The e-PCA and m-PCA: Dimension Reduction of Parameters by Information Geometry

Shotaro Akaho
Neuroscience Research Institute
The National Institute of Advanced Industrial Science and Technology
Tsukuba, Japan 3058568
E-mail: s.akaho@aist.go.jp

***Abstract*— We propose a method for extracting a low dimensional structure from a set of parameters of probability distributions. By an information geometrical interpretation, we show that there exist two kinds of possible flat structures for fitting (e-PCA and m-PCA). We derive alternating procedures to find the low dimensional structures. Each alternating procedure can be written in a nonlinear equation. It can be solved analytically in some special cases. Otherwise, we need to apply gradient type methods that we also derive. Since the overall algorithm may converge to a local optimum, we propose a method to find a good initial solution by using metric information.**

I. Introduction

Principal component analysis (PCA) is widely used for extracting a low dimensional linear subspace that is underlying in a set of spatial data. PCA minimizes the reconstruction error defined by the sum of the squared distance between a data point and its projection onto the subspace.

In this paper, we consider the case in which a data point is represented as a parameter of a probability distribution. For example, suppose we have a population and each individual is represented as a stochastic model. We would like to extract features that are common in the population, when the parameter of each individual has been already identified.

Let us consider a simple example. Suppose there are five individuals 'A', 'B', . . . , 'E', each of which is characterized by a normal distribution $\mathcal{N}[\mu, \sigma^2]$. In Fig. 1, they are represented by points on the (σ^2, μ) space. What we would like to do here is to find a lower dimensional structure of the population. In this example, we can consider zero or one dimensional subspace: The 0-dim space is a point representing a center of the population, and the 1-dim space is a line (or curve) fitting the population.

Another example can be found in the field of bioinformatics where a structure is to be extracted from a set of kernel matrices, each of which represents a similarity of genes or proteins[16]. The kernel matrix can be regarded as a covariance parameter of a normal distribution with zero mean.

In those examples, PCA does not always give a good result, mainly because the squared distance is not an appropriate metric in the space of distributions. Furthermore, the projection point may exceed the domain of the parameter space.

Let us consider the example of Fig. 1 again. The solid line represents a one dimensional subspace obtained by PCA. In this space, only the right half-plane $\sigma^2 > 0$ is meaningful.

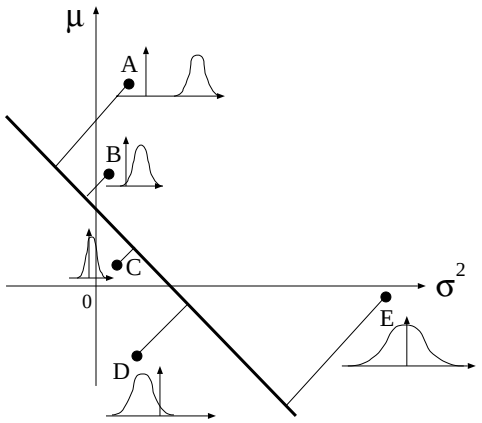


Fig. 1. Schematic figure of PCA for the parameters of a distribution

However, the projection points of 'A' and 'B' are in the negative variance region. Even if the projection point is luckily in the positive region, we do not know whether the vertical projection is reasonable especially for the point that is distant from the subspace.

To cope with such problems, we apply the information geometrical framework[2], [4] that gives natural geometry of probability distributions. The information geometry has been successfully applied to analyze and interpret many kinds of learning methods, for instance, the EM algorithm[3], mean field approximation[14], turbo coding[9] and boosting[11].

We assume the data points belong to an exponential family that is a basic space in the information geometry. Then the task is to find a linear subspace of the exponential family that fits to data points. An extension of PCA to the exponential family was studied in the pioneer work of Collins et al[7]. Our framework is more general as shown in sec. VIII. Further, our information geometrical interpretation can provide with understandings as follows: Firstly, we show the existence of another set of linear subspace, so we have two kinds of subspaces unlike the original PCA. Secondly, there are some cases in which we can calculate a closed form solution in each step. Thirdly, we can obtain a good initial solution by using metric information.

II. Dualistic geometry of the exponential family

To obtain a flat submanifold fitting in the space of an exponential family, we need to extend the notion of flatness and projection. For that purpose, we briefly summarize the information geometry of the exponential family.

The exponential family $\mathcal{S}$ is a set of distributions of random variable $\boldsymbol{x}$ with a density,

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\{\sum_{i=1}^{d} \theta_i F_i(\boldsymbol{x}) + C(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\}, \qquad (1)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ is called the natural parameter. A set of parameters of the exponential family can be regarded as a Riemannian space specified by the local coordinate system $\boldsymbol{\theta}$. We can take another coordinate system $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)^\top$ that is dual to $\boldsymbol{\theta}$ and defined by $\eta_i = \mathbb{E}_{\boldsymbol{\theta}}[F_i(\boldsymbol{x})]$, where $\mathbb{E}_{\boldsymbol{\theta}}$ denotes the expectation with respect to the distribution $p(\boldsymbol{x};\boldsymbol{\theta})$. Thus $\boldsymbol{\eta}$ is called the expectation parameter. $\boldsymbol{\theta}$ coordinate and $\boldsymbol{\eta}$ coordinate are related one-to-one, hence there is coordinate transformation maps $\boldsymbol{\theta}(\boldsymbol{\eta})$ and $\boldsymbol{\eta}(\boldsymbol{\theta})$. Note that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ cannot take all values in $\mathbb{R}^d$ in general.

The geodesic is a generalized notion of 'straight curve' in the Riemannian space. We can define dually coupled geodesics for the space of probability distributions: an exponential geodesic (e-geodesic) and a mixture geodesic (m-geodesic). The e-geodesic is a linear curve in the $\boldsymbol{\theta}$ coordinate, i.e., the e-geodesic connecting $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be written as $\boldsymbol{\theta}(t) = t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2$, where $t \in T \subset \mathbb{R}$, and T is an interval such that $\boldsymbol{\theta}(t)$ is defined. On the other hand, the m-geodesic is a linear curve in the $\boldsymbol{\eta}$ coordinate, and the m-geodesic connecting $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ can be written as $\boldsymbol{\eta}(t) = t\boldsymbol{\eta}_1 + (1-t)\boldsymbol{\eta}_2$.

In our problem, we consider a 'flat' submanifold, which can also be defined in dual ways. The e-flat submanifold is defined as a submanifold $\mathcal{M}$ in which the e-geodesic connecting any two points of $\mathcal{M}$ is included in $\mathcal{M}$ again. The m-flat submanifold is defined similarly in terms of the m-geodesic. The whole space of the exponential family is both e-flat and m-flat, hence it is called dually flat. However, the e-flat submanifold is not m-flat in general and vice versa. The conditions that they coincide are partly understood for the space of positive semidefinite matrices[12].

The notion of projection is also extended in dual ways. The e-projection from a point $\boldsymbol{\theta} \in \mathcal{S}$ to a submanifold $\mathcal{M}$ is the e-geodesic connecting $\boldsymbol{\theta}$ to $\hat{\boldsymbol{\theta}} \in \mathcal{M}$ which is orthogonal at $\hat{\boldsymbol{\theta}}$ with respect to the Riemannian metric $g_{jk}(\hat{\boldsymbol{\theta}})$ for $\boldsymbol{\theta}$ coordinate (e-metric), which is equal to the Fisher information,

$$g_{jk}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_j}\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_k}\right] \qquad (2)$$

The m-projection from a point $\boldsymbol{\eta} \in \mathcal{S}$ to a submanifold $\mathcal{M}$ is defined in a similar way with respect to the m-geodesic and the Riemannian metric $g^{jk}(\boldsymbol{\eta})$ for $\boldsymbol{\eta}$ coordinate (m-metric), which is given by the inverse matrix of $g_{jk}(\boldsymbol{\theta}(\boldsymbol{\eta}))$.

The e- or m-projection is obtained by the following proposition.

Proposition 1 (Amari[2]): The e-projection from a point $\boldsymbol{\theta} \in \mathcal{S}$ to a submanifold $\mathcal{M}$ is given by finding a point $\hat{\boldsymbol{\theta}} \in \mathcal{M}$ that is a critical point of the e-divergence defined by the Kullback-Leibler divergence

$$k(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}_{\hat{\boldsymbol{\theta}}}[\log p(\boldsymbol{x};\hat{\boldsymbol{\theta}}) - \log p(\boldsymbol{x};\boldsymbol{\theta})]. \qquad (3)$$

In particular, if $\mathcal{M}$ is m-flat, the e-projection is unique and it minimizes the e-divergence.

Similarly, the m-projection from a point $\boldsymbol{\eta} \in \mathcal{S}$ to a submanifold $\mathcal{M}$ is given by finding a point $\hat{\boldsymbol{\eta}} \in \mathcal{M}$ that is a critical point of the m-divergence defined by $k(\boldsymbol{\theta}(\boldsymbol{\eta}), \boldsymbol{\theta}(\hat{\boldsymbol{\eta}}))$. If $\mathcal{M}$ is e-flat, the m-projection is unique and it minimizes the m-divergence. (Note that e-divergence is not equal to m-divergence because of asymmetry of Kullback-Leibler divergence.) ■

This proposition gives a relation between the divergence and the projection, and also it guarantees that there exists a projection from any point of $\mathcal{S}$, even when $\mathcal{S}$ is not defined for the whole $\mathbb{R}^d$ space, while it is not true for the conventional Euclidean projection.

III. e-PCA and m-PCA

The information geometrical consideration suggests that there are two possible ways of a flat submanifold, i.e. e-flat and m-flat. Let e-PCA denote the e-flat submanifold fitting, and m-PCA denote the m-flat submanifold fitting.

Here we describe only the e-PCA by space limitation. The m-PCA can be derived just in a similar (dual) way by just exchanging m- and e- in the following description.

The h dimensional e-flat submanifold $\mathcal{M} \subset \mathcal{S}$ can be expressed by the set of points,

$$\boldsymbol{\theta}(\boldsymbol{w}; U) = \sum_{j=1}^{h} w_j \boldsymbol{u}_j + \boldsymbol{u}_0, \qquad (4)$$

where $U = [\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_h] \in \mathbb{R}^{d\times h}$ is a set of basis vectors and $\boldsymbol{w} = (w_1, \ldots, w_h)^\top \in \mathbb{R}^h$ is a local coordinate of $\mathcal{M}$. Note that the constant vector $\boldsymbol{u}_0$ plays a more significant role than the original PCA, because the projection points move nonlinearly by changing $\boldsymbol{u}_0$.

It is easy to show that $\mathcal{S}$ is convex with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ coordinate, hence $\mathcal{M}$ is also convex with respect to $\boldsymbol{w}$.

In the case of e-PCA, it is natural to take an m-projection since the Proposition 1 holds. Now suppose we have n sample points $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)} \in \mathcal{S}$ expressed by $\boldsymbol{\theta}$ coordinate. Then the m-projection of each sample $\boldsymbol{\theta}^{(i)}$ onto the submanifold $\mathcal{M}$ is given by $\hat{\boldsymbol{\theta}}^{(i)} = \boldsymbol{\theta}(\hat{\boldsymbol{w}}^{(i)}; U)$, where $\hat{\boldsymbol{w}}^{(i)}$ is the $\boldsymbol{w}^{(i)}$ that minimizes the m-divergence $k(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}(\boldsymbol{w}^{(i)}; U))$.

Next, we need to define a cost function to optimize the submanifold bases U. It is convenient to take the sum of the m-divergence for all samples as a cost function, because the overall optimization is obtained by minimizing

$$L(U, W) = \sum_{i=1}^{n} k(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}(\boldsymbol{w}^{(i)}; U)) \qquad (5)$$

simultaneously with respect to $U \in \mathcal{U}$ and $W \in \mathcal{M}^n$, where $W = [\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(n)}]$, and $\mathcal{U}$ is the domain of U such that $\mathcal{M} \neq \emptyset$.

IV. Alternating procedures

Since it is difficult to optimize U and W simultaneously, we apply alternating procedures where one parameter is optimized with fixing the other parameter and vice versa.

In this section, we derive a solution for each alternating procedure with respect to a mixed coordinate introduced by Amari[5]. We focus only on the e-PCA case. The m-flat case can be obtained straightforwardly.

First, let us consider the optimization of W with fixing U. In this case, each $\boldsymbol{w}^{(i)}$ can be optimized independently. We consider the bases $U^* = (\boldsymbol{u}_{h+1}, \ldots, \boldsymbol{u}_n)$ of the complement space of the column space of U, where we assume U has a full rank. Then $\boldsymbol{\theta}$ can be written in the form

$$\boldsymbol{\theta} = (U\ U^*) \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{w}^* \end{pmatrix} + \boldsymbol{u}_0, \tag{6}$$

where $\boldsymbol{w}^*$ is a coefficient vector for U^*. On the submanifold $\mathcal{M}$ spanned by U, $\boldsymbol{w}^*$ is equal to $\mathbf{0}$. If U is full-rank, $(U\ U^*)$ is invertible and

$$\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{w}^* \end{pmatrix} = \begin{pmatrix} V \\ V^* \end{pmatrix} (\boldsymbol{\theta} - \boldsymbol{u}_0), \tag{7}$$

where V and V^* are a partition of $(U\ U^*)^{-1}$.

Since the new coordinate system $(\boldsymbol{w}, \boldsymbol{w}^*)$ is a linear function of $\boldsymbol{\theta}$, $(\boldsymbol{w}, \boldsymbol{w}^*)$ is also a natural coordinate of $p(\boldsymbol{x})$. We can easily show that the corresponding dual parameter of $(\boldsymbol{w}, \boldsymbol{w}^*)$ is given by

$$\begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{v}^* \end{pmatrix} = \begin{pmatrix} U^\top \\ U^{*\top} \end{pmatrix} \boldsymbol{\eta},$$

where $\boldsymbol{v}$ is dual to $\boldsymbol{w}$ and $\boldsymbol{v}^*$ is dual to $\boldsymbol{w}^*$.

Let us introduce the mixed coordinate system $[\boldsymbol{v}; \boldsymbol{w}^*]$.

Proposition 2 (Amari[5]): Any point in $\mathcal{S}$ is uniquely represented by the mixed coordinate $\boldsymbol{\beta} = [\boldsymbol{v}; \boldsymbol{w}^*]$. The m-projection of the point onto the submanifold $\mathcal{M}$ spanned by U is expressed by the mixed coordinate as

$$\hat{\boldsymbol{\beta}} = [\boldsymbol{v}; \mathbf{0}], \tag{8}$$

i.e., the dual coordinate part $\boldsymbol{v}$ is the same as $\boldsymbol{\beta}$ and the natural coordinate part $\boldsymbol{w}^*$ is $\mathbf{0}$. ■

From this proposition, the m-projection of $\boldsymbol{\theta}^{(i)}$ onto $\mathcal{M}$ spanned by U satisfies the equations

$$V^*(\hat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{u}_0) = \mathbf{0}, \qquad U^\top \hat{\boldsymbol{\eta}}^{(i)} = \boldsymbol{\eta}^{(i)}, \tag{9}$$

where $\boldsymbol{\eta}^{(i)}$ is $\boldsymbol{\eta}(\boldsymbol{\theta}^{(i)})$ and $\hat{\boldsymbol{\eta}}^{(i)} = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}^{(i)})$.

Similarly, the case of the optimization of U with fixing W can be formalized. However, in this case, all columns of U are coupled, and we have to consider the product space $\mathcal{U} \subset \mathcal{S}^n$ of the whole samples.

A similar equation to (6) is given by

$$\begin{pmatrix} \boldsymbol{\theta}^{(1)} \\ \vdots \\ \boldsymbol{\theta}^{(n)} \end{pmatrix} = (A\ A^*) \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_h \\ \boldsymbol{u}_0 \\ \boldsymbol{u}^* \end{pmatrix}, \tag{10}$$

where A is a matrix which depends only on W, and A^* spans the complement space of the column space of A. We assume A has a full rank, which means all $\boldsymbol{w}^{(i)}$ are located in general positions without any collinearity.

Equations (6) and (10) are not always analytically solvable in general. In such a case, we can use a gradient descent algorithm or Newton-type method described in the next section.

V. Iterative algorithm

If a closed form solution is not available, we need to use an iterative method.

First let us derive a simple gradient type algorithm for the e-PCA. Suppose we have a current candidate $\tilde{\boldsymbol{\theta}}^{(i)} = \boldsymbol{\theta}(\tilde{\boldsymbol{w}}^{(i)}; U)$ of the m-projection point $\hat{\boldsymbol{\theta}}^{(i)}$. The m-divergence is expressed explicitly as

$$k(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}(\tilde{\boldsymbol{w}}^{(i)}; U)) = (\boldsymbol{\theta}^{(i)} - \tilde{\boldsymbol{\theta}}^{(i)})^\top \boldsymbol{\eta}^{(i)} - \psi(\boldsymbol{\theta}^{(i)}) + \psi(\tilde{\boldsymbol{\theta}}^{(i)}). \tag{11}$$

Using a useful relation $\boldsymbol{\eta} = \partial\psi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, we easily get the derivative of $L(U, W)$ by

$$\frac{\partial L(U,W)}{\partial w_j^{(i)}} = \boldsymbol{u}_j^\top (\tilde{\boldsymbol{\eta}}^{(i)} - \boldsymbol{\eta}^{(i)}) \tag{12}$$

$$\frac{\partial L(U,W)}{\partial \boldsymbol{u}_j} = \sum_{i=1}^{n} w_j^{(i)} (\tilde{\boldsymbol{\eta}}^{(i)} - \boldsymbol{\eta}^{(i)}) \tag{13}$$

where $\tilde{\boldsymbol{\eta}}^{(i)}$ denotes $\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}}^{(i)})$. The derivative by $\boldsymbol{u}_0$ is given by $n(\tilde{\boldsymbol{\eta}}^{(i)} - \boldsymbol{\eta}^{(i)})$. We can see clear duality between the derivatives by U and W. This is related to the gradient flow of the alternating procedures that appears in the EM algorithm[6], [3]. In a simple gradient algorithm, $w_j^{(i)}$ and $\boldsymbol{u}_j$ are modified slightly to the opposite direction given by (12) and (13).

In the next step, we consider a higher order algorithm. Here we describe only the case of optimizing W with fixing U. The case of optimizing U can be derived in a similar way.

Suppose the current candidate $\tilde{\boldsymbol{\theta}}^{(i)}$ is close to the optimal value $\hat{\boldsymbol{\theta}}^{(i)}$, i.e., $\hat{\boldsymbol{\theta}}^{(i)} = \tilde{\boldsymbol{\theta}}^{(i)} + d\boldsymbol{\theta}$ for small $d\boldsymbol{\theta}$. Then (10) can be written as

$$V^*(\tilde{\boldsymbol{\theta}}^{(i)} + d\boldsymbol{\theta} - \boldsymbol{u}_0) = \mathbf{0}, \qquad U^\top(\tilde{\boldsymbol{\eta}}^{(i)} + d\boldsymbol{\eta}) = \boldsymbol{\eta}^{(i)}, \tag{14}$$

where $d\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}^{(i)} - \tilde{\boldsymbol{\eta}}^{(i)}$. By using the relation $\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta} = G(\boldsymbol{\theta})$ where $G(\boldsymbol{\theta})$ is a matrix of the metric (2), we have

$$d\boldsymbol{\eta} = G(\tilde{\boldsymbol{\theta}}^{(i)}) d\boldsymbol{\theta} + o(|d\boldsymbol{\theta}|). \tag{15}$$

Neglecting small order terms, we have a linear equation for $d\boldsymbol{\theta}$,

$$\begin{pmatrix} V^* \\ U^\top G(\tilde{\boldsymbol{\theta}}^{(i)}) \end{pmatrix} d\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{u}_0 - V^* \tilde{\boldsymbol{\theta}}^{(i)} \\ \boldsymbol{\eta}^{(i)} - U^\top \tilde{\boldsymbol{\eta}}^{(i)} \end{pmatrix}. \tag{16}$$

Since we used the approximation, we should update $\tilde{\boldsymbol{\theta}}^{(i)}$ slightly to the direction obtained by the above equation for stable convergence.

This algorithm behaves like Newton method that is a second-order algorithm and faster than the simple gradient algorithm. In practical applications, however, the simple gradient algorithm is much simpler for implementation. In particular, since optimizing U with fixing W requires a calculation of inverse of large matrix, we apply the simple gradient algorithm in simulations.

We need to control these algorithms so that they do not exceed the domain of W and U. In such a case, we need to change the learning constant sufficiently small.

VI. Parameter initialization

The algorithm described in the previous section is not guaranteed to converge to the global optimum, although each alternating procedure converges to the global optimum. Therefore, it is important to get a good initial solution. A simple idea is to take the solution of the original PCA in $\boldsymbol{\theta}$ space in the case of e-PCA. However, we can obtain a better solution, because we have information of Riemannian metric $g_{jk}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ coordinate (e-metric).

Since it is easy to calculate the metric values at sample points, let us assume a locally constant metric around the sample points[1]. Then the distance between a sample point $\boldsymbol{\theta}^{(i)}$ and another point $\boldsymbol{\theta}$ is given by $(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)})^\top G(\boldsymbol{\theta}^{(i)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i)})$. Let $\tilde{\boldsymbol{\theta}}^{(i)} = \boldsymbol{\theta}(\tilde{\boldsymbol{w}}^{(i)}; U)$ be the projection point which minimizes this distance. Then $\tilde{\boldsymbol{w}}^{(i)}$ can be expressed explicitly as a function of U and $\boldsymbol{\theta}^{(i)}$ under the assumption. Using this fact, we obtain a naive extension of PCA that minimizes the reconstruction error measured with the metric $\sum_{i=1}^n (\tilde{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{(i)})^\top G(\boldsymbol{\theta}^{(i)})(\tilde{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{(i)})$. However, an efficient algorithm is not known to minimize this cost, and in fact it may have many local optima.

Therefore, we simplify the problem further by replacing $G(\boldsymbol{\theta}^{(i)})$ by $\lambda^{(i)} G^\dagger$ which is an essentially constant metric except for the scale factor $\lambda^{(i)} \in \mathbb{R}$, i.e., the cost function is written by

$$\sum_{i=1}^n \lambda^{(i)} (\tilde{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{(i)})^\top G^\dagger (\tilde{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{(i)}), \tag{17}$$

which can be solved by the original PCA for the transformed sample $\boldsymbol{\xi}^{(i)} = \sqrt{\lambda^{(i)}} G^{\dagger 1/2} \boldsymbol{\theta}^{(i)}$. The constant shift $\boldsymbol{u}_0$ is given by the weighted sum $\boldsymbol{u}_0 = \sum_{i=1}^n \lambda^{(i)} \boldsymbol{\theta}^{(i)} / \sum_{i=1}^n \lambda^{(i)}$. Letting K be a matrix with the first h principal directions of $\boldsymbol{\xi}^{(i)}$, we get initial solutions,

$$U = G^{\dagger -1/2} K, \qquad \tilde{\boldsymbol{w}}^{(i)} = \frac{1}{\sqrt{\lambda^{(i)}}} K^\top \boldsymbol{\xi}^{(i)}. \tag{18}$$

Specifically, in this paper, we take the average metric $G^\dagger = \sum_{i=1}^n G(\boldsymbol{\theta}^{(i)})$, and $\lambda^{(i)} = \det(G(\boldsymbol{\theta}^{(i)}))$.

Note that $\tilde{\boldsymbol{\theta}}^{(i)}$ does not always belong to the domain $\mathcal{M}$. In such a case, we need to define some projection operator which maps a point $\boldsymbol{w}^{(i)} \in \mathbb{R}^h$ onto $\mathcal{M}$.

This initialization method would work well when the principal directions of G_i are not so different. Otherwise, the original PCA sometimes may give a better solution.

VII. A simple case: Finding a center

The 0-dimensional subspace fitting is the most simple example of submanifold fitting. In this case, only we need to find is $\boldsymbol{u}_0$ that minimizes the sum of e- or m-divergence. Intuitively, it is a center or representative point of samples.

Let e-center (m-center) be a point that minimizes the sum of e-divergence (m-divergence). We can easily show that e-center (m-center) is given by the arithmetic mean of samples in e-coordinate (m-coordinate),

$$\hat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\theta}^{(i)}, \qquad (\hat{\boldsymbol{\eta}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\eta}^{(i)}), \tag{19}$$

Unlike PCA, the e-PCA and m-PCA do not have hierarchical structure, i.e., a higher dimensional submanifold does not include a lower dimensional submanifold. Therefore, e-center (or m-center) is not included in higher dimensional submanifold.

VIII. A special case: Dimension reduction of random variables

Collins et al.[7] proposed a generalization of PCA for random samples generated from an exponential family. We show that this generalization is a special case of the e-PCA

Suppose we have random samples $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$, where the sample $\boldsymbol{x}^{(i)}$ is generated from unknown different distribution $p(\boldsymbol{x}; \boldsymbol{\theta}^{(i)})$ belonging to the exponential family. Even though we have only one sample for each parameter, each sample can be located as a point on the space of the exponential family by $\boldsymbol{\eta}^{(i)} = \mathbf{F}(\boldsymbol{x}^{(i)})$. In order to reduce the dimensionality of $\boldsymbol{x}$, they consider only the case $F_j(\boldsymbol{x}) = x_j$. However, the restriction can be loosen such as the number of sufficient statistics is less than or equal to the dimensionality of $\boldsymbol{x}$. They propose a method to find a linear submanifolds in e-coordinate. Therefore, their framework can be regarded as a special case of e-PCA where each point is calculated from only one sample.

This method is equivalent to the original PCA in the following case: the exponential family is a space of mean parameter of Gaussian distribution with a unit variance, $\mathcal{N}[\boldsymbol{\mu}, I]$. This space is equivalent to the Euclidean space, whose metric is used for calculating reconstruction errors.

Note that $\boldsymbol{x}^{(i)}$ are not samples from a fixed parameter. Therefore, it cannot be applied to the case that $\boldsymbol{x}^{(i)}$'s are samples from a fixed distribution $p(\boldsymbol{x}; \boldsymbol{\theta})$.

IX. An example: 1D normal distribution

Here we give numerical results for one dimensional normal distribution $\mathcal{N}(\mu, \sigma^2)$, a typical example of the exponential family.

First we summarize definitions of variables that are necessary for implementation. The density function of the normal distribution is written as

$$p(x;\mu,\sigma^2) = \exp\{-\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\}. \qquad (20)$$

Therefore, letting $F_1(x) = x^2, F_2(x) = x$, we have the natural parameters

$$\theta_1 = -\frac{1}{2\sigma^2}, \qquad \theta_2 = \frac{\mu}{\sigma^2}, \qquad (21)$$

and the expectation parameters

$$\eta_1 = \mu^2 + \sigma^2, \qquad \eta_2 = \mu. \qquad (22)$$

The domain of parameters $\mathcal{S}$ is $\theta_1 < 0$ for $\boldsymbol{\theta}$ coordinate and $\eta_1 > {\eta_2}^2$ for $\boldsymbol{\eta}$ coordinate. The coordinate transforms are $\theta_1(\boldsymbol{\eta}) = -1/2(\eta_1 - {\eta_2}^2)$, $\theta_2(\boldsymbol{\eta}) = \eta_2/\eta_1 - {\eta_2}^2$, $\eta_1(\boldsymbol{\theta}) = {\theta_2}^2/4{\theta_1}^2 - 1/2\theta_1$, $\eta_2(\boldsymbol{\theta}) = -\theta_2/2\theta_1$. The e-metric and m-metric matrices are given by

$$G(\boldsymbol{\theta}) = -\frac{1}{2{\theta_1}^3}\begin{pmatrix} {\theta_2}^2 - \theta_1 & -\theta_1\theta_2 \\ -\theta_1\theta_2 & {\theta_1}^2 \end{pmatrix}, \qquad (23)$$

$$G^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta})) = \frac{1}{(\eta_1 - {\eta_2}^2)^2}\begin{pmatrix} 1/2 & -\eta_2 \\ -\eta_2 & \eta_1 + {\eta_2}^2 \end{pmatrix}. \qquad (24)$$

First, let us consider the 0-dimensional fitting. The e-center is given by solving the following equation for $\hat{\sigma}^2$ and $\hat{\mu}$,

$$\hat{\theta}_1 = -\frac{1}{2\hat{\sigma}^2} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2(\sigma^{(i)})^2}, \qquad (25)$$

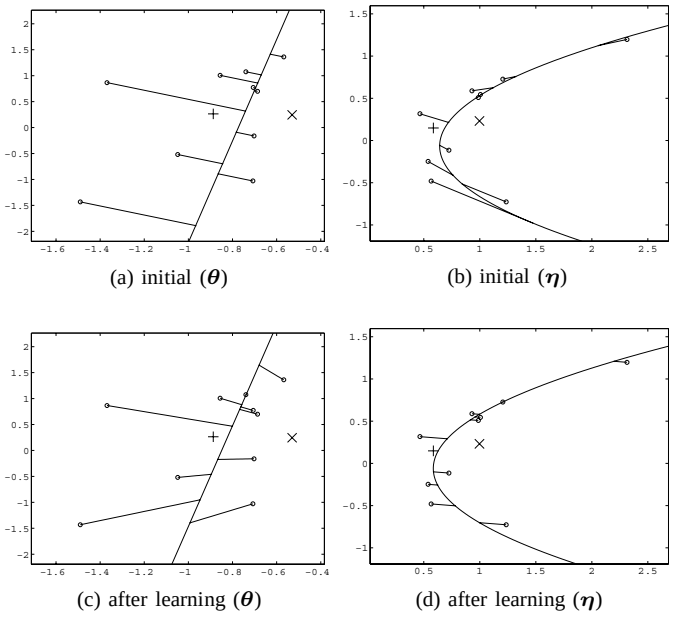
$$\hat{\theta}_2 = \frac{\hat{\mu}}{2\hat{\sigma}^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mu^{(i)}}{2(\sigma^{(i)})^2}. \qquad (26)$$

Similarly, the m-center is given by solving

$$\hat{\eta}_1 = \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n}({\sigma^{(i)}}^2 + {\mu^{(i)}}^2), \qquad (27)$$

$$\hat{\eta}_2 = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu^{(i)}. \qquad (28)$$

Next let us consider one dimensional fitting ($h = 1$). In this special example, the projection of sample points to the submanifold $\mathcal{M}$ can be obtained in a closed form by the method described in sec. IV. Therefore, we only use the gradient descent method to optimize U. The domain of the bases $\mathcal{U}$ is given by the condition that the line of the submanifold intersects the domain $\mathcal{M}$. In the case of e-PCA, almost all flat submanifold intersects the domain unless $\boldsymbol{u}_1 \propto (0,1)^\top$ and $(u_0)_1 \geq 0$. In the case of m-flat submanifold fitting, the condition can be written by the existence of solution of a quadratic equation. If the gradient algorithm exceeds the domain $\mathcal{U}$, we project the bases to some close point in the domain (we cannot project to the boundary, because the domain is an open set in this case).



(a) initial ($\boldsymbol{\theta}$) (b) initial ($\boldsymbol{\eta}$)

(c) after learning ($\boldsymbol{\theta}$) (d) after learning ($\boldsymbol{\eta}$)

Fig. 2. The e-PCA: circles represent sample points, $+$ represents e-center, $\times$ represents m-center. (a) and (b) are the initial solution obtained by the extension of PCA. (c) and (d) are the solution after 1000 steps of gradient descent learning. (a) and (c) are plotted in $\boldsymbol{\theta}$ space and (b) and (d) are in $\boldsymbol{\eta}$ space

In numerical simulations, we generated 10 points randomly from $\theta_1 = \epsilon_1$ and $\theta_2 = \theta_1 + \epsilon_2$, where ϵ_1 is generated randomly from the uniform distribution on $[-1.5, -0.5)$ and ϵ_2 from the normal distribution $N(0, 0.5^2)$. First we initialized the parameter values by the extension of PCA described in sec. VI. Then we applied the gradient algorithm for 100 steps by taking learning rate 0.01.

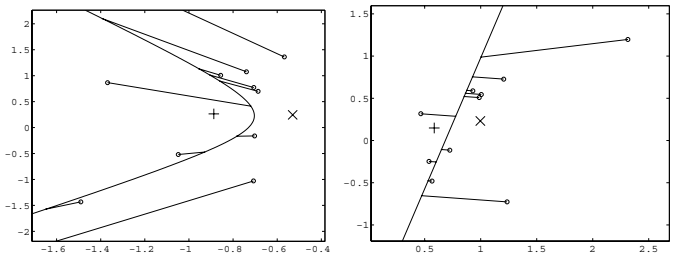
The result of e-PCA is shown in Fig. 2. In those figures, e-center and m-center are marked by '+' and '$\times$' respectively. The upper figures ((a) and (b)) represent the initial solution which are plotted in $\boldsymbol{\theta}$ space ((a)) and $\boldsymbol{\eta}$ space ((b)). In those figures, the samples (represented by circles) and the projection points are connected by the m-geodesic which are straight lines in $\boldsymbol{\eta}$ coordinate but not straight in $\boldsymbol{\theta}$ coordinate. The solution after learning is shown in (c) and (d). In this example, apparent superiority of the proposed initialization method was not clear because deviance of G_i is larger than the difference between G_i and I.

Fig. 3 shows similar figures for m-flat submanifold fitting for the same dataset. We see that the fitting curve is very different from the e-PCA case.

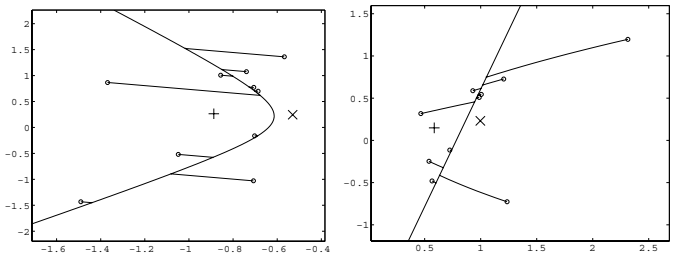
The behavior of cost functions are given in Fig. 4. It seems to decrease quickly in spite of the simple gradient algorithm.

X. Models with hidden variables

The framework described in this paper is not applicable directly to more complex models such as hidden Markov models (HMM) and mixture models, because they are not

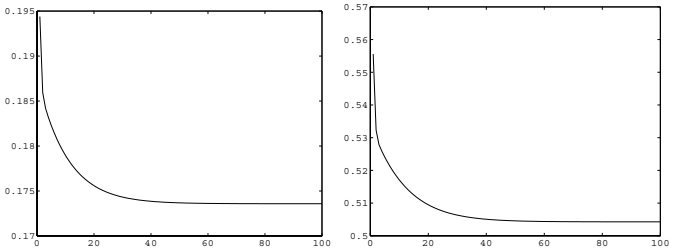


(a) initial ($\boldsymbol{\theta}$) (b) initial ($\boldsymbol{\eta}$)



(c) after learning ($\boldsymbol{\theta}$) (d) after learning ($\boldsymbol{\eta}$)

Fig. 3. The m-PCA. The arrangement follows Fig. 2



(a) e-PCA (b) m-PCA

Fig. 4. The behavior of cost function through learning. The cost functions are the sum of m-divergence for e-PCA, and e-divergence for m-PCA.

members of exponential family but a curved exponential family.

However, such models can often be regarded as an exponential family by introducing hidden random variables, i.e., even if $p(\boldsymbol{x};\boldsymbol{\theta})$ is not belonging to an exponential family, $p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ can be an exponential family, where $\boldsymbol{z}$ is a hidden variable. Therefore, we can apply our framework in the space of $p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$. This replacement is possible because we have a set of parameters not random variables. Therefore, it does not affect the inference.

XI. Concluding remarks and future directions

In this paper, we gave an information geometrical view to the submanifold fitting. We proposed fitting algorithm for two kinds of dual flat submanifolds. It raises a question about which submanifold we should choose. If we only consider the projection of one point, the m-projection is more natural from a statistical point of view. The m-projection $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is equal to the maximum likelihood estimation of $\boldsymbol{\theta}$. However, it is also known e-projection is also first-order efficient, therefore the difference is expected to be slight in the sense of asymptotic statistics. Therefore, in practical applications, we can choose a model which has better goodness of fit.

In this paper, we are only concerned with submanifold fitting. However, we can directly apply the submanifold fitting to the classification problem like as the original PCA is applied to the subspace method. We may also consider extension to classification methods other than the subspace method, such as logistic regression, linear discriminant analysis, and large margin classifiers. For that purpose, we need to define different cost functions to optimize the bases U, though it has been the same both for U and W in this paper for simplicity. In particular, when the cost function is a nonlinear function of divergence as the cost function, the resulting gradient algorithm is essentially similar to the case of e- and m-PCA.

Another direction is to extend the framework to nonflat submanifold fitting. Recently, the nonlinear surface fitting in the Euclidean or functional space has been extensively studied from various approaches, such as kernel PCA[10], principal surface[8], ISOMAP[15] and locally linear embedding[13]. The extension of those methods to a Riemannian space will give more natural fitting in some applications, and it remains as future works.

References

[1] S. Akaho, SVM that maximizes the margin in the input space, *Proc. of ICONIP 2002* (2002)

[2] S. Amari, *Differential Geometrical Methods in Statistics*, Springer-Verlag (1985)

[3] S. Amari, Information geometry of the EM and em algorithms for neural networks, *Neural Networks*, 8 (1995)

[4] S. Amari, H. Nagaoka, *Methods of Information Geometry*, AMS and Oxford university press (2000)

[5] S. Amari, Information Geometry on Hierarchy of Probability Distributions, *IEEE Trans. on Information Theory*, 47 (2001)

[6] I. Csiszár, G. Tusnády, Information geometry and alternating minimization procedures, *Statistics and Decisions, Supplement Issue*, 1 (1984)

[7] M. Collins, S. Dasgupta, R.E. Schapire, A Generalization of Principal Component Analysis to the Exponential Family, *Advances in Neural Information Processing Systems*, 14 (2002)

[8] T. Hastie and W. Stuetzle, Principal curves, *Journal of the American Statistical Association*, 84 (1989)

[9] S. Ikeda, T. Tanaka, S. Amari, Information geometrical framework for analyzing belief propagation decoder, *Advances in Neural Information Processing Systems*, 14 (2002)

[10] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, G. Ratsch, Kernel PCA and de-noising in feature spaces, *Advances in Neural Information Processing Systems* 11 (1999)

[11] N. Murata, T. Takenouchi, T. Kanamori, S. Eguchi, Information geometry of U-Boost and Bregman divergence, *Research Memorandum No.860*, Institute of Statistical Mathematics (2002)

[12] A. Ohara, Information geometric analysis of an interior point method for semidefinite programming, In O.E. Barndorff-Nielsen and E.B. Vedel Jensen (eds), *Geometry in Present Day Science*, World Scientific (1999).

[13] S. Roweis, L. Saul, Nonlinear dimension reduction by locally linear embedding, *Science*, 290 (2000)

[14] T. Tanaka, Information geometry of mean-field approximation, In M. Opper and D. Saad (eds) *Advanced Mean Field Methods — Theory and Practice*, MIT Press (2001)

[15] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, 290 (2000)

[16] K. Tsuda, S. Akaho, K. Asai, The *em* Algorithm for Kernel Matrix Completion with Auxiliary Data, *J. Machine Learning Research*, 4 (2003)