

# A Computational Approach to Fisher Information Geometry with Applications to Image Analysis

Washington Mio<sup>1</sup>, Dennis Badlyans<sup>2</sup>, and Xiuwen Liu<sup>3</sup>

<sup>1</sup> Department of Mathematics, Florida State University  
Tallahassee, FL 32306-4510, USA  
[mio@math.fsu.edu](mailto:mio@math.fsu.edu)

<sup>2</sup> Department of Mathematics, Florida State University  
Tallahassee, FL 32306-4510, USA  
[dbadlyan@math.fsu.edu](mailto:dbadlyan@math.fsu.edu)

<sup>3</sup> Department of Computer Science, Florida State University  
Tallahassee, FL 32306-4530, USA  
[liux@cs.fsu.edu](mailto:liux@cs.fsu.edu)

**Abstract.** We develop a computational approach to non-parametric Fisher information geometry and algorithms to calculate geodesic paths in this geometry. Geodesics are used to quantify divergence of probability density functions and to develop tools of data analysis in information manifolds. The methodology developed is applied to several image analysis problems using a representation of textures based on the statistics of multiple spectral components. Histograms of filter responses are viewed as elements of a non-parametric statistical manifold, and local texture patterns are compared using information geometry. Appearance-based object recognition experiments, as well as region-based image segmentation experiments are carried out to test both the representation and metric. The proposed representation of textures is also applied to the development of a *spectral cartoon model* of images.

## 1 Introduction

Large ensembles of data are often modeled as random samples of probability distributions. As such, the algorithmic analysis of complex data sets naturally leads to the investigation of families of probability density functions (PDFs). Of particular interest is the development of metrics to quantify divergence of PDFs and to model similarities and variations observed within a family of PDFs.

*Information geometry* studies differential geometric structures on *manifolds of probability distributions* and provides important tools for the statistical analysis of families of probability density functions. While this area has experienced a vigorous growth on the theoretical front in recent decades (see e.g. [1, 2, 18]), the development of corresponding computational tools for data analysis is still somewhat incipient. In this paper, we develop novel computational methods and

strategies for the algorithmic study of non-parametric Fisher information geometry and investigate applications to problems in image analysis.

As geodesics are natural interpolators in Riemannian manifolds, a key element developed in this paper is an algorithm to calculate geodesics and geodesic distances in the geometry associated with Fisher information. This basic tool will allow us to devise computational approaches to problems such as: (i) quantifying similarity and divergence of PDFs; (ii) interpolating and extrapolating PDFs; (iii) clustering probability density functions; (iv) dimensionality reduction in the representation of families of PDFs; (v) development of statistical models that account for variability observed within a class of density functions. For simplicity, we shall focus our investigation on PDFs defined on a finite interval, which will be normalized to be  $I = [0, 1]$ , with respect to the Lebesgue measure. However, the techniques apply to more general settings.

Since the introduction of the discrete *cartoon model* of images by Geman and Geman [8] and Blake and Zisserman [3], and its continuous analogue by Mumford and Shah [17], many variants followed and have been applied to a wide range of image processing tasks [4]. In these models, an image is typically viewed as composed of two basic elements: (i) a *cartoon* formed by regions bounded by sharp edges, within which the variation of pixel values is fairly smooth; (ii) a *texture pattern* within each region, which is frequently modeled as white noise. A drawback in such approaches is the texture model adopted; the view that texture is not noise, but some form of *structured appearance* is becoming prevalent. To address this problem, models such as the *spectrogram model* [15, 24] have been proposed (see also [9]). A common strategy in texture analysis has been to decompose images into their spectral components using bandpass filters and utilize histograms of filter responses to represent textures. Zhu et al. [23] have shown that marginal distributions of spectral components are sufficient to characterize homogeneous textures; other studies of the statistics of spectral components include [19, 6, 22]. Experiments reported in [14] offer empirical evidence that the same applies to non-homogeneous textures if adequate boundary conditions are available; that is, enough pixel values near the boundary of the image domain are known.

In this paper, we model local and global texture patterns using histograms of spectral components viewed as elements of a non-parametric information manifold. Geodesic distances in this manifold are used to quantify texture similarity and divergence. Multi-scale texture representation and analysis can be carried out within this framework by restricting spectral components to sub-windows of the image domain of varying sizes. Several experiments involving appearance-based object classification and recognition, as well as region-based image segmentation are carried out to test the proposed representation and methodology. We also introduce a multi-scale *spectral cartoon model* of images based on this information theoretical representation of textures.

The paper is organized as follows. In Sec. 2, we briefly review basic facts about Fisher information geometry for parametric families of PDFs. Secs. 3 and 4 are devoted to a computational treatment of non-parametric Fisher informa-

tion and to the development of an algorithm to calculate geodesics in information manifolds. Secs. 5.1, 5.2 address the calculation of basic statistics of collections of PDFs such as means and tangent-space covariance; this is followed by a discussion of clustering techniques in Sec. 5.3. In the remaining sections, the techniques developed are applied to various problems in image analysis including appearance-based recognition and segmentation of images. In the last section, we propose a new *spectral cartoon model* of images, which is a modified version of the well-known Mumford-Shah model that includes a novel representation of textures using histograms of spectral components viewed as elements of a non-parametric information manifold.

## 2 Fisher Information

Let  $I = [0, 1]$  and  $p: I \times \mathbb{R}^k \rightarrow \mathbb{R}^+$ ,  $(x, \theta) \mapsto p(x; \theta)$ , a  $k$ -dimensional family of positive probability density functions parameterized by  $\theta \in \mathbb{R}^k$ . In classical information geometry, the Riemannian structure on the parameter space  $\mathbb{R}^k$  defined by the *Fisher information matrix*  $g$ , whose  $(i, j)$ -entry is

$$g_{ij}(\theta) = \int_0^1 \left( \frac{\partial}{\partial \theta_i} \log p(x; \theta) \right) \left( \frac{\partial}{\partial \theta_j} \log p(x; \theta) \right) p(x; \theta) dx,$$

is regarded as the most natural Riemannian structure on the family from the viewpoint of information theory (see e.g. [2]). Recall that if  $p_1, p_2: I \rightarrow \mathbb{R}$  are positive PDFs, the *Kullback-Leibler (KL) divergence* is defined by

$$KL(p_1, p_2) = \int_0^1 \log \left( \frac{p_1(x)}{p_2(x)} \right) p_1(x) dx.$$

If restricted to the family  $p(x; \theta)$ , the *KL divergence* may be viewed as a function of parameters, with domain  $\mathbb{R}^k \times \mathbb{R}^k$ . If  $\theta, \xi \in \mathbb{R}^k$ , we use the notation  $KL(\theta, \xi)$  for  $KL(p(\cdot, \theta), p(\cdot, \xi))$ . Infinitesimally, the double of the *KL divergence* is known to coincide with the quadratic form

$$ds^2 = \sum_{i,j=1}^k g_{ij}(\theta) d\theta_i d\theta_j$$

associated with the Fisher information matrix  $g$ . That is, if  $E_\theta: \mathbb{R}^k \rightarrow \mathbb{R}$  is the energy functional  $E_\theta(\xi) = KL(\theta, \xi)$ , the Hessian of  $E_\theta$  at the point  $\xi = \theta$  is given by  $g(\theta)$ . This fact is often expressed as

$$KL(\theta, \theta + d\theta) = \frac{1}{2} ds^2.$$

In [5], Dawid suggested the investigation of non-parametric analogues of this geometry and such model was developed by Pistone and Sempi in [18]. This extension to non-parametric families of a.e. positive density functions involves the

study of infinite-dimensional manifolds. For technical reasons, the geometry that generalizes Fisher information falls in the realm of infinite-dimensional Banach manifolds, spaces whose geometries are more difficult to analyze. One of our main goals is to develop computational approaches to discrete versions of non-parametric Fisher information obtained by sampling density functions  $p: I \rightarrow \mathbb{R}$  at a finite set of points.

### 3 Non-Parametric Information Manifolds

We investigate a non-parametric statistical manifold  $\mathcal{P}$  whose elements represent the log-likelihood of positive probability density functions  $p: I \rightarrow \mathbb{R}^+$ . The manifold  $\mathcal{P}$  will be endowed with an information-theoretic geometric structure which, among other things, will allow us to quantify variations and dissimilarities of PDFs.

Each tangent space  $T_\varphi\mathcal{P}$  will be equipped with a natural inner product  $\langle \cdot, \cdot \rangle_\varphi$ . Although a Hilbert-Riemannian structure might seem to be the natural geometric structure on  $\mathcal{P}$  to expect, one is led to a manifold locally modeled on Banach spaces [18]. Since, in this paper, we are primarily interested in computational aspects of information geometry, we construct finite-dimensional analogues of  $\mathcal{P}$  by sampling probability density functions uniformly at a finite set of points under the assumption that they are continuous. Then, arguing heuristically, we derive an expression for the inner product on the tangent space  $T_\varphi\mathcal{P}$ , which induces a Riemannian structure on finite-dimensional, non-parametric analogues of  $\mathcal{P}$ . From the viewpoint of information theory, the geodesic distance between two PDFs can be interpreted as a measurement of the *uncertainty* or *unpredictability* in a density function relative to the other. Throughout the paper, we abuse notation and refer to both continuous and discrete models with the same symbols; however, the difference should be clear from the context.

Positive PDFs will be represented via their log-likelihood  $\varphi(x) = \log p(x)$ . Thus, a function  $\varphi: I \rightarrow \mathbb{R}$  represents an element of  $\mathcal{P}$  if and only if it satisfies

$$\int_I e^{\varphi(x)} dx = 1. \quad (1)$$

*Remark.* In the discrete formulation,  $\varphi$  denotes the vector  $(\varphi(x_1), \dots, \varphi(x_n))$ , where  $0 = x_1 < x_2 < \dots < x_n = 1$  are  $n$  uniformly spaced points on the unit interval  $I$ .

Tangent vectors  $f$  to the manifold  $\mathcal{P}$  at  $\varphi$  represent infinitesimal (first-order) deformations of  $\varphi$ . Using a “time” parameter  $t$ , write such variation as  $\varphi(x, t)$ ,  $x \in I$  and  $t \in (-\epsilon, \epsilon)$ , where

$$\varphi(x, 0) = \varphi(x) \quad \text{and} \quad f(x) = \frac{d}{dt} \varphi(x, 0).$$

Differentiating constraint (1) with respect to  $t$  at  $t = 0$ , it follows that  $f: I \rightarrow \mathbb{R}$  represents a tangent vector at  $\varphi$  if and only if

$$\int_I f(x) e^{\varphi(x)} dx = 0. \quad (2)$$

This simply means that  $f(x)$  has null expectation with respect to  $e^{\varphi(x)} dx$ . Thus, the tangent space  $T_\varphi\mathcal{P}$  to the manifold  $\mathcal{P}$  at  $\varphi$  can be described as

$$T_\varphi\mathcal{P} = \{f: I \rightarrow \mathbb{R} \mid \int_0^1 f(x)e^{\varphi(x)} dx = 0\}.$$

What is the natural inner product on  $T_\varphi\mathcal{P}$  that generalizes Fisher information? In Sec. 2, we remarked that the Fisher information matrix can be viewed as the Hessian of an energy functional associated with the *KL* divergence. In the non-parametric setting, for  $\varphi \in \mathcal{P}$ , the corresponding energy is given by

$$E_\varphi(\psi) = KL(e^\varphi, e^\psi) = \int_0^1 (\varphi(x) - \psi(x)) e^{\varphi(x)} dx.$$

Calculating the Hessian of  $E_\varphi$  at  $\psi = \varphi$ , it follows that the inner product induced on  $T_\varphi\mathcal{P}$  is given by

$$\langle v, w \rangle_\varphi = \int_I v(x)w(x)e^{\varphi(x)} dx, \quad (3)$$

which agrees with Fisher information on parametric submanifolds. A similar calculation with the *Jensen-Shannon (JS) entropy divergence*, which is a symmetrization of *KL*, leads to the same inner product, up to a multiplicative factor independent of  $\varphi$ . This means that both *KL* and *JS* essentially yield the same infinitesimal measure of relative uncertainty or lack of information.

Continuing with this informal argument, Eq. 2 can be rewritten as  $\langle f, 1 \rangle_\varphi = 0$ , where 1 denotes the constant function 1. Thus,  $f$  is tangent to  $\mathcal{P}$  if and only if it is orthogonal to 1.

To discretize this model, let  $0 = x_1 < x_2 < \dots < x_{n-1} < x_n = 1$  be  $n$  uniformly spaced points on the interval  $I$ . Heretoforth, all functions  $\psi: I \rightarrow \mathbb{R}$  will be viewed as  $n$ -vectors obtained by sampling the function at these points; that is,  $\psi = (\psi_1, \dots, \psi_n) \in \mathbb{R}^n$ , with  $\psi_i = \psi(x_i)$ . Eqn. 3 suggests that, at each  $\varphi \in \mathbb{R}^n$ , we consider the inner product

$$\langle f, g \rangle_\varphi = \sum_{i=1}^n f_i g_i e^{\varphi_i}, \quad (4)$$

where  $f, g \in \mathbb{R}^n$ . From (1), it follows that  $\varphi \in \mathbb{R}^n$  represents a (discretized) PDF if and only if  $\sum_{i=1}^n e^{\varphi_i} = 1$ . More formally, consider the function  $F: \mathbb{R}^n \rightarrow \mathbb{R}$  given by

$$F(\varphi) = \sum_{i=1}^n e^{\varphi_i}. \quad (5)$$

The differential of  $F$  at  $\varphi$  evaluated at a vector  $f \in \mathbb{R}^n$  is

$$dF_\varphi(f) = \sum_{i=1}^n f_i e^{\varphi_i} = \langle f, 1 \rangle_\varphi,$$

which shows that the gradient of  $F$  at  $\varphi$  with respect to the inner product  $\langle \cdot, \cdot \rangle_\varphi$  is  $\nabla F(\varphi) = (1, \dots, 1)$ , for any  $\varphi$ . This implies that the level sets of  $F$  are  $(n - 1)$ -dimensional submanifolds of  $\mathbb{R}^n$ . Of particular interest, is the manifold

$$\mathcal{P}_n = F^{-1}(1),$$

which is our finite-dimensional analogue of  $\mathcal{P}$ . As in (2), the tangent space  $T_\varphi \mathcal{P}_n$  consists of all vectors  $f \in \mathbb{R}^n$  satisfying  $\langle f, 1 \rangle_\varphi = 0$ ; that is, vectors  $f \in \mathbb{R}^n$  orthogonal to  $(1, \dots, 1)$  with respect to the inner product  $\langle \cdot, \cdot \rangle_\varphi$ . The *geodesic distance* between  $\varphi, \psi \in \mathcal{P}_n$  will be denoted  $d(\varphi, \psi)$ .

## 4 Geodesics in $\mathcal{P}_n$

We are interested in developing an algorithm to calculate geodesics in  $\mathcal{P}_n$  (with respect to the Levi-Civita connection) with prescribed boundary conditions; that is, with given initial and terminal points  $\varphi$  and  $\psi$ , respectively. Following a strategy similar to that adopted in [13] for planar shapes, we propose to construct geodesics in two stages. First, we describe a numerical strategy to calculate geodesics in  $\mathcal{P}_n$  with prescribed initial position  $\varphi_0$  and initial velocity  $f_0$ .

### 4.1 Geodesics with Prescribed Initial Conditions

Recall that  $\mathcal{P}_n$  is a submanifold of the Riemannian manifold  $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_\varphi)$ . From (4),  $\mathbb{R}^n$  with this metric can be expressed as the  $n$ -fold Cartesian product of the real line  $\mathbb{R}$  equipped with the Riemannian metric

$$\langle u, v \rangle_x = uv e^x,$$

$x \in \mathbb{R}$ . This allows us to easily calculate the differential equation that governs geodesics in  $\mathbb{R}^n$  with this non-standard (flat) metric and derive explicit expressions for geodesics with initial conditions prescribed to first order. To solve the corresponding initial value problem in  $\mathcal{P}_n$ , we adopt the following strategy:

- (i) Infinitesimally, follow the geodesic path  $\alpha_0$  in  $\mathbb{R}^n$  satisfying the given initial conditions.
- (ii) The end point  $\tilde{\varphi}_1$  of this small geodesic arc in  $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_\varphi)$  will typically fall slightly off of  $\mathcal{P}_n$ ; to place it back on the level set  $\mathcal{P}_n = F^{-1}(1)$  (i.e., to have equation  $F(\tilde{\varphi}_1) - 1 = 0$  satisfied), we use Newton's method. Since the gradient of  $F$  is  $1 = (1, \dots, 1)$  at any point, this projection can be accomplished in a single step since it is equivalent to simply adding a constant to  $\tilde{\varphi}_1$  so that  $\int_0^1 e^{\tilde{\varphi}_1(x)} dx = 1$ . This gives  $\varphi_1 \in \mathcal{P}_n$ .
- (iii) To iterate the construction, we need to parallel transport the velocity vector  $f_0$  to the new point  $\varphi_1$  along the estimated geodesic arc. As an approximation to the parallel transport, from the velocity vector of  $\alpha_0$  at the end point  $\tilde{\varphi}_1$ , subtract the component normal to  $\mathcal{P}_n$  at  $\varphi_1$  and rescale it to have the same magnitude as  $f_0$  to obtain the velocity vector  $f_1$  at  $\varphi_1$ . This is done because

geodesics have constant speed. One can show that this approximation of the parallel transport is a mild variant of Euler’s method applied to the differential equation of geodesics in  $\mathcal{P}_n$ .

(iv) Iterate the construction.

We denote this geodesic by  $\Psi(t; \varphi_0, f_0)$ , where  $t$  is the time parameter. The position  $\Psi(1; \varphi_0, f_0)$  of the geodesic at time  $t = 1$  is known as the exponential of  $f_0$  and denoted

$$\exp_{\varphi_0}(f_0) = \Psi(1; \varphi_0, f_0).$$

One often refers to  $f_0$  as a logarithm of  $\varphi_1 = \Psi(1; \varphi_0, f_0)$ , denoted  $f_0 = \log_{\varphi_0} \varphi_1$ . The procedure just described can be interpreted as a first-order numerical integration of the differential equation that governs geodesics in  $\mathcal{P}_n$ . Higher-order methods can be adapted similarly.

## 4.2 Geodesics with Boundary Conditions

Given two points  $\varphi, \psi \in \mathcal{P}_n$ , how to find a geodesic in  $\mathcal{P}_n$  connecting them? Similar to the strategy for computing geodesics in shape manifolds developed in [13], we propose to use a *shooting method*. If we solve the equation

$$\Psi(1; \varphi, f) = \psi \tag{6}$$

for  $f \in T_{\varphi}\mathcal{P}_n$  (i.e., if we find the correct direction  $f_0$  to shoot a geodesic to reach  $\psi$  in unit time), then  $\Psi(t; \varphi, f_0)$ ,  $0 \leq t \leq 1$ , gives the desired geodesic. Solving Eqn. 6 is equivalent to finding the zeros of the *miss function*

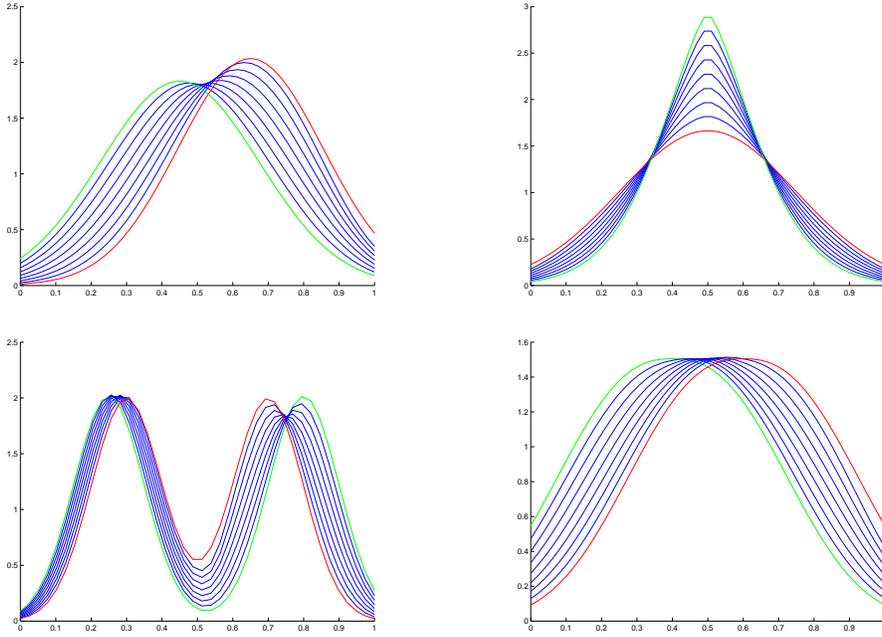
$$E(f) = \|\Psi(1; \varphi, f) - \psi\|^2$$

on the tangent space  $T_{\varphi}\mathcal{P}_n$ , where  $\|\cdot\|$  denotes the standard Euclidean norm. This problem can be approached numerically via Newton’s method.

Fig. 1 shows some examples of geodesics in  $\mathcal{P}_n$  computed with the algorithmic procedure described above. The examples involve probability density functions obtained by truncating Gaussians, generalized Laplacians and mixtures of Gaussians.

## 5 Data Analysis

Probability density functions are often used to model large ensembles of data viewed as random samples of the model. To compare different ensembles, it is desirable to adapt existing data analysis tools to the framework of Fisher information. In this section, we consider the problem of defining and computing means and covariance of families of PDFs, as well as extending clustering techniques to the manifold  $\mathcal{P}_n$ . Note that each PDF will be treated as a point on an information manifold.



**Fig. 1.** Examples of geodesics in  $\mathcal{P}_n$ .

### 5.1 Fréchet Means

We begin with the problem of defining and finding the mean of a family of PDFs on the interval  $I = [0, 1]$ .

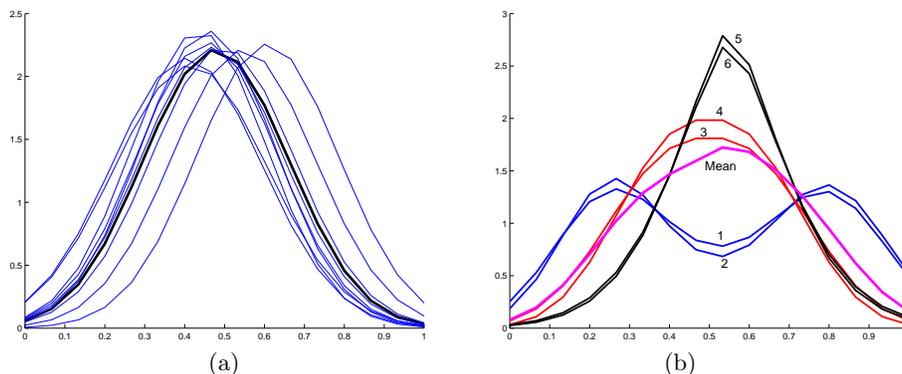
Let  $S = \{\varphi_1, \dots, \varphi_\ell\} \subset \mathcal{P}_n$  be a collection of  $\ell$  probability distributions over  $I$  represented by their discretized log-likelihood functions. We are interested in defining and computing sample statistics such as mean and covariance of  $S$ . We propose to use the intrinsic notion of *Fréchet mean* [12], which is defined as a (local) minimum of the *total variance function*

$$V(\varphi) = \frac{1}{2} \sum_{i=1}^{\ell} d^2(\varphi, \varphi_i),$$

where  $d$  denotes geodesic distance in  $\mathcal{P}_n$ . It can be shown [12] that if  $f_i$  is the initial velocity of the geodesic that connects  $\varphi$  to  $\varphi_i$  in unit time, then

$$\nabla V(\varphi) = - \sum_{i=1}^{\ell} f_i.$$

Thus, if we compute the velocity vectors  $f_i$ ,  $1 \leq i \leq \ell$ , using the algorithmic procedure described in Sec. 4, the calculation of Fréchet means can be approached with gradient methods. Fig. 2 shows examples of Fréchet means computed with the techniques just described.



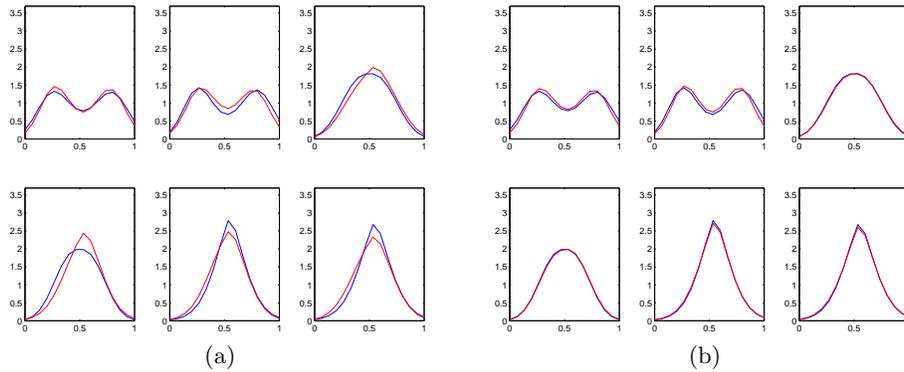
**Fig. 2.** Examples of Fréchet means of PDFs.

## 5.2 Covariance, Dimension Reduction and Learning

Given  $S = \{\varphi_1, \dots, \varphi_\ell\} \subset \mathcal{P}_n$ , let  $\hat{\varphi} \in \mathcal{P}_n$  be a Fréchet mean of the collection  $S$ , and let  $f_i = \log_{\hat{\varphi}} \varphi_i$  be the initial velocity of the geodesic that connects  $\hat{\varphi}$  to  $\varphi_i$  in unit time, which can be calculated with the techniques of Sec. 4.

The vectors  $f_i$ ,  $1 \leq i \leq \ell$ , yield an alternative *tangent-space representation* of the elements of  $S$  as vectors in  $T_{\hat{\varphi}}\mathcal{P}_n$ . This type of representation of data on Riemannian manifolds via the inverse exponential map was introduced in the context of shape analysis using landmark representations of shapes (see e.g. [7]). Note that the original data point  $\varphi_i$  can be recovered from  $f_i$  via the exponential map. The advantage of this representation is that each  $f_i$  lies in the inner-product space  $(T_{\hat{\varphi}}\mathcal{P}_n, \langle \cdot, \cdot \rangle_{\hat{\varphi}})$ , where classical data analysis techniques such as Component Analysis can be used. This tangent-space representation may, in principle, distort the geometry of the data somewhat. However, the distortion is small if the data does not exhibit very large spread. This is often the case, in practice, if we assume that the data has been pre-clustered and we are analyzing individual clusters separately.

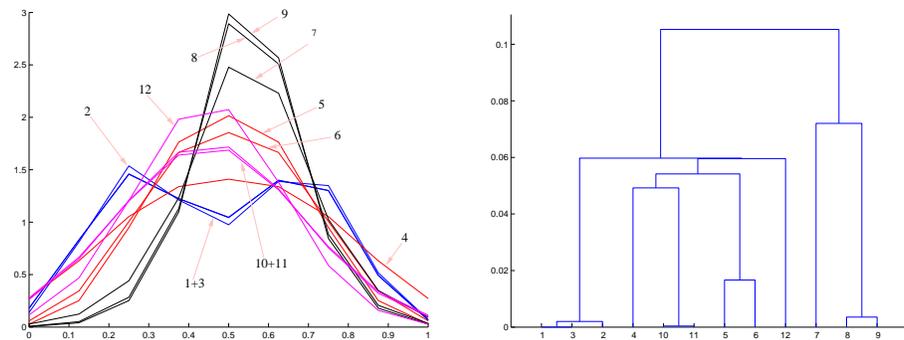
Once the data has been lifted to  $(T_{\hat{\varphi}}\mathcal{P}_n, \langle \cdot, \cdot \rangle_{\hat{\varphi}})$ , covariance can be defined as usual. One can learn probability models for the family  $\{f_i, 1 \leq i \leq \ell\}$  using standard techniques (see e.g. [7, 21]). For example, *Principal Component Analysis* (PCA) can be applied to the tangent-space representation of  $S$  to derive a Gaussian model. Fig. 2(b) shows a set of six PDFs and their mean. Figs. 3 (a) and (b) show tangent-space PCA reconstructions of the data using projections over one and three principal directions, respectively. Other well-known data analysis methods such as *Independent Component Analysis* and *Kernel PCA* (see e.g. [11, 20, 10]) can be applied to tangent-space representations, as well.



**Fig. 3.** Reconstructing the data shown in Fig. 2(b) with tangent-space Principal Component Analysis using (a) one and (b) three principal components, respectively.

### 5.3 Clustering

Classical clustering algorithms can be adapted to the present setting to group large collections of PDFs into smaller subclasses. For example, starting with single-element clusters formed by each element in a dataset, one can use hierarchical clustering techniques to successively combine clusters using the geodesic distance between clusters in  $\mathcal{P}_n$  as a merging criterion. Fig. 4 shows the clustering dendrogram obtained for a family twelve PDFs using the nearest-neighbor merging criterion. The techniques for calculating Fréchet means described in



**Fig. 4.** Twelve PDFs and a “nearest-neighbor” hierarchical clustering dendrogram.

Sec. 5.1 allow us to adapt the classical *k-Means Clustering Algorithm* to families of PDFs using the geodesic distance, as well.

## 6 Spectral Representation and Clustering of Texture

In this section, we employ histograms of spectral components to represent texture patterns in images at different scales. We carry out appearance-based object recognition experiments to test the representation, geodesic metric, and geodesic interpolation. We also use the geodesic metric and a variant of the clustering techniques discussed in Sec. 5.3 to identify regions of similar appearance in images.

### 6.1 Texture Representation

Given a bank of filters  $\mathcal{F} = \{F^j, 1 \leq j \leq K\}$  and an image  $I$ , let  $I^j$  be the associated spectral components obtained by applying filter  $F^j$  to the image. Assume that the histogram of the  $j$ th spectral component is modeled on a PDF with log-likelihood  $\varphi_j \in \mathcal{P}$ . The (texture of) image  $I$  will be represented by the  $K$ -tuple  $\Phi = (\varphi^1, \dots, \varphi^K) \in \mathcal{P} \times \dots \times \mathcal{P} = \mathcal{P}^K$ . We should point out that this is a global representation of the image  $I$ , but the same construction applied to local windows leads to *multi-scale representations* of texture patterns. If  $\Phi_A, \Phi_B \in \mathcal{P}^K$  represent images  $I_A$  and  $I_B$ , respectively, let  $d_T$  be the root-mean-square geodesic distance

$$d_T(\Phi_A, \Phi_B) = \left( \frac{1}{K} \sum_{j=1}^K d^2(\varphi_A^j, \varphi_B^j) \right)^{1/2}, \quad (7)$$

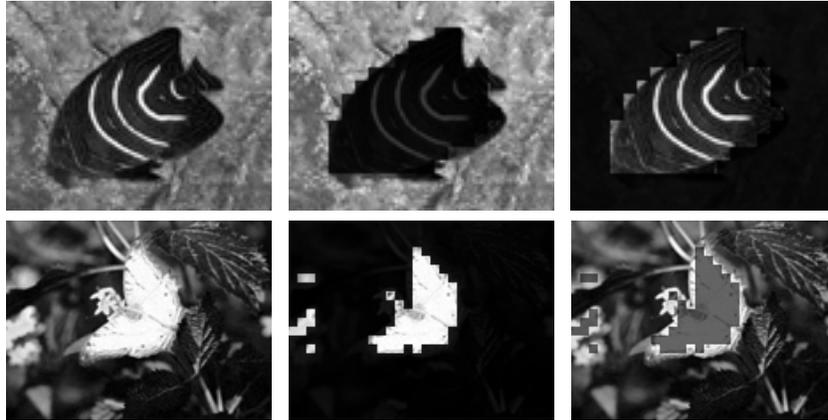
which defines a metric on the space  $\mathcal{P}^K$  of texture representations.

*Remark.* In specific applications, one may wish to attribute different weights to the various summands of  $d_T(\Phi_A, \Phi_B)$  in order to emphasize particular filters.

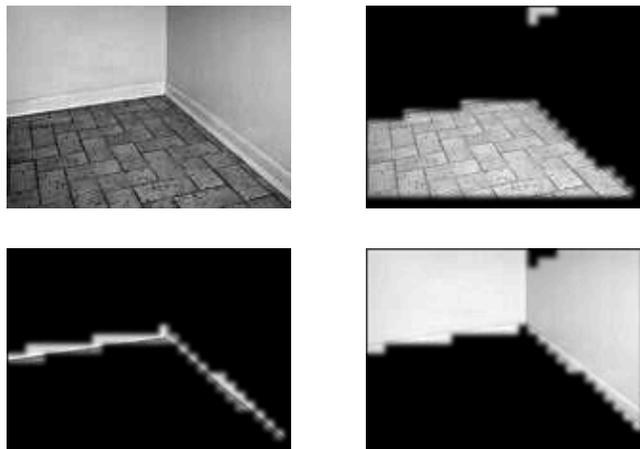
### 6.2 Region-Based Segmentation

In this section, we present results obtained in image segmentation experiments with the ideas discussed above. To illustrate the ability of the metric  $d_T$  to discern and classify local texture patterns, we grouped the pixels of some images into two clusters. We used local histograms associated with five distinct spectral components and the metric  $d_T$  as measure of dissimilarity; a hierarchical ‘‘centroid’’ clustering was adopted.

On each row of Fig. 5, the leftmost panel shows the original images. The other two panels display the two clusters obtained highlighted in different ways; observe that clusters may be disconnected as in the image with a butterfly. Since clustering was performed at a low resolution, the boundaries of the regions are somewhat irregular. Note that because of the resolution and the local window size utilized in the spectral analysis, the relatively thin white stripes on the fish are clustered with the rest of the fish, not with the background. Fig. 6 shows the results of a similar experiment, where the image was decomposed into three regions.



**Fig. 5.** In each row, the leftmost panel displays a test image. On the other panels, the regions obtained by clustering the pixels into two clusters using histograms of local responses to 5 filters are highlighted in two different ways.



**Fig. 6.** A low-resolution segmentation of an image into three regions by clustering pixels using 5 spectral components and the geodesic metric derived from Fisher information.

### 6.3 Appearance-Based Recognition

As an illustration of possible uses of the proposed spectral representation of textures using information manifolds, we carried out a small object recognition experiment using 10 objects from the COIL-100 database. Each object in the database is represented by 72 images taken at successive views that differ by 5-degree angles. We used histograms of 39 spectral components, as well as the histograms of the original images, so that each image is represented as an element of  $\mathcal{P}^{40}$ .

In the first recognition experiment, the training set consisted of 4 images corresponding to 90-degree rotations of the objects, with the 68 remaining images used as test images. Table 1 compares the recognition rates achieved with 4 training images for each object to those obtained by estimating four additional views using geodesic interpolations. Similar results are shown for a training set of 8 images. Examples of histograms of intermediate views estimated using geodesic

**Table 1.** Recognition rates in an experiment with 10 objects from the COIL-100 database.

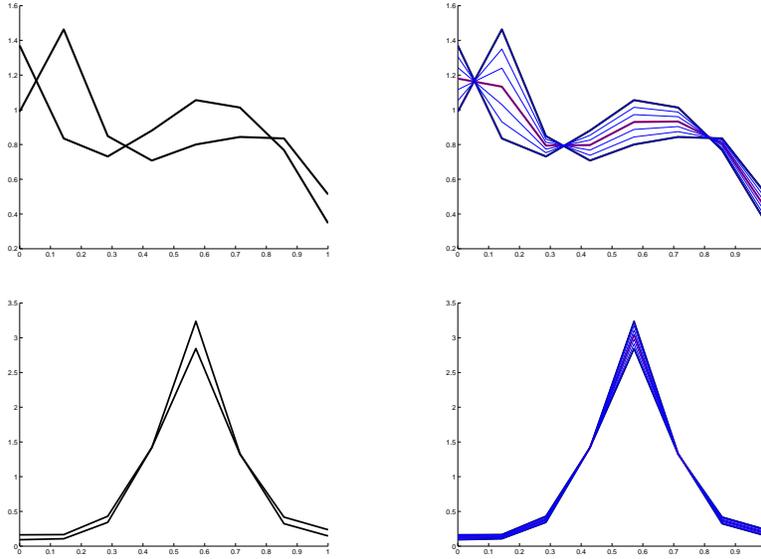
# of Training Images	# of Test Images	Performance with no Interpolations	Performance with Interpolations
4	68	93%	95%
8	64	97%	100%

interpolations are shown in Fig. 7. On the first row, we display histograms of images of an object from angles differing by  $90^\circ$  and the corresponding interpolation. A similar illustration for a spectral component of the image is shown on the second row.

## 7 The Spectral Cartoon Model

To model texture patterns using *multi-resolution* spectral components of images, we localize the notion of appearance, as follows. Given a bank of filters  $\mathcal{F} = \{F_1, \dots, F_K\}$  and an image  $I$ , let  $I^j$ ,  $1 \leq j \leq K$  be the associated spectral components. For a pixel  $p$ , consider a window of fixed size (this determines the scale) centered at  $p$  and let  $h_p^j$  be the histogram of  $I^j$  restricted to this window. The histograms  $h_p^j$  yield an  $s$ -tuple  $\Phi_p = (\varphi_p^1, \dots, \varphi_p^K) \in \mathcal{P}^K$ , which encodes the local texture pattern near the pixel  $p$ . If  $p, q$  are two pixels, we use the distance  $d_T(\Phi_p, \Phi_q)$  defined in (7) to quantify texture divergence.

To simplify the discussion, we consider a binary model and assume that the image consists of two main regions: a background and a single foreground element, which are separated by a closed contour  $C$ . The proposed model can be



**Fig. 7.** First row: the left panel shows histograms of images of an object taken from different views and the right panel displays histograms of intermediate views estimated using geodesic interpolations. Second row: similar illustration for a spectral component.

modified to allow more complex configurations as in [17]. A key difference to be noted is that unlike the classical Ising image model, where a binary cartoon is adopted (see e.g. [15]), we make a similar assumption at the level of spectral representations, so that even cartoons can be non-trivially textured. Thus, variations of pixel values, often treated as white noise, will be modeled on random fluctuations of more richly structured texture patterns.

Let  $I: D \rightarrow \mathbb{R}$  be an image, where  $D$  is the image domain, typically a rectangle in  $\mathbb{R}^2$ . Consider triples  $(\Phi_{in}, \Phi_{out}, C)$ , where  $C$  is a closed contour in  $D$ , and  $\Phi_{in}, \Phi_{out} \in \mathcal{P}^K$  represent cartoon models for the local texture patterns in the regions inside and outside  $C$ , respectively. We adopt a Bayesian model, with a prior that assumes that  $C$  is not “unnecessarily” long, so that the prior energy will be a multiple of the length  $\ell(C)$ . This can be easily modified to accommodate other commonly used priors such as the elastic energy; a probabilistic interpretation of the elastic energy is given in [16]. The proposed data likelihood energy is of the form

$$\begin{aligned}
 E_d(I|\Phi_{in}, \Phi_{out}, C) &= \alpha \int_{D_{in}} d_T^2(\Phi_p, \Phi_{in}) dp \\
 &+ \beta \int_{D_{out}} d_T^2(\Phi_p, \Phi_{out}) dp,
 \end{aligned} \tag{8}$$

where  $\alpha, \beta > 0$ , and  $D_{in}, D_{out}$  are the regions inside and outside  $C$ , respectively. The idea is that  $E_d$  will measure the compatibility of local texture patterns in an image  $I$  with the texture of a proposed cartoon. The *spectral cartoon* of  $I$  is represented by the triple  $(\Phi_{in}, \Phi_{out}, C)$  that minimizes the posterior energy

$$E(\Phi_{in}, \Phi_{out}, C|I) = \alpha \int_{D_{in}} d_T^2(\Phi_p, \Phi_{in}) dp + \beta \int_{D_{out}} d_T^2(\Phi_p, \Phi_{out}) dp + \gamma \ell(C), \quad (9)$$

$\gamma > 0$ .

Since the estimation of the triple  $(\Phi_{in}, \Phi_{out}, C)$  may be a costly task, one may modify the model, as follows. For a given curve  $C$ , the optimal  $\Phi_{in}$  can be interpreted as the average value of  $\Phi_p$  in the region  $D_{in}$ , and the integral  $\int_{D_{in}} d_T^2(\Phi_p, \Phi_{in}) dp$  as the total variance of  $\Phi_p$  in the region. We propose to replace  $d^2(\Phi_p, \Phi_{in})$ , the distance square to the mean, with the average distance square from  $\Phi_p$  to  $\Phi_q$ , for  $q \in D_{in}, q \neq p$ , which is given by

$$\frac{1}{P_{in} - 1} \sum_{\substack{q \in D_{in} \\ q \neq p}} d^2(\Phi_p, \Phi_q).$$

Here,  $P_{in}$  is the number of pixels in  $D_{in}$ . Proceeding similarly for the region outside, the task is reduced to the simpler *maximum-a-posteriori* estimation of the curve  $C$ ; that is, the curve that minimizes the energy functional

$$E(C|I) = \frac{\alpha}{P_{in} - 1} \sum_{\substack{p, q \in D_{in} \\ q \neq p}} d^2(\Phi_p, \Phi_q) + \frac{\beta}{P_{out} - 1} \sum_{\substack{p, q \in D_{out} \\ q \neq p}} d^2(\Phi_p, \Phi_q) + \gamma \ell(C).$$

## Summary and Comments

We introduced new computational methods and strategies in non-parametric Fisher information geometry. A basic tool developed is an algorithm to calculate geodesics in information manifolds that allows us to address computational problems arising in the analysis of families of probability density functions such as clustering PDFs and the calculation of means and covariance of families of PDFs. To demonstrate its usefulness, the methodology developed was applied to various image analysis problems such as appearance-based recognition of imaged objects and image segmentation based on local texture patterns. A *spectral cartoon model* of images was proposed using a new representation of local texture patterns. More extensive testing of the methods introduced in the context of image analysis will be carried out in future work.

*Acknowledgment.* This work was supported in part by NSF grants CCF-0514743 and IIS-0307998, and ARO grant W911NF-04-01-0268.

## References

1. S. Amari, *Differential-geometrical methods of statistics*, Lecture Notes in Statistics, vol. 28, Springer, Berlin, 1985.
2. S. Amari and H. Nagaoka, *Methods of information geometry*, AMS and Oxford University Press, New York, 2000.
3. A. Blake and A. Zisserman, *Visual reconstruction*, The MIT Press, 1987.
4. T. Chan, J. Shen, and L. Vese, *Variational PDE models in image processing*, Notices Amer. Math. Soc. **50** (2003), 14–26.
5. A. P. Dawid, *Discussion of “defining a curvature of a statistical problem (with applications to second-order efficiency)” by B. Efron*, The Annals of Statistics **3** (1975), 1231–1234.
6. D. L. Donoho and A. G. Flesia, *Can recent innovations in harmonic analysis ‘explain’ key findings in natural image statistics?*, Network: Computation in Neural Systems **12** (2001), no. 3, 371–393.
7. I. L. Dryden and K. V. Mardia, *Statistical shape analysis*, John Wiley & Son, 1998.
8. S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **6** (1984), no. 6, 721–741.
9. T. Hoffmann and J. M. Buhmann, *Pairwise data clustering by deterministic annealing*, IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997), 1–14.
10. A. Hyvarinen, *Survey on independent component analysis*, Neural Computing Surveys **2** (1999), 94–128.
11. I. T. Jolliffe, *Principal component analysis*, Springer-Verlag, New York, 1986.
12. H. Karcher, *Riemann center of mass and mollifier smoothing*, Comm. Pure Appl. Math. **30** (1977), 509–541.
13. E. Klassen, A. Srivastava, W. Mio, and S. Joshi, *Analysis of planar shapes using geodesic paths on shape manifolds*, IEEE Trans. on Pattern Analysis and Machine Intelligence **26** (2004), 372–383.
14. X. Liu and L. Cheng, *Independent spectral representations of images for recognition*, J. Optical Soc. of America **20** (2003), no. 7.
15. D. Mumford, *The Bayesian rationale for energy functionals*, Geometry-Driven Diffusion in Computer Vision (B. Romeny, ed.), Kluwer Academic, 1994, pp. 141–153.
16. ———, *Elastica and computer vision*, pp. 491–506, Springer, New York, 1994.
17. D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math. **42** (1989), 577–685.
18. B. Pistone and C. Sempì, *An infinite-dimensional geometric structure on the space of all probability measures equivalent to a given one*, The Annals of Statistics **23** (1995), no. 5, 1543–1561.
19. J. Portilla and E. P. Simoncelli, *A parametric texture model based on joint statistics of complex wavelet coefficients*, International Journal of Computer Vision **40** (2000), no. 1, 49–70.
20. B. Schölkopf, A. Smola, and K. R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural Computation **10** (1998), 1299–1319.

21. A. Srivastava, S. Joshi, W. Mio, and X. Liu, *Statistical shape analysis: Clustering, learning and testing*, IEEE Trans. on Pattern Analysis and Machine Intelligence **27** (2005), 590–602.
22. A. Srivastava, X. Liu, and U. Grenander, *Universal analytical forms for modeling image probability*, IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2002), no. 9, 1200–1214.
23. Y. N. Wu, S. C. Zhu, and X. Liu, *Equivalence of Julesz ensembles and FRAME models*, International Journal of Computer Vision **38** (2000), no. 3, 247–265.
24. S. C. Zhu, Y. Wu, and D. Mumford, *Filters, random fields and maximum entropy (FRAME)*, International Journal of Computer Vision **27** (1998), 1–20.