_Y(y)\, p_Y(y)\, dy, \tag{21}$$

where $p_Y(y)$ is the density of data points in the observable data space Y. Note that the density satisfies

$$p_X(x) = p_Y(y)\big(\det JJ^T\big)^{1/2}, \tag{22}$$

where $p_X(x)$ is the density of points in the original space X. The function q_Y can be viewed as a function on X via

$$q_X(x) = q_Y\big(f(x)\big) = q_Y(y). \tag{23}$$

Substituting $y = f(x)$ in the integral (21) while employing (17) results in a Laplace type integral

$$\int_X \exp\left\{-\frac{\|x - \xi\|^2 + O(\|x - \xi\|^4)}{2\varepsilon}\right\} h(x)\, dx, \tag{24}$$

where $h(x) = q_X(x)p_X(x)$, and $\xi = x^{(i)}$. Substituting $z = (x - \xi)/\sqrt{\varepsilon}$ in (24) gives

$$\varepsilon^{n/2}\int_{\mathbb{R}^n} \exp\big\{-\|z\|^2/2 + O\big(\varepsilon\|z\|^4\big)\big\} h(\xi + \sqrt{\varepsilon}z)\, dz, \tag{25}$$

where the $O(\varepsilon\|z\|^4)$ term also include curvatures terms, because we replaced the domain of integration with the tangent space. The expansion of h and the exponent near ξ is given by

$$h(\xi+\sqrt{\varepsilon}z) = h(\xi) + \sqrt{\varepsilon}\sum_i h_i(\xi)z^i + \frac{\varepsilon}{2}\sum_{ij} h_{ij}(\xi)z^i z^j + O\big(\varepsilon^{3/2}\big), \tag{26}$$

and

$$\exp\big\{-\|z\|^2/2 + O\big(\varepsilon\|z\|^4\big)\big\} = \exp\big\{-\|z\|^2/2\big\}\bigg[1+\varepsilon\sum_{ijkl}\alpha_{ijkl}(\xi)z^i z^j z^k z^l + O\big(\varepsilon^2\big)\bigg], \tag{27}$$

where the coefficients $\alpha_{ijkl}(\xi)$ depend on the partial derivatives of g as determined by the remainder term of (14). The exact value of α_{ijkl} is not important, because it will cancel out upon normalization of the kernel. Substituting (26) and (27) in (25), and observing that odd monomials of z integrate to zero, yields

$$(2\pi\varepsilon)^{n/2}\bigg\{h(\xi) + \frac{\varepsilon}{2}\big[E(\xi)h(\xi) + \Delta h(\xi)\big] + O\big(\varepsilon^2\big)\bigg\}, \tag{28}$$

where Δ is the Laplacian

$$\Delta h = \sum_i h_{ii},$$

and $E(\xi)$ is a scalar function that depends on the mapping f through

$$E(\xi) = \frac{1}{(2\pi)^{n/2}}\sum_{ijkl}\alpha_{ijkl}(\xi)\int_{\mathbb{R}^n}\exp\big\{-\|z\|^2/2\big\}z^i z^j z^k z^l\,dz.$$

The continuous limit of the discrete graph Laplacian corresponding to data points lying on a low-dimensional Riemannian manifold also gives rise to a similar scalar potential E, which is due to the fact that Euclidean distances in the ambient space are second-order approximations of geodesic distances over the manifold [11,19]. This emphasizes once again why a second-order approximation of the Euclidean distance (17) is a necessity rather than a choice.

Normalizing the affinity matrix W to be row stochastic and subtracting the identity matrix yields the discrete normalized graph Laplacian L

$$L = D^{-1}W - I, \tag{29}$$

where D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. The row stochastic matrix $D^{-1}W$ can be viewed as a transition probability matrix of an anisotropic Markov jump process over the data points due to the local Jacobian scaling. The continuous limit of the discrete graph Laplacian L is obtained from (28)

$$\begin{aligned}\sum_j L_{ij}q_Y\big(y^{(j)}\big) &\to \frac{q_X(\xi)p_X(\xi) + \frac{\varepsilon}{2}[E(\xi)q_X(\xi)p_X(\xi) + \Delta(q_X(\xi)p_X(\xi))] + O(\varepsilon^2)}{p_X(\xi) + \frac{\varepsilon}{2}[E(\xi)p_X(\xi) + \Delta p_X(\xi)] + O(\varepsilon^2)} - q_X(\xi)\\ &= \frac{\varepsilon}{2}\big[\Delta q_X + 2\nabla(\log p_X)\cdot\nabla q_X\big] + O\big(\varepsilon^2\big).\end{aligned} \tag{30}$$

This shows that the graph Laplacian converges to a backward Fokker–Planck operator $\mathcal{L}$ on X

$$\mathcal{L}q = \Delta q - \nabla U\cdot\nabla q,$$

where the potential is proportional to the logarithm of the density $U = -2\log p_X$. Thus, the anisotropic diffusion over Y is realized as an isotropic diffusion over X with an additional drift term due to the density in X.

As the original Itô processes x^i are independent, it follows that their stationary density is multiplicative

$$p_X(x) = \prod_i p^i\big(x^i\big),$$

where $p^i(x^i)$ is the stationary density of the ith process. Therefore, the potential U is additive,

$$U(x) = \sum_i U^i\big(x^i\big),$$

and the backward Fokker–Planck operator separates into n one-dimensional operators. It follows that the eigenfunctions of the Fokker–Planck operator are in a separation of variables form, and give the original coordinates as shown in [5].

5. Numerical example

Consider the Brownian motion $(x_1, x_2) = (w_1, w_2)$[1] in the unit square $[0,1] \times [0,1]$ with normal reflection at the boundary. The stationary density is the uniform distribution $x_1, x_2 \sim U[0,1]$. Suppose that instead of (x_1, x_2) we observe their non-linear mapping

$$y_1 = x_1 + x_2^3, \qquad y_2 = x_2 - x_1^3. \tag{31}$$

This cubic transformation maps the unit square to a mushroom like domain, see Fig. 1. Although the original points are uniformly distributed in the unit square, the density of the mapped points is non-uniform and is given by

$$\det J^{-1} = \begin{vmatrix} 1 & 3x_2^2 \\ -3x_1^2 & 1 \end{vmatrix}^{-1} = \left[1 + 9(x_1 x_2)^2\right]^{-1}. \tag{32}$$

Thus, the density at the leftmost point $(0,0) = f(0,0)$ is 10 times bigger than the density at the rightmost point $(2,0) = f(1,1)$, as can be observed in Fig. 1.

The mapped coordinates (y_1, y_2) satisfy the dynamics derived by Itô's lemma (4)

$$dy_1 = 3x_2\,dt + dw_1 + 3x_2^2\,dw_2, \tag{33}$$

$$dy_2 = -3x_1\,dt - 3x_1^2\,dw_1 + dw_2, \tag{34}$$

where x_1 and x_2 are now viewed as functions of y_1 and y_2. For example, $x_1 = x_1(y_1, y_2)$ is the solution of the 9th degree polynomial

$$x_1 = y_1 - x_2^3 = y_1 - \left(y_2 + x_1^3\right)^3.$$

Therefore, the processes (33)–(34) are fully coupled. We want to find the inverse mapping that decouples them to the independent processes $x_1 = x_1(y_1, y_2)$ and $x_2 = x_2(y_1, y_2)$.

We conduct the following numerical experiment. We randomly generate $N = 2000$ points uniformly distributed in the unit square, $(x_1^{(i)}, x_2^{(i)})$, $i = 1, \ldots, N$, and non-linearly map them according to (31) to obtain $(y_1^{(i)}, y_2^{(i)})$ (Fig. 1). For every $i = 1, \ldots, N$, we run $N_c = 200$ stochastic simulations for only a short period of time $\Delta t = 0.01$, all initiating at $(x_1^{(i)}, x_2^{(i)})$. We then map the simulated trajectories to the mushroom. At time Δt we end up with a point cloud

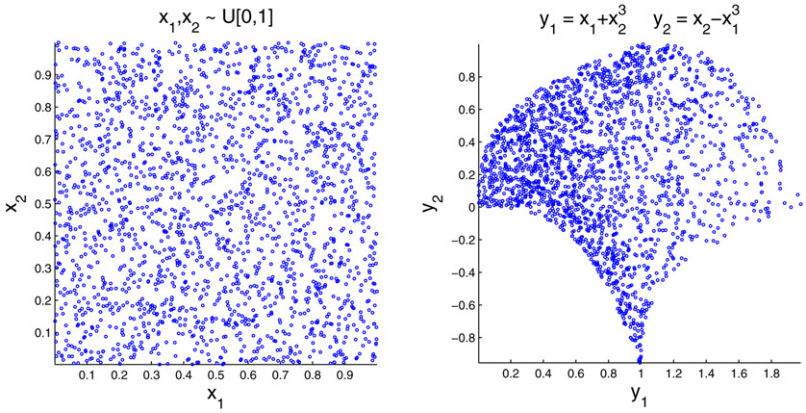

Fig. 1. $N = 2000$ uniformly sampled points in the unit square (left), and their image under the non-linear mapping (31) (right).

[1] We changed coordinates to subscript notation.

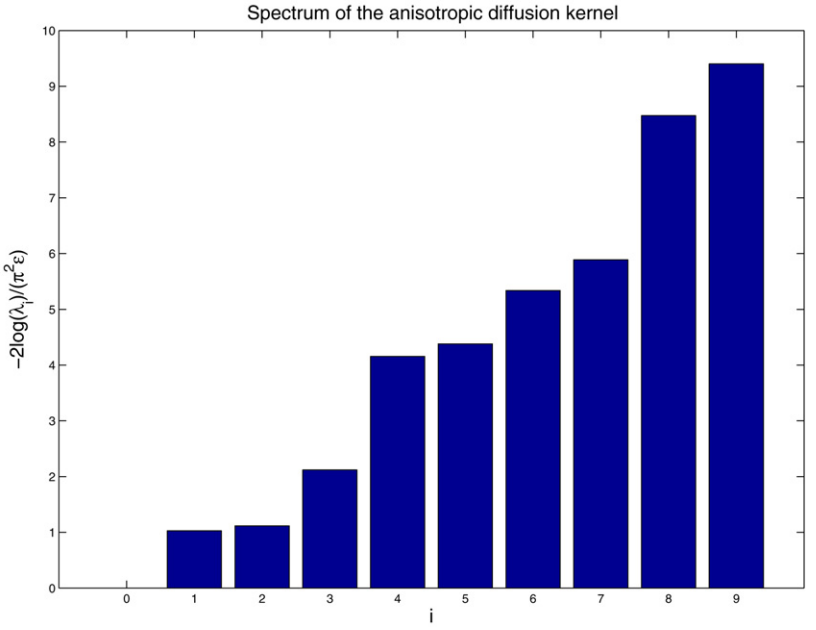

Fig. 2. The numerical spectrum of the anisotropic diffusion kernel: λ_i are the eigenvalues of the 2000×2000 row stochastic matrix $D^{-1}W$.

consisting of N_c mapped simulated points in the (y_1, y_2) plane. We calculate the sample covariance $C^{(i)}$ of the point cloud and estimate the local Jacobian $J(x_1^{(i)}, x_2^{(i)})$ using (10)

$$JJ^T\left(x_1^{(i)}, x_2^{(i)}\right) = C^{(i)}/\Delta t, \quad i = 1, \ldots, N. \tag{35}$$

We overall produce $NN_c = 4 \times 10^5$ data points by running those short time simulation bursts. The 2000 estimated 2×2 covariance matrices are inverted for the calculation of the anisotropic kernel (20) with $\varepsilon = 0.005$. The kernel is normalized to be row stochastic and the first few eigenvectors ϕ_j $(j = 0, 1, \ldots)$ of L (29) are computed.

According to (30), $D^{-1}W \approx \exp\{-\frac{\varepsilon}{2}\mathcal{L}\}$, where $\mathcal{L}$ is the Fokker–Planck operator on the parametric manifold. In our case, the parametric manifold is the unit square with a uniform density, so that $\mathcal{L}$ is merely the Laplacian of the unit square, whose eigenvalues (for the Neumann boundary conditions) are $\mu_{n,m} = \pi^2(n^2 + m^2)$ $(n, m = 0, 1, \ldots)$. Therefore, the eigenvalues λ_i of $D^{-1}W$ are expected to be $\lambda_i \approx \exp\{-\frac{\varepsilon}{2}\mu_{m,n}\}$. Fig. 2 shows $-2\log(\lambda_i)/(\pi^2\varepsilon)$ for the 2000×2000 matrix of the numerical example. The spectral lines $n^2 + m^2 = 0, 1, 1, 2, 4, 4, 5, 5, 8, 9, \ldots$ are easily identified.

The eigenfunctions of the Neumann Laplacian in the unit square are

$$\phi_{n,m}(x_1, x_2) = \cos(n\pi x_1)\cos(m\pi x_2).$$

In particular, the second eigenfunction $\phi_1 = \cos(\pi x_1)$ is only a function of x_1 (but not of x_2!) and $\phi_2 = \cos(\pi x_2)$ is only a function of x_2 (but not of x_1!).[2] We have succeeded in decoupling the processes and finding the original coordinates x_1 and x_2 in the form of ϕ_1 and ϕ_2 (up to a cosine scaling). Fig. 3 (left column) shows a color map of ϕ_1 (top) and ϕ_2 (bottom) as functions of the original data points. It is apparent that ϕ_1 is a function of x_2 while ϕ_2 is a function of x_1. Alternatively, Fig. 4 (left column) shows the embedding $(x_1, x_2) \mapsto (\phi_1, \phi_2)$. Again, it is evident that the unit square is mapped to a square. The density of points is not preserved by the embedded due to the cosine scaling that shifts more points towards the boundary.

[2] Due to multiplicity of the second eigenvalue, we find linear combinations of ϕ_1 and ϕ_2, but ϕ_3 is used to give the correct linear transformation. See [5] for more details.

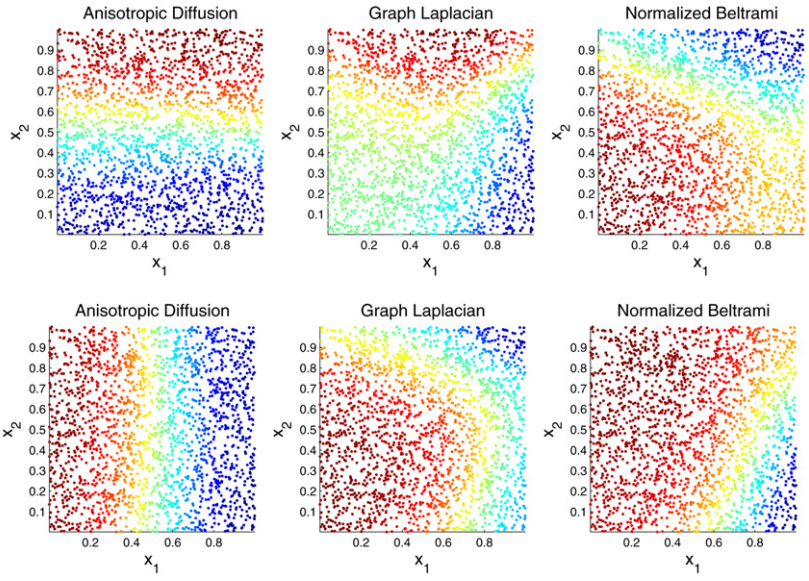

Fig. 3. Top: color map of ϕ_1. Bottom: color map of ϕ_2.

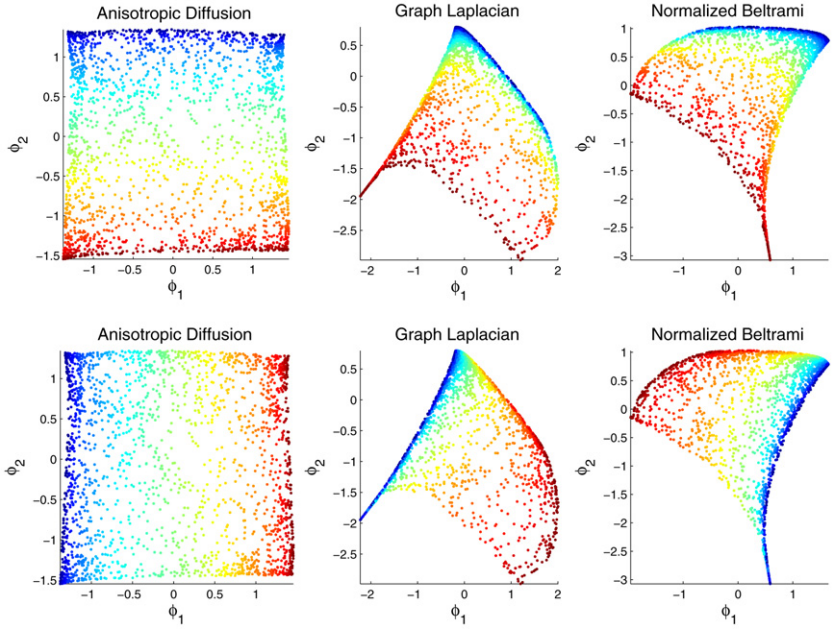

Fig. 4. Top: color map of x_1. Bottom: color map of x_2.

Figs. 3 and 4 also depict the classical isotropic diffusion map (center column) and the normalized Beltrami map (right column) [11,12]. The eigenvectors of the isotropic diffusion map approximate the eigenfunctions of the Fokker–Planck operator on the mushroom (rather than the square) with a potential derived from the non-uniform density (32). The eigenvectors of the normalized Beltrami approximate the eigenfunctions of the Laplacian on the mushroom. The figures clearly show that in both cases the computed eigenfunctions are some non-linear mixing functions of x_1 and x_2.

6. From the observable manifold to the parametric manifold

The non-linear ICA example shows that in some cases it is much favorable to approximate the Laplacian on the parametric manifold rather than on the observable manifold. In general, we want to construct the weight matrix

$$W_{i,j} = \exp\left\{-\frac{\|x^{(i)} - x^{(j)}\|^2}{2\varepsilon}\right\} \tag{36}$$

that corresponds to the parameters $x^{(i)}$ instead of the weight matrix

$$W_{i,j} = \exp\left\{-\frac{\|y^{(i)} - y^{(j)}\|^2}{2\varepsilon}\right\} \tag{37}$$

that corresponds to the observables $y^{(i)} = f(x^{(i)})$. Unfortunately, this usually cannot be done: the distances $\|x^{(i)} - x^{(j)}\|$ cannot be computed, because the parameters $x^{(i)}$ are unknown. In fact, in most applications, the whole point is to find them! But in some cases, as was demonstrated in the non-linear ICA problem of Itô processes, there is hope: instead of computing the distances $\|x^{(i)} - x^{(j)}\|$, we show a good way of approximating them, which eventually enables the computation of the Laplacian over the parametric manifold. The key is the local Jacobian, which we revisit from a different perspective.

Throughout this section we assume that we observe data points $y^{(i)}$, but we cannot access the corresponding parameters $x^{(i)}$. An additional assumption is that we can detect data points in the observable space that are the result of mapping a small ball of known size in the parameter space. This assumption enables the estimation of the local Jacobian, as explained below. We do not assume that the points in the parameter space are uniformly distributed or any other knowledge of their density. We remark that in most cases of unsupervised learning the local Jacobian cannot be estimated from the observable data, as the assumption for detecting data points that originate from small parametric balls is not realistic. However, this assumption is realistic in other cases, such as inverse problems and semi-supervised learning, as well as for the non-linear ICA problem of Itô processes, as was demonstrated in Sections 2–5.

6.1. Local principal component analysis: the Jacobian, balls and ellipsoids

The basic idea is that any smooth map between smooth manifolds $f : \mathcal{M}_X \mapsto \mathcal{M}_Y$ can be linearly approximated in a local neighborhood of any given point by its differential. The first-order Taylor expansion near x_0 reads

$$y = f(x) = y_0 + J_f(x_0)(x - x_0) + O\big(\|x - x_0\|^2\big), \tag{38}$$

where $J_f(x_0)$ is the Jacobian of f at x_0 and $y_0 = f(x_0)$.[3] This gives a first-order approximation for the distances

$$\|y - y_0\|^2 = \big\|J_f(x_0)(x - x_0)\big\|^2 + O\big(\|x - x_0\|^3\big). \tag{39}$$

Similarly, for the inverse map $f^{-1} : \mathcal{M}_Y \mapsto \mathcal{M}_X$ we have

$$\|x - x_0\|^2 = \big\|J_{f^{-1}}(y_0)(y - y_0)\big\|^2 + O\big(\|x - x_0\|^3\big), \tag{40}$$

where $J_{f^{-1}}(y_0)$ is the Jacobian of the inverse map. This means that the image of the small ball

$$\big\{x \in \mathcal{M}_X\colon\ \|x - x_0\|^2 \leqslant \delta^2\big\}$$

is the (approximately) small ellipsoid

$$\big\{y \in \mathcal{M}_Y\colon\ (y - y_0)^T J_{f^{-1}}(y_0)^T J_{f^{-1}}(y_0)(y - y_0) \leqslant \delta^2\big\}.$$

Fig. 5 illustrates that small discs are mapped to small ellipses for a specific non-linear planar map.

We view the data points $y^{(1)}, \ldots, y^{(N)} \in \mathcal{M}_Y$ as points in $\mathbb{R}^n$, but their sources $x^{(1)}, \ldots, x^{(N)} \in \mathcal{M}_X$ are not available to us. The sources

$$\big\{x^{(j)}\colon\ \big\|x^{(j)} - x^{(i)}\big\| < \delta\big\}$$

[3] The Jacobian is also known as the differential of f at x_0, which is a linear map from $T_{x_0}M_X$ (the tangent space of $\mathcal{M}_X$ at x_0) to $T_{y_0}\mathcal{M}_Y$ (the tangent space of $\mathcal{M}_Y$ at y_0). Other frequently used notations for the Jacobian are Df_{x_0}, $(f_*)_{x_0}$ and $f'(x_0)$.

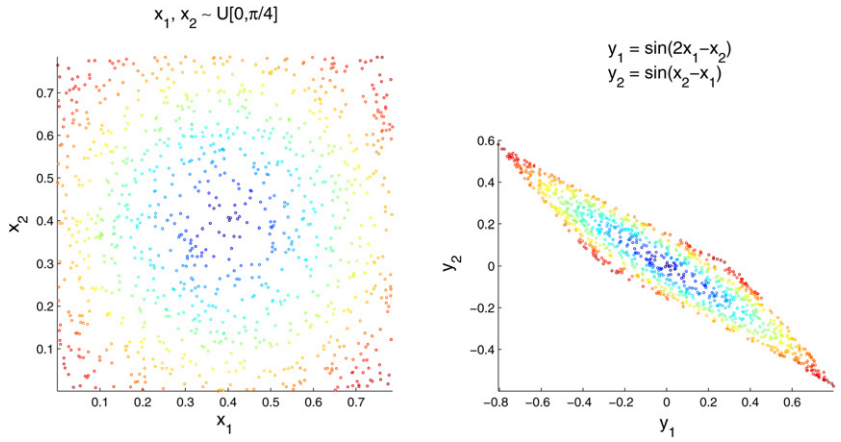

Fig. 5. Small discs are mapped into small ellipses.

inside a small ball $\mathcal{B}_{x^{(i)},\delta}$ of radius δ centered at $x^{(i)}$ are mapped by f to the data points

$$\{y^{(j)} = f(x^{(j)}):\ x^{(j)} \in \mathcal{B}_{x^{(i)},\delta}\}$$

inside the (approximately) small ellipsoid $\mathcal{E}_{y^{(i)},\delta} = f(\mathcal{B}_{x^{(i)},\delta} \cap \mathcal{M}_X)$ centered at $y^{(i)}$. Suppose that we can identify which data points $y^{(j)}$ belong to the ellipsoid $\mathcal{E}_{y^{(i)},\delta}$ and which reside outside it. For example, in Fig. 5 we identify the points inside the blue ellipse, because they were colored for us. In such a case, where the ellipsoid is known, we can translate distances over $\mathcal{M}_Y$ to those over $\mathcal{M}_X$ by contracting the major axis coordinates and expanding the minor axis coordinates.

The ellipsoid $\mathcal{E}_{y^{(i)},\delta}$ is also identified with the covariance matrix $C_{i,\delta}$ of the data points inside it

$$\begin{aligned} C_{i,\delta} &= \mathbb{E}\big[(y - y^{(i)})(y - y^{(i)})^T\big] \approx \mathbb{E}\big[J_f(x^{(i)})(x - x^{(i)})(x - x^{(i)})^T J_f^T(x^{(i)})\big] \\ &= J_f(x^{(i)})\mathbb{E}\big[(x - x^{(i)})(x - x^{(i)})^T\big] J_f^T(x^{(i)}), \end{aligned} \tag{41}$$

where the approximation is due to the linear approximation (38) and we used the linearity of expectation. For sufficiently small δ, the random variable x is uniformly distributed in a d-dimensional ball of radius δ in the tangent space $T_{x^{(i)}}\mathcal{M}_X$ ($d = \dim \mathcal{M}_X = \dim \mathcal{M}_Y$). The symmetry of the ball implies that the covariance is proportional to the $d \times d$ identity matrix I

$$\mathbb{E}\big[(x - x^{(i)})(x - x^{(i)})^T\big] = c_{d,\delta} I, \tag{42}$$

where the constant $c_{d,\delta}$ is directly calculated by integration[4]

$$c_{d,\delta} = \frac{\int_{\mathcal{B}_\delta} x_1^2\, dx}{\mathrm{vol}(\mathcal{B}_\delta)} = \frac{1}{d}\frac{\int_0^\delta r^2 r^{d-1}\omega_d\, dr}{\int_0^\delta r^{d-1}\omega_d\, dr} = \frac{\delta^2}{d+2}. \tag{43}$$

Plugging (42)–(43) into (41) shows that we can approximate $J_f(x^{(i)})J_f^T(x^{(i)})$ with the covariance matrix

$$J_f(x^{(i)})J_f^T(x^{(i)}) = \frac{d+2}{\delta^2} C_{i,\delta} + O(\delta). \tag{44}$$

For that reason, we also refer to the matrix $J_f(x^{(i)})J_f^T(x^{(i)})$ as the covariance matrix.

Even though the covariance matrix $C_{i,\delta}$ is an $n \times n$ semi-positive matrix, its rank should only be d, because the data points approximately lie on the d-dimensional tangent space $T_{y^{(i)}}\mathcal{M}_Y$. Its largest d eigenvalues $\mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant \mu_d$ are the squares of the semi-principal axes of the ellipsoid, while the corresponding eigenvectors $v_1, v_2, \ldots, v_d$ are the principal components. The remaining $n - d$ eigenvalues should be close to 0 (they do not completely vanish due

[4] Notation: $x = (x_1, x_2, \ldots, x_d)$, $r^2 = x_1^2 + x_2^2 + \cdots + x_d^2$ and ω_d is the surface area of the d-dimensional unit sphere $S^{d-1} \subset \mathbb{R}^d$.

to deviations from linearity and noisy data measurements). Therefore, the spectral decomposition of the covariance matrix is

$$C_{i,\delta} = \sum_{m=1}^{n} \mu_m v_m v_m^T \approx \sum_{m=1}^{d} \mu_m v_m v_m^T. \tag{45}$$

The procedure we have just described is no other than local principal component analysis (PCA) of the ellipsoid point cloud.

The inverse function theorem implies that $J_{f^{-1}}^T(y^{(i)}) J_{f^{-1}}(y^{(i)}) = (J_f(x^{(i)}) J_f^T(x^{(i)}))^{-1}$, and we can now approximate local distances on $\mathcal{M}_X$ near $x^{(i)}$ using our estimation for the covariance matrix

$$\|x - x^{(i)}\|^2 \approx (y - y^{(i)})^T \big(J_f(x^{(i)}) J_f^T(x^{(i)})\big)^{-1} (y - y^{(i)}) \approx \frac{\delta^2}{d+2} (y - y^{(i)})^T C_{i,\delta}^{-1} (y - y^{(i)}). \tag{46}$$

We need to clarify what do we exactly mean by $C_{i,\delta}^{-1}$ in (46), because $C_{i,\delta}$ is a matrix of rank $d \ll n$. By $C_{i,\delta}^{-1}$ we actually mean the pseudo-inverse $C_{i,\delta}^{\dagger}$ of $C_{i,\delta}$ on the d-dimensional subspace of principal components

$$C_{i,\delta}^{\dagger} = \sum_{m=1}^{d} \mu_m^{-1} v_m v_m^T, \tag{47}$$

so (46) reads

$$\|x - x^{(i)}\|^2 \approx \frac{\delta^2}{d+2} (y - y^{(i)})^T C_{i,\delta}^{\dagger} (y - y^{(i)}). \tag{48}$$

The approximation (48) is valid only for points y in the local neighborhood of $y^{(i)}$, and is interpreted as follows. Ignoring the remaining $n - d$ components is equivalent to projecting the point y onto the d-dimensional tangent space of principal components, and distances are measured in that tangent space. The principal coordinates $(y - y^{(i)})^T v_m$ $(m = 1, \ldots, d)$ are scaled with their corresponding semi-principal axes $\sqrt{\mu_m}$, thus stretching and contracting the ellipsoid in all d directions so it becomes a ball.

The Euclidean chordal distance $\|x - x^{(i)}\|$ between points in the ambient space is a second-order approximation of the geodesic distance $d_g(x, x^{(i)})$ over the low-dimensional intrinsic manifold $\mathcal{M}_X$

$$d_g(x, x^{(i)}) = \|x - x^{(i)}\| + O(\|x - x^{(i)}\|^3).$$

Therefore, in order to approximate the Laplace–Beltrami operator over $\mathcal{M}_X$, a second-order approximation of the distance must be used, for otherwise a different limiting differential operator is recovered. However, the linear approximation (40) is only a first-order approximation. We construct a second-order approximation, without including second-order derivatives, by using symmetrization

$$\|x - x^{(i)}\|^2 = \frac{1}{2} \|J_{f^{-1}}(y^{(i)})(y - y^{(i)})\|^2 + \frac{1}{2} \|J_{f^{-1}}(y)(y - y^{(i)})\|^2 + O(\|x - x^{(i)}\|^4) \tag{49}$$

which in terms of the covariance matrices is rewritten as

$$\|x^{(j)} - x^{(i)}\|^2 = \frac{1}{2} \frac{\delta^2}{d+2} (y^{(j)} - y^{(i)})^T [C_{i,\delta}^{\dagger} + C_{j,\delta}^{\dagger}](y^{(j)} - y^{(i)}) + O(\|x^{(j)} - x^{(i)}\|^4).$$

We construct the parametric graph Laplacian using the $N \times N$ weight matrix W,

$$W_{ij} = \exp\left\{ -\frac{\|J_{f^{-1}}(y^{(i)})(y^{(j)} - y^{(i)})\|^2 + \|J_{f^{-1}}(y^{(j)})(y^{(j)} - y^{(i)})\|^2}{4\varepsilon} \right\}, \tag{50}$$

or equivalently, using the covariance matrices

$$W_{ij} = \exp\left\{ -\frac{\delta^2}{d+2} \frac{(y^{(j)} - y^{(i)})^T [C_{i,\delta}^{\dagger} + C_{j,\delta}^{\dagger}](y^{(j)} - y^{(i)})}{4\varepsilon} \right\}. \tag{51}$$

Even though the data points $y^{(i)}$ are given on $\mathcal{M}_Y$, the parametric graph Laplacian $L = D^{-1}W - I$ approximates the Laplace–Beltrami on $\mathcal{M}_X$ rather than that on $\mathcal{M}_Y$, as required.

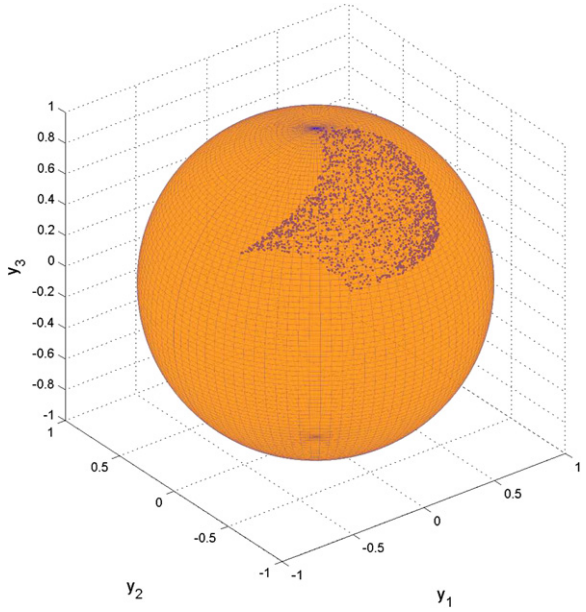

Fig. 6. Points (y_1, y_2, y_3) on the unit sphere are obtained by mapping uniformly sampled points in the unit square by (52).

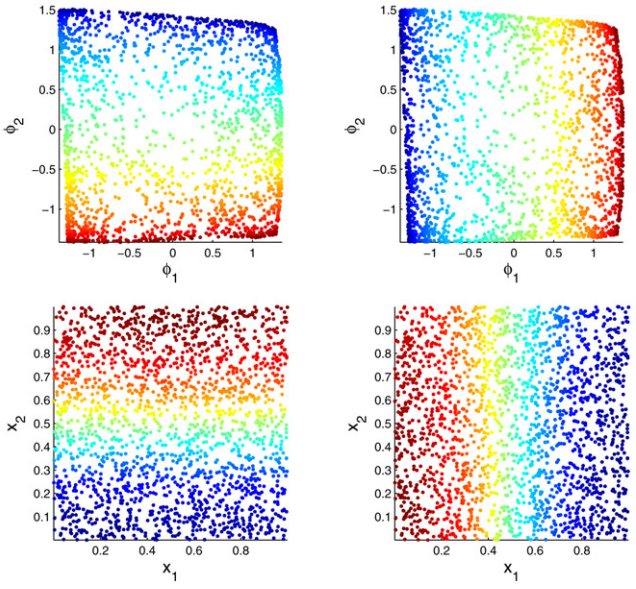

Fig. 7. Top: color map of x_1 (left) and x_2 (right) as a function of the embedded coordinates ϕ_1 and ϕ_2. Bottom: color map of ϕ_1 (left) and ϕ_2 (right) as a function of the original sampled points in the unit square.

6.2. *Numerical example*

We illustrate uniformization through the following example. We map $N = 2000$ uniformly sampled points (x_1, x_2) in the unit square to a mushroom-like manifold on the 2-sphere in $\mathbb{R}^3$ $(y_1^2 + y_2^2 + y_3^2 = 1)$

$$y_1 = \frac{x_1 + x_2^3}{\sqrt{(x_1 + x_2^3)^2 + (x_2 - x_1^3)^2 + 1}},$$

$$y_2 = \frac{x_2 - x_1^3}{\sqrt{(x_1 + x_2^3)^2 + (x_2 - x_1^3)^2 + 1}},$$
$$y_3 = \frac{1}{\sqrt{(x_1 + x_2^3)^2 + (x_2 - x_1^3)^2 + 1}} \tag{52}$$

(see Fig. 6). For every point we compute the 3×3 covariance matrix C_i of a local burst of $N_c = 1000$ simulated points with $\Delta t = 0.001$ (under the assumption of the standard dynamics $dx_1 = dw_1$ and $dx_2 = dw_2$, as in Section 5). We compute the pseudo-inverse $C_i^\dagger$ by taking the two principal components (the third singular value of the covariance matrix is much smaller because the manifold is two-dimensional). The eigenvectors of the normalized anisotropic Laplacian (51) with $\varepsilon = 0.005$ are computed. The resulted embedding is shown in Fig. 7 where it is clear that the computed coordinates ϕ_1, ϕ_2 recover the original coordinates x_1, x_2.

Acknowledgments

We would like to thank Yosi Keller, Yoel Shkolnisky, Mauro Maggioni, Ioannis Kevrekidis and Peter Jones for valuable discussions. Research supported by NGA NURI 2006.

References

[1] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[2] D.L. Donoho, C. Grimes, Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, Proc. Nat. Acad. Sci. 100 (10) (2003) 5591–5596.
[3] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[4] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30.
[5] A. Singer, Spectral independent component analysis, Appl. Comput. Harmon. Anal. 21 (1) (2006) 135–144.
[6] A. Hyvärinen, P. Pajunen, Non linear independent component analysis: Existence and uniqueness results, Neural Networks 12 (3) (1999) 429–439.
[7] J. Hogden, P. Rubin, E. McDermott, S. Katagiri, L. Goldstein, Inverting mappings from smooth paths through R^n to paths through R^m: A technique applied to recovering articulation from acoustics, Speech Commun. 49 (2007) 361–383.
[8] C. Jutten, J. Karhunen, Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures, Int. J. Neural Systems 14 (5) (2004) 267–292.
[9] L. Almeida, Nonlinear Source Separation. Synthesis Lectures on Signal Processing, Morgan & Claypool Publishers, 2005, 114 pp.
[10] A. Singer, R. Erban, I. Kevrekidis, R.R. Coifman, Detecting the slow manifold by anisotropic diffusion maps, preprint.
[11] S. Lafon, Diffusion maps and geometric harmonics, Ph.D. dissertation, Yale University, 2004.
[12] B. Nadler, S. Lafon, R.R. Coifman, I. Kevrekidis, Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA, 2006, pp. 955–962.
[13] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmon. Anal. 21 (1) (2006) 113–127.
[14] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems 17 (NIPS'04), MIT Press, Cambridge, MA, 2005, pp. 1601–1608.
[15] S. Basu, Y. Bresler, Feasibility of tomography with unknown view angles, IEEE Trans. Image Process. 9 (6) (2000) 1107–1122.
[16] R.R. Coifman, Y. Shkolnisky, F.J. Sigworth, A. Singer, Graph Laplacian tomography from unknown random projections, IEEE Trans. Image Process., in press.
[17] A. Jansen, P. Niyogi, Intrinsic Fourier analysis on the manifold of speech sounds, in: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2006), 2006.
[18] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York.
[19] 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-Verlag, Berlin, 2005, pp. 486–500.