J Math Imaging Vis (2009) 34: 270–290
DOI 10.1007/s10851-009-0146-1

Statistical M-Estimation and Consistency in Large Deformable Models for Image Warping

Jérémie Bigot · Sébastien Gadat · Jean-Michel Loubes

Published online: 20 March 2009

Abstract The problem of defining appropriate distances between shapes or images and modeling the variability of natural images by group transformations is at the heart of modern image analysis. A current trend is the study of probabilistic and statistical aspects of deformation models, and the development of consistent statistical procedure for the estimation of template images. In this paper, we consider a set of images randomly warped from a mean template which has to be recovered. For this, we define an appropriate statistical parametric model to generate random diffeomorphic deformations in two-dimensions. Then, we focus on the problem of estimating the mean pattern when the images are observed with noise. This problem is challenging both from a theoretical and a practical point of view. M-estimation theory enables us to build an estimator defined as a minimizer of a well-tailored empirical criterion. We prove the convergence of this estimator and propose a gradient descent algorithm to compute this M-estimator in practice. Simulations of template extraction and an application to image clustering and classification are also provided.

Keywords Image warping · Template extraction · Random diffeomorphism · Large deformable models · M-estimation · Asymptotic statistics · Clustering

1 Introduction

Image analysis and pattern recognition has been an increasing field of motivation in statistics over the last decade. One of the main difficulty comes from the choice of a proper definition for the model generating the images. Several methods have been investigated, each one dealing with a different point of view in statistics.

In practice, we always observe noisy images. The noise may be due either to the measurement devices or to the way images are generated, which makes their comparison difficult. One of the main difficulty in image analysis is the definition of a distance to compare the different observations. Several choices can be made and recently, originating in Grenander's pattern theory [17], new distances have been investigated. Such distances are based on the use of deformation groups to model the variability of natural images (see e.g. [15, 33, 34]).

In this paper we will mainly be concerned by the estimation of a mean template while observing similar noisy images. There are not so many results in the statistical literature dealing with the problem of building appropriate models to reflect the variability of natural images due to the presence of local deformations between them. A first attempt in this direction is the statistical framework based on penalized maximum likelihood proposed in [16] (see also the discussion therein) to approximate the mean of a set of images.

More recently, [1] have proposed a statistical model using Bayesian modeling and maximum likelihood estimation in the context of small parametric deformations. The approach proposed in [1] yields a consistent estimator of a mean of a set of images, and shows interesting classification performances. An extension of this work [4] uses a stochastic algorithm for approximating a maximum a posterior estimator. However, in all these non-rigid deformation approaches the transformations used to model the images variability are not constrained to be one-to-one, and therefore these approaches fail in generating diffeomorphic stochastic models. Note that

J. Bigot · S. Gadat (✉) · J.-M. Loubes
Institut de Mathématiques de Toulouse, Université de Toulouse, Toulouse, France
e-mail: sebastien.gadat@free.fr

a recent work [8] proposes also to use an infinitesimal gradient descent with respect to the Hausdorff topology to define the empirical mean and covariance of shapes but without giving any one-to-one matching between points of random shapes. Recently, statistical interpretation of the landmark matching problem with a random model for generating diffeomorphisms has been proposed in [28] and [29] but this approach has not been applied to image template estimation.

On the other hand, numerous works have been proposed to generate diffeomorphisms using flows governed by appropriate time-dependent vector fields (we refer to [22, 23, 30, 33] for further details). A current trend is the study of probabilistic and statistical aspects of deformation models, and the development of consistent statistical procedure for the estimation of template images. Some works in this direction [9, 40] have been recently published where the authors define probabilistic models of shapes or images that could be used to generate new data.

Our objective is therefore to combine powerful approaches for generating diffeomorphisms with an automatic statistical estimation of image mean and deformations. More precisely, our goal is to provide a statistical model to generate random images that yield new matching criterions to align a set of images.

For this, we define a general procedure to generate random diffeomorphic deformations, and we consider a statistical model for a set of images randomly warped from an unknown mean template. We then focus on the estimation of the mean pattern of these (possibly noisy) images. This problem is challenging both from a theoretical and a practical point of view. M-estimation theory (see e.g. [37]) enables us to build an estimator defined as a minimizer of a well-tailored empirical criterion. This generic method has been successfully applied in [20] and [6] to define the Fréchet mean of a set of curves or to describe central tendency of random curves. Fields of applications are numerous ranging from pattern recognition, brain atlas construction and computational anatomy to name but a few (see the various examples discussed in [15]).

Our contribution is the following. First we propose a new random diffeomorphic model for noisy images and we prove the convergence of our estimator to some mean pattern image when the number of observations (images) goes to infinity. Our estimator can be interpreted as the Fréchet mean of a set of images based on a distance involving diffeomorphic deformations. Consistency of Fréchet mean for curves and shapes has been investigated in [6] and [20], but to the best of our knowledge Fréchet mean for images using diffeomorphisms has not been investigated from a statistical point of view. We also present a new class of matching functionals that allows to easily incorporate penalization terms to control the amplitude of the estimated deformations and the amount of noise in the reconstructed mean pattern. A new gradient descent algorithm is finally proposed to minimize such functionals. This approach is also shown to be useful for clustering and classification problems in pattern recognition.

This article falls into the following parts. Section 2 deals with the definition of a new warping model. In Sect. 3, we state our statistical problem, and we study the asymptotic properties of various estimators of a mean pattern. In Sect. 4, we discuss some theoretical and practical aspects of our procedure, and we compare them with those of the Bayesian approach of [1]. Section 5 is devoted to the description of the algorithm needed to construct this estimate. Section 6 presents some experiments with simulated and real images. We also focus on clustering and classification problems to illustrate the usefulness of our methodology. We end the paper by a concluding section with a discussion on further developments of this work.

2 Model for Image Deformation

We start with discussing our random model of image deformation. Consider a two dimensional gray-level image as a real function defined on a compact set $\Omega \subset \mathbb{R}^2$. For sake of simplicity, we will set $\Omega = [0;1]^2$ and the generic notation for images will be $I : [0;1]^2 \to \mathbb{R}$. Assume moreover that I is a bounded function, which is not too restrictive since gray-level images typically take values between 0 and 255.

2.1 A Large Deformation Model with O.D.E.

Our goal is to generate a large enough deformation Φ to model the variability between observed images, but still being a diffeomorphism of $[0;1]^2$ in order to provide non ambiguous point displacements. These deformations will later be combined with a template $I^\star$ to generate a set of warped images, $I^\star \circ \Phi$. For this, we follow the approach proposed in [41] and [33].

Definition 1 (Diffeomorphism Φ_v^t) Let v be a smooth vector field from $[0;1]^2 \to \mathbb{R}^2$ vanishing on the boundary of this domain i.e.:

$$v_{|_{\partial[0;1]^2}} = 0. \tag{2.1}$$

Define a sequence of diffeomorphisms of $[0;1]^2$ denoted by $\{\Phi_v^t, t \in [0;1]\}$, as the solution of the following ordinary differential equation (O.D.E.):

$$\Phi_v^0(x) = x \quad \text{and} \quad \frac{d\Phi_v^t(x)}{dt} = v\left(\Phi_v^t(x)\right) \tag{2.2}$$

where t ranges over $[0;1]$ and $x \in [0;1]^2$.

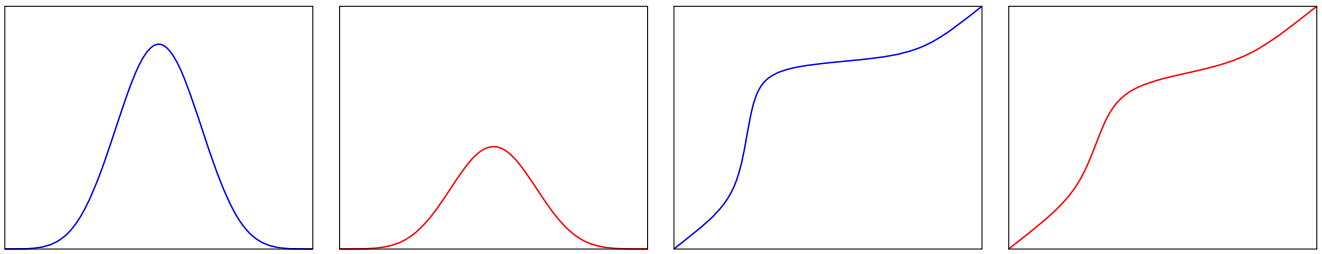

Fig. 1 A one-dimensional example of two vector fields with different amplitudes (*left images*) and corresponding diffeomorphisms at time $t = 1$ (*right images*)

As we want to have a deformation which remains in $[0;1]^2$, we have imposed that $\Phi^1_{v|_{\partial[0;1]^2}} = Id$, meaning that our diffeomorphism is the identity at the boundaries of $[0;1]^2$. Note that in the above definition, the vector field is not time dependent and in what follows, such vector fields will be called homogeneous. Moreover, as usual, by smooth we mean a C^∞ function.

The solution at time $t = 1$ denoted by Φ^1_v of the above O.D.E. is a diffeomorphic transformation of $[0;1]^2$ generated by the vector field v, which will be used to model image deformations. One can easily check (see [41]) that the vanishing conditions (2.1) on the vector field v imply that $\Phi^1_v([0;1]^2) = [0;1]^2$ and that Φ^t_v is a diffeomorphism for all time $t \in [0,1]$. Thus Φ^1_v is a convenient object to generate diffeomorphisms.

To illustrate the influence of the choice of the vector field v on the shape of the deformation Φ^1_v, we consider a simple example in one-dimension (i.e. for $v : [0,1] \to \mathbb{R}$ which generates a diffeomorphism of the interval $[0,1]$). In Fig. 1, we display two vector fields that have the same support on $[0,1]$ but different amplitudes, and we plot the corresponding deformation Φ^1_v. One can see that the amount of deformations (measured as the local distance between Φ^1_v and the identity) depends on the amplitude of the vector field. In the intervals where v is zero, then the deformation is locally equal to the identity. Hence, choosing compactly supported vector fields allows one to generate local deformations.

To generate random diffeomorphisms, we propose to use a parametric class of diffeomorphisms. Consider an integer K and some basis functions (not necessarily linearly independent) $e_k : \mathbb{R}^2 \to \mathbb{R}^2$ whose choice will be discussed later on. We then decompose the former vector field v on the set of functions $e_k = (e^1_k, e^2_k)$. The random deformations are generated as follows. Let (a^1_k, a^2_k), $k = 1, \ldots, K$ be random coefficients drawn independently from a distribution P_A with compact support included in $[A, A]$ for given real $A > 0$. Then, we define a random vector field v_a as

$$\forall x \in [0;1]^2 \quad v_a(x) = \begin{pmatrix} \sum_{k=1}^K a^1_k e^1_k(x) \\ \sum_{k=1}^K a^2_k e^2_k(x) \end{pmatrix}. \tag{2.3}$$

Finally, one has just to run the previously defined O.D.E. (2.2) to produce a random deformation, Φ_{v_a}.

2.1.1 Choice of Prior Distribution P_A

Choosing the prior distribution of the coefficients of the vector field v_a determines the corresponding deformation. For example, one can take for P_A the uniform distribution on $[-A, A]$ i.e. $a^i_k \sim \mathcal{U}_{[-A,A]}$, $i = 1, 2$. However, it should mentioned that P_A can be any distribution on $\mathbb{R}$ provided it has a compact support. The compact support assumption for P is mainly used to simplify the proof for the consistency of our estimator. Hence, the parameter A can be a viewed as an a priori on the size of the deformations, and be considered as a kind of regularizing parameter. More discussion on the role on the parameter A and other regularizing parameters to control the amplitude of deformations is deferred to Sect. 4.

2.1.2 Choice of Basis Functions e_k

In order to get a smooth bijection of $[0;1]^2$, the e_k should be at least differentiable. Such functions are built as follows. First, we choose a set of one-dimensional B-splines functions (of degree at least 2) whose supports are included in $[0;1]$. To form two-dimensional B-splines, the common way is to use tensor products for each dimension. Recall that to define B-splines, one has to fix a set of control points and to define their degree. Further details are provided in [11] and we will fix these parameters in the section dealing with experiments.

We use B-splines functions because they are compactly supported with a local effect on the knots positions (see [11] for instance). This local influence is very useful for some problems in image warping where the deformation must be the identity on large parts of the images together with a very local and sharp effect at some other locations. The choice of the knots and the B-spline functions allows one to control the support of the vector field and therefore to define a priori the areas of the images that should be transformed.

In Fig. 2 we display an example of a basis $e^1_k = e^2_k$, $k = 1, \ldots, K$ for vector fields generated by the tensor product of two one-dimensional B-splines (hence $K = 4$). An example of deformation of the classical Lena image is shown

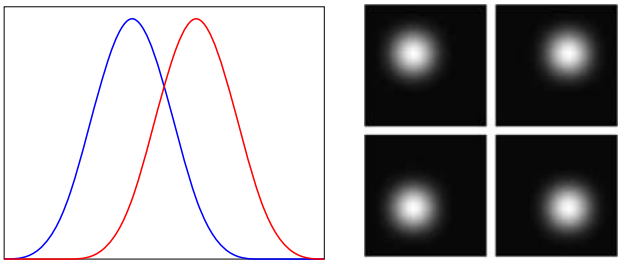

Fig. 2 *Left*: two 1D B-splines./ *Right*: corresponding basis $e_k^1 : [0,1]^2 \to \mathbb{R}$, $k = 1, \ldots, 4$ generated by tensor products of two 1D B-splines

Fig. 3 Random deformation of the Lena image with $A = 0.1$ and $A = 0.5$

in Fig. 3 with two different sets of coefficients a_k sampled from a uniform distribution on $[-A, A]$ (corresponding to different values for the amplitude A, a small and a large one). The amount of deformation depends on the amplitude of A, while the choice of the B-spline functions allows one to localize the deformation.

2.2 Random Image Warping Model with Additive Noise

Given a discretization of $[0;1]^2$ as a $N_1 \times N_2$ square grid of $N = N_1 N_2$ pixels, we will generically denote a pixel position by p. Once the deformation by random parametric diffeomorphisms with the O.D.E. method are generated, we can define the general warping model by:

Definition 2 (Noisy Random Deformation of Image) Fix an integer K and a real $A > 0$, we define a noisy random deformation of the mean template $I^\star$ as

$$I_{\varepsilon,a}(p) = I^\star \circ \Phi^1_{v_a}(p) + \varepsilon(p), \quad p \in [0,1]^2,$$

where $a \sim P_A^{\otimes 2K}$ and ε is an additive noise independent from the coefficients a. The new image $I_{\varepsilon,a}$ is generated by deforming the template $I^\star$ (using the composition rule $\circ$) and by adding a white noise at each pixel of the image.

In our theoretical approach, we consider the pixels p as a discretization of the set $[0;1]^2$ since our applications will be set up in this framework. It is often the case in the statistical literature on image analysis. However, our model could be formulated in a continuous setting using the continuous white noise model and a decomposition of the images in a wavelet basis as described in Sect. 3.3. This model involves the use of an integration measure over $[0;1]^2$ instead of sums over the pixels p of the image, see e.g. [7] for further details. Finally, remark that the image $I^\star$ is considered as a function of the whole square $[0;1]^2$, giving sense to $I^\star(\Phi^1_u(x))$.

In what follows, we denote by $\Phi_a(p) = \Phi^1_{v_a}(p)$ the solution of the following equation (starting from pixel p at time $t = 0$)

$$\forall p \in [0;1]^2 \quad \Phi^1_{v_a}(p) = p + \int_0^1 v_a(\Phi^t_{v_a}(p))dt. \tag{2.4}$$

Using this property, we consider now a set of n noisy images that are random deformations of the same unknown template $I^\star$ as follows:

$$I_{a^i,\varepsilon^i}(p) = I^\star \circ \Phi^1_{a^i}(p) + \varepsilon^i(p), \quad i = 1, \ldots, n, \tag{2.5}$$

where ε^i are i.i.d. unknown observation noise and a^i are i.i.d unknown coefficients sampled as $P_A^{\otimes K \times n}$. Our goal is to estimate the mean template image $I^\star$.

2.3 Mathematical Assumptions

For our theoretical study, we will need some mathematical assumptions:

A1 There exists a constant C such that

$$|\varepsilon| < C.$$

A2 $I^\star$ is L-Lipschitz.

Assumption A1 means that the level of noise is bounded which seems reasonable since we generally observe gray-level images which take values on a finite discrete set. Assumption A2 is more questionable. Indeed it implies that $I^\star$ is continuous, which seems impossible for natural models of images with structural discontinuities (think of the space of bounded variation (BV) functions for instance). However,

Fig. 4 Naive mean (*right image*) of a set of 10 images (mnist database, 28 × 28 pixels images, see [24] for more details on this data set)

one can view $I^\star$ as a map from all points in $[0;1]^2$ rather than just a function defined on the pixels. On $[0;1]^2$, it is more likely to suppose that $I^\star$ is the result of the convolution of $\mathcal{C}^\infty$-filters with captors measurements, which yields a smooth differentiable map on $[0;1]^2$. We refer to [13] for further comments on this assumption.

3 Statistical Estimation of a Mean Pattern

Consider a set of n noisy images $I_1, \dots I_n$. Assume first that these images are independent realizations from the model (2.5). We aim at constructing an estimate of the reference image $I^\star$. Without any convex structure on the images, averaging directly the observations is likely to blur the n images without yielding a sharp "mean shape". Indeed, computing the arithmetic mean of a set of images to estimate the mean pattern does not make sense as the space of deformed images $I^* \circ \Phi_v^1$ and the space of diffeomorphisms are not vectorial spaces, as shown in Fig. 4. To have a consistent estimation of $I^\star$, one needs to solve an inverse problem as stated in [6] and [20] derived from the random deformable model (2.4).

In our framework, estimating the pattern $I^\star$ involves finding a best image that minimizes an energy for the best transformation which aligns the observations onto the candidate. So, following [37], we will therefore define an estimator of $I^\star$ as a minimum of an empirical contrast function F_n (based on the observations $I_1, \dots, I_n$) which converges, under mild assumptions, toward a minimum of some contrast F.

3.1 A New Contrast Function for Estimating a Mean Pattern

Definition 3 (Contrast Function) Denote by $\mathcal{Z} = \{Z : [0,1]^2 \to \mathbb{R}\}$ a set of images uniformly bounded (e.g. by the maximum gray-level). Note that $\mathcal{Z}$ does not need to contain the true image $I^\star$. Assume also that $\mathcal{Z}$ is compact for the supremum norm on $[0;1]^2$. Then, define $\mathcal{V}_A$ as the set of vector fields given by (2.3). An element v_a in $\mathcal{V}$ can thus be written as

$$v_a = \left(\sum_{k=1}^{K} a_k^1 e_k^1, \sum_{k=1}^{K} a_k^2 e_k^2 \right), \quad \text{for some } a_k^i \in [-A, A].$$

Recall that N is the number of pixels. For an image $Z \in \mathcal{Z}$, a vector field $v_a \in \mathcal{V}_A$, and a given reference image $I^\star$, we define the following function f as

$$f(a, \varepsilon, Z) = \min_{v \in \mathcal{V}_A} \sum_{p=1}^{N} \left(I_{a,\varepsilon}(p) - Z \circ \Phi_v^1(p) \right)^2. \tag{3.1}$$

Thus f measures the cost of optimally aligning the image Z onto the image $I_{a,\varepsilon}$ using a diffeomorphic transformation. Note that this minimum is computed over a finite set of bounded coefficients $[-A;A]^{2K}$. Moreover, one can prove using [42] that this energy is a continuous function of v and thus of the set of coefficients $(a_k^i)_{1 \leqslant k \leqslant K; 1 \leqslant i \leqslant 2}$. This minimum is therefore reached at some $v_a \in \mathcal{V}_A$. For sake of simplicity, we introduce a notation that corresponds to a discretized norm over the pixels:

$$\left| I_{a,\varepsilon} - Z \circ \Phi_v^1 \right|_{\mathcal{P}}^2 = \sum_{p=1}^{N} \left(I_{a,\varepsilon}(p) - Z \circ \Phi_v^1(p) \right)^2.$$

At last, we define the mean contrast function F given by

$$F(Z) = \int_{[-A;A]^{2K} \times \mathbb{R}^N} f(a, \varepsilon, Z) dP(a, \varepsilon)$$

where $dP(a, \varepsilon)$ is the product measure on a and ε.

The interpretation of $F(Z)$ is the following: it measures "on average" how far an image Z is from the image $I_{a,\varepsilon}$ generated from our random warping model using an optimal alignment of Z onto $I_{a,\varepsilon}$. Our goal is to estimate a mean pattern image $Z^\star$ (possibly not unique) which corresponds to the minimum of the contrast function F when $I^\star$ is unknown.

Note that we only observe realizations $I_1, \dots, I_n$ that have been generated with the parameters $a^1, \dots, a^n$ and $\varepsilon^1, \dots, \varepsilon^n$. To estimate $Z^\star$, it is therefore natural to define the following empirical mean contrast:

Definition 4 (Empirical Mean Contrast) We define the measure $\mathbf{P}_n$ and the empirical contrast F_n as

$$\mathbf{P}_n(a, \varepsilon) = \frac{1}{n} \sum_{i=1}^{n} \delta_{a^i, \varepsilon^i} \quad \text{and}$$

$$F_n(Z) = \int f(a, \varepsilon, Z) d\mathbf{P}_n(a, \varepsilon).$$

Note that even if we do not observe the deformation parameters a^i and the noise ε^i, it is nevertheless possible to optimize $F_n(Z)$ with respect to Z since it can be written as:

$$F_n(Z) = \frac{1}{n}\sum_{i=1}^{n} \min_{v_i \in \mathcal{V}_A} \left| I_i - Z \circ \Phi_{v_i}^1 \right|_{\mathcal{P}}^2. \tag{3.2}$$

Note that the expression $|I - Z \circ \Phi_v^1|_{\mathcal{P}}$ does not define a distance between images I and Z since obviously $|I - Z \circ \Phi_v^1|_{\mathcal{P}} = 0$ can occur even if $I \neq Z$. Moreover, this expression is not symmetric in I and Z.

Moreover, note that in the above equation *it is not required* to specify the law P_A or the law of the additive noise to compute the criterion $F_n(Z)$. We then introduce quite naturally a sequence of sets of estimators

$$\hat{Q}_n = \arg\min_{Z \in \mathcal{Z}} F_n(Z) \tag{3.3}$$

and we will theoretically compare the asymptotic behavior of these sets with the deterministic one

$$Q_0 = \arg\min_{Z \in \mathcal{Z}} F(Z). \tag{3.4}$$

Remark that both sets $\hat{Q}_n$ and Q_0 are not necessarily restricted to a singleton, but these sets are obviously not invariant with respect to any smooth deformation Φ_v^1 since the way we generate diffeomorphisms does not provide any group structure. Consequently, if $Z \in Q_0$, it is not clear whether $Z \circ \Phi_v^1$ is in Q_0 or not. However, for any generated deformation Φ_v^1, there exists some other vector field v' such that $\Phi_v^1 \circ \Phi_{v'}^1$ is closed to the identity provided the basis used to generate the deformation is reach enough. Hence, even if for any $Z \in Q_0$ and any vector field v, $Z \circ \Phi_v^1$ does not belong necessary to Q_0, probably it is possible to find some other v_a such that $Z \circ \Phi_{v_a}^1$ is closed enough to Q_0. This uniqueness issues disappear by the addition of a regularization term on the norm of the diffeomorphism as it is done in Sect. 3.3.

3.2 Convergence of the Estimator

The following theorem gives sufficient conditions to ensure the convergence of the M-estimator in the sense of Theorem 1. The proof is deferred to Appendix A.

Theorem 1 *Assume that conditions* A1 *and* A2 *hold, then*

$$\hat{Q}_\infty \subset Q_0 \quad a.s.,$$

where $\hat{Q}_\infty$ is defined as the set of accumulation points of the $\hat{Z}_n$, i.e. the limits of convergent subsequences $\hat{Z}_{n_k}$ of minimizers $\hat{Z}_n \in \hat{Q}_n$.

This theorem ensures that the M-estimator, when constrained to lie in a fixed compact set of images, converges to a minimizer $Z^\star$ of the limit contrast function $F(Z)$. It seems therefore natural to ask how one chooses the compact set $\mathcal{Z}$ in practice, and also to determine the relationship between $Z^\star$ and the mean pattern $I^\star$. These problems will be discussed in the next sections.

Remark that Theorem 1 only proves the consistency of our estimator when the observed images comes from the true distribution (2.4). This assumption is obviously quite unrealistic, since in practice the observed images generally come from a distribution that is different from the model (2.5). In Sect. 3.3, we therefore address the problem of studying the consistency of our procedure when the observed images $I_i, i = 1, \ldots, n$ are an i.i.d. sample from an unknown distribution on $\mathbb{R}^N$ (see Theorem 2).

3.3 Penalization through Basis Expansions

The first M-estimator (3.3) minimizes a rough criterion, hence the minimum Z^* may be very different from the original image $I^\star$, leading to very poor estimate. This behavior is well known in statistics, see for instance [36], and the empirical mean contrast (3.2) has often to be balanced by a penalty which regularizes the matching criterion. In a Bayesian framework, it is well known that this penalized point of view can be interpreted as a special choice of a prior distributions. In nonparametric statistics, this regularization often takes the form of a penalized criterion which enforces the estimator to belong to a specific space satisfying appropriate regularity conditions. In our setting one needs to control both the smoothness of the estimated mean pattern and the amount of deformation allowed to align a set of images.

3.3.1 Penalization on the Deformations

To impose regularity on the deformations, we propose to add a penalty term to the matching criterion to exclude unlikely large warping (see e.g. [2]). For this, let Γ a symmetric positive definite matrix, and define

$$\mathrm{pen}_1(v) = \sum_{i=1}^{2}\sum_{k,k'=1}^{K} a_k^i \Gamma_{k,k'} a_{k'}^i.$$

This choice for pen_1 means that one can incorporate spatial dependencies through the use of the matrix Γ. Choosing such a penalty function implies that we do not assume anymore that all deformations have the same weight, as done in the original definition of $F_n(Z)$.

3.3.2 Penalization on the Images

To control the smoothness of the mean pattern, we have chosen to expand the images $Z \in \mathcal{Z}$ into a set of wavelet basis functions $(\psi_\lambda)_{\lambda \in \Lambda}$, since these functions are well suited

for image processing (see e.g. [27])). Here, the set Λ can be finite or not. This means that any image Z can be written as $Z = Z_\theta = \sum_{\lambda\in\Lambda} \theta_\lambda \psi_\lambda$, where the θ_λ's are the coefficients of Z in the wavelet basis. Estimating a noisy image expanded in a wavelet basis is generally done via an appropriate thresholding of its wavelet coefficient, and it is well known (see [3, 25]) that soft-thresholding estimator correspond to the use of the following penalty function on the θ_λ's

$$\text{pen}_2(\theta) = \sum_{\lambda\in\Lambda} |\theta_\lambda|.$$

Soft-thresholding estimators enable to incorporate some sparsity constraint on the set $\mathcal{Z}$ and have good properties for image smoothing. We could have chosen to follow some decomposition in some reproducing kernel Hilbert space with a finite set of control points as in [1]. But to the best of our knowledge, the effect of penalization in RKHS with a quadratic penalty is not really well suited to image analysis, whereas soft-thresholding methods have been shown to produce sparse representation of an image in a wavelet basis and have thus extremely good approximation and statistical properties (see e.g. [27]).

Note that other choices of penalty can be studied for practical applications. In what follows, we provide a general consistency result that is stated for general penalties. Let λ_1 and λ_2 be two smoothing parameters that we use to balance the contribution of the empirical mean contrast (3.2) and the penalties. Then, define the following penalized estimator $\hat{Z}_n = \sum_{\lambda\in\Lambda} \hat{\theta}_\lambda \psi_\lambda$, with

$$\hat{\theta}_n \in \arg\min_{\theta\in\mathbb{R}^\Lambda} \frac{1}{n}\sum_{i=1}^{n} \min_{v_i\in\mathcal{V}_A} (|I_i - Z_\theta \circ \Phi^1_{v_i}|^2_{\mathcal{P}} + \lambda_1 \text{pen}_1(v_i)) + \lambda_2 \text{pen}_2(\theta). \tag{3.5}$$

The above minimum may not be unique. However, some special conditions on λ_1, λ_2 and Λ could ensure uniqueness of $\hat{\theta}_n$ but studying such issue is beyond the scope of this paper.

Note that high values of λ_1 and λ_2 impose further regularity constraints on the mean pattern and the deformations. The numerical advantages of incorporating such penalization terms are studied in Sect. 6.3. The effects of adding such extra terms can also be studied from a theoretical point of view. If the smoothing parameters λ_1 and λ_2 are held fixed (they do not depend on n) then it is possible to study the converge of $\hat{\theta}_n$ as n grows to infinity under appropriate conditions on the penalty terms and the set Λ.

More precisely, we address now the problem of studying the consistency of our M-estimator when the observed images (viewed as random vectors in $\mathbb{R}^N$) come from an *unknown distribution* P, that does not necessarily correspond to the model (2.5). For sake of simplicity we still use the notation f introduced in (3.1). However within a penalized framework with unknown P, the dependency on ε disappears, and f is now defined as

$$f(I, Z_\theta) = \min_{v\in\mathcal{V}_A} \left[\|I - Z_\theta \circ \Phi^1_v\|^2_{\mathcal{P}} + \lambda_1 \text{pen}_1(v)\right] + \lambda_2 \text{pen}_2(\theta), \tag{3.6}$$

where $\lambda_1, \lambda_2 \in \mathbb{R}^+$, $\text{pen}_1(v) := \text{pen}_1(a) : \mathbb{R}^{2K} \to \mathbb{R}^+$, and $\text{pen}_2(\theta) : \mathbb{R}^\Lambda \to \mathbb{R}^+$. For any θ that "parametrizes" the image Z_θ in the basis $(\psi_\lambda)_{\lambda\in\Lambda}$, let F denote the general contrast function

$$F(Z_\theta) = \int f(I, Z_\theta) dP(I), \tag{3.7}$$

and F_n the empirical one defined as

$$F_n(Z_\theta) = \frac{1}{n}\sum_{i=1}^{n} f(I_i, Z_\theta).$$

The following theorem, whose proof is deferred to Appendix A, provides sufficient conditions to ensure the consistency of our estimator in the simple case when $F(Z_\theta)$ has a unique minimum at $Z_{\theta^\star}$ for $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^\Lambda$ is a compact set, and Λ is finite.

Theorem 2 *Assume that Λ is finite, that the set of vector fields $v = v_a \in \mathcal{V}$ is indexed by parameters a which belong to a compact subset of $\mathbb{R}^{2K}$, that $a \mapsto \text{pen}_1(v_a)$ and $\theta \mapsto \text{pen}_2(\theta)$ are continuous. Moreover, assume that $F(Z_\theta)$ has a unique minimum at $Z_{\theta^\star}$ for $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^\Lambda$ is a compact set. Finally, assume that the basis $(\psi_\lambda)_{\lambda\in\Lambda}$ and the set Θ are such that there exists two positive constants M_1 and M_2 which satisfy for any $\theta \in \Theta$*

$$M_1 \sup_{\lambda\in\Lambda} |\theta_\lambda| \leqslant \sup_{x\in[0,1]^2} |Z_\theta(x)| \leqslant M_2 \sup_{\lambda\in\Lambda} |\theta_\lambda|. \tag{3.8}$$

Then, if P satisfies the following moment condition,

$$\int \|I\|^2_{\infty,N} dP(I) < \infty,$$

where $\|I\|_{\infty,N} = \max_{p=1,\ldots,N} |I(p)|$, the M-estimator defined by $\hat{Z}_n = Z_{\hat{\theta}_n}$ where

$$\hat{\theta}_n \in \arg\min_{\theta\in\Theta} F_n(Z_\theta)$$

is consistent for the supremum norm of functions defined on $[0, 1]^2$ i.e.

$$\lim_{n\to\infty} \|\hat{Z}_n - Z_{\theta^\star}\|_\infty = 0 \quad a.s.$$

Two remarks on the last theorem can be made. First, the hypothesis on the uniqueness assumption can be substituted assuming that the set of minimum of F does not have some accumulation point:

$$\exists \eta > 0 \ \forall \theta \quad \text{such that } \|\theta^\star - \theta\| < \eta,$$
$$\theta \neq \theta^\star \quad F(Z_{\theta^\star}) < F(Z_\theta).$$

Secondly, the hypothesis on the existence of M_1 and M_2 will be here rather trivial since we will decompose our images in some finite wavelet basis Λ.

4 Discussion

4.1 Comparison with a Bayesian Approach

We discuss here the differences and the similarities between our approach and the Bayesian model proposed in [1].

First, assume that we do not use a penalization term on the deformations and images (λ_1, λ_2 are set to 0). Then, an important question raised by our model is the problem of deciding if the true template $I^\star$, used to generate the observed images, belongs to the set of minimizers of the limit criterion $F(Z)$ i.e. if $I^\star \in Q_0$ where $Q_0 = \arg\min_{Z\in\mathcal{Z}} F(Z)$. Obviously, the set Q_0 depends both on the choice of the compact set $\mathcal{Z}$ of candidate images, and on the level of noise. Determining the distance between an image $Z^\star \in Q_0$ and the mean pattern $I^\star$ is rather difficult in the presence of additive noise. Thus, if we consider a simple model without additive noise, then our limit criterion becomes $F(Z) = \mathbb{E}_a \min_{v\in\mathcal{V}_A} |I_a - Z\circ\Phi_v^1|_{\mathcal{P}}^2$ where $I_a = I^\star \circ \Phi_{v_a}^1$. Therefore, if the set $\mathcal{Z}$ contains $I^\star$, then the set of global minima of $F(Z)$ is the "orbit of $I^\star$" with respect to the "action" of Φ_v^1. In this setting our procedure is consistent in the sense as the number of images grows to infinity then the estimated image is the mean pattern $I^\star$. Of course here, we do not have any group action since the composition $\Phi_{v_1}^1 \circ \Phi_{v_2}^1$ is not necessarily equal to some Φ_w^1. We thus use the "orbit" term to design all images I such that $I = I^\star \circ \Phi_v^1$.

Now, using penalization terms, the limit criterion becomes

$$F(Z_\theta) = \mathbb{E}_a \min_{v\in\mathcal{V}_A} \left| I_a - Z_\theta \circ \Phi_v^1 \right|_{\mathcal{P}}^2 + \lambda_1 \text{pen}_1(v) + \lambda_2 \text{pen}_2(\theta).$$

In this case, $I^\star$ is not guaranteed to be a minimizer of F but arguing as in Sect. 3.1, if the basis is rich enough, we believe that $\arg\min F$ is closed enough to $I^\star$.

The approach proposed in [1] can also be interpreted from the M-estimation point of view. Note that their proofs of consistency relies on Wald's theorem which is a classical technique to prove the convergence of M-estimators, see e.g. [37]. Their estimated mean template is obtained via the minimization of an empirical criterion $G_n(\theta)$ depending on an image $Z = Z_\theta = \sum_{b=1}^B \theta_b \psi_b$ that is decomposed into a set of basis functions $\psi_b, b = 1, \ldots, B : \mathbb{R}^2 \to \mathbb{R}$. It is shown that as n grows to infinity then $\arg\min_{\theta\in\Theta} G_n(\theta)$ converges to the set $\arg\min_{\theta\in\Theta} G(\theta)$ where $G(\theta)$ correspond to the limit of $G_n(\theta)$ and Θ is some compact set of parameters. However, their construction of the criterion $G(\theta)$ and $G_n(\theta)$ is derived through Bayesian arguments, which therefore leads to different matching functionals. More precisely, in our notations their Bayesian model is the following

$$I(p) = I^*(p - u_\beta(p)) + \sigma\epsilon(p), \quad p = 1, \ldots, N, \tag{4.1}$$

where $\epsilon(p) \sim_{i.i.d.} N(0,1)$, $I^*(p) = \sum_{b=1}^B \theta_b^* \psi_b(p)$, and u_β is a deformation field parametrized by set of coefficients β. If a Gaussian prior is set on $\beta \sim N(0,\Gamma)$ (which yields random deformations), then [1] propose to estimate the coefficients $\theta^\star$ via maximization of the incomplete likelihood (for simplicity we assume hereafter that Γ and σ are known):

$$q(I|\theta) \propto \int e^{-\frac{1}{2}|I - Z_{\theta,\beta}|_{\mathcal{P}}^2 - \frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\beta^t\Gamma^{-1}\beta} d\beta, \tag{4.2}$$

where $Z_{\theta,\beta}(p) = \sum_{b=1}^B \theta_b \psi_b(p - u_\beta(p))$ for each pixel p. This yields the following MAP estimator

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} G_n(\theta) = \arg\min_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^n \log q(I_i|\theta)$$

and their limit criterion is thus of the form

$$G(\theta) = -\mathbb{E}\log q(I|\theta),$$

where the expectation is taken over random image I following the model (4.1). They also consider the case where the observed images follows another distribution P which is not necessarily the one induced by (4.1), and they study the consistency of their M-estimator in this case.

Explicit computation of $q(I|\theta)$ requires an integration over the hidden variables β which can be done numerically via an EM algorithm, but no analytical formula of this integral is available. Moreover, a natural question is to ask whether the true parameter $\theta^\star$ used to generate the observed images is a minimizer of $G(\theta)$. This problem still remains an open issue since such minimizers depend on $\theta^\star$ in a complicated way, through the law of the noise and the deformation. Note that this problem is also not solved in [1] or [4] since their consistency theorems only assert that $\hat{\theta}_n$ converges to a minimizer of $G(\theta)$.

However, following the arguments in Appendix B of [1], one can approximate the integral (4.2) by

$$\log q(I|\theta) \approx U(\beta^*), \tag{4.3}$$

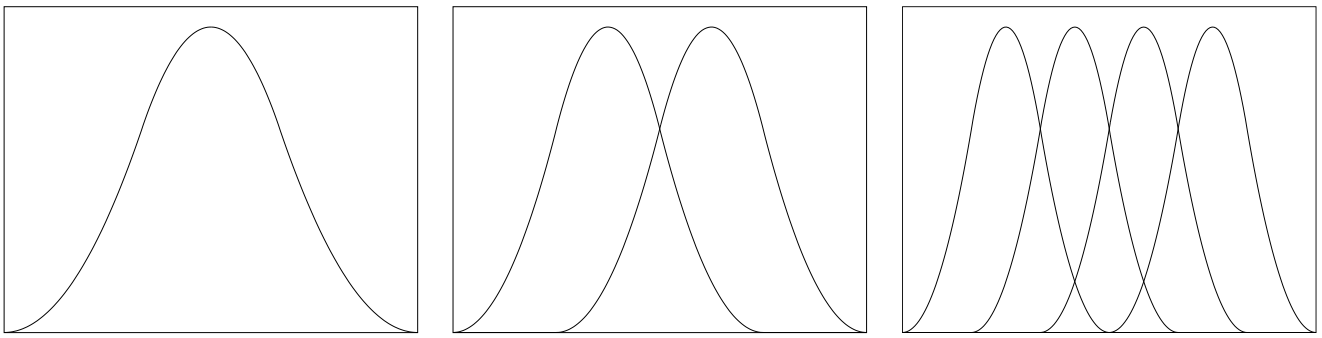

Fig. 5 An example of multiscale B-splines $\phi_{j,\ell}$, $\ell = 0, \ldots, 2^j - 1$ with $J = 3$ and $s = 3$, ordered *left to right*, $j = 0, 1, 2$

where $U(\beta) = -\frac{1}{2}|I - Z_{\theta,\beta}|^2_{\mathcal{P}} - \frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\beta^t\Gamma^{-1}\beta$ and $\beta^* = \arg\min U(\beta)$. Therefore, using the above approximation and if we eliminate the terms not depending on θ and β, then

$$\hat{\theta}_n \approx \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \min_{\beta_i} \left(|I_i - Z_{\theta,\beta_i}|^2_{\mathcal{P}} + \beta^t\Gamma^{-1}\beta \right)$$

and the limit criterion is therefore of the form:

$$G(\theta) \approx \mathbb{E} \min_{\beta} \left(|I - Z_{\theta,\beta}|^2_{\mathcal{P}} + \beta^t\Gamma^{-1}\beta \right),$$

where again the expectation is taken over a random image I following some distribution P. Hence, using a first order approximation for the integration over the hidden variable β, $G(\theta)$ is exactly our matching criterion $F(Z)$ (if the image Z is decomposed into some set of basis functions), with an additional penalty $\beta^t\Gamma^{-1}\beta$ on the parameters controlling the deformation. These arguments illustrate the classical interpretation of MAP estimate as a penalized likelihood estimator for suitable choices of the a priori distributions. Again, if we consider a simplest model with no additive noise and do not impose any penalization on the parameters of the deformation, then $\theta^\star \in \arg\min G(\theta)$. However, if one keeps the penalization term $\beta^t\Gamma^{-1}\beta$, then in the absence of noise there is no reason to believe that $\theta^\star \in \arg\min G(\theta)$ since the minimizers of $G(\theta)$ depends on the balance between image alignment and the amount of deformation.

4.2 Choice of the Basis Functions for the Vector Field and the Regularizing Parameter λ_1 and λ_2

Our estimation procedure obviously depends on the choice of the basis functions $e_k = (e_k^1, e_k^2)$ that generate the vector fields. In our simulations, we have chosen to use tensor products of one-dimensional B-spline organized in a multiscale fashion. Let s be some integer that represents a given order of the B-spline and, let $J \geqslant 1$ be some positive integer. For each scale $j = 0, \ldots, J-1$, we denote by $\phi_{j,\ell}$, $\ell = 0, \ldots, 2^j - 1$ the 2^j the B-spline functions obtained by taking $2^j + s$ knots points equispaced on $[0, 1]$ (see [11]). This gives a set of functions organized in a multiscale fashion, and in our numerical experiments we took $s = 3$ and $J = 3$ as shown in Fig. 5. Note that as j increases the support of the B-spline decreases which makes them more localized.

For $j = 0, \ldots, J-1$, we then generate a multiscale basis $\phi_{j,\ell_1,\ell_2} : [0, 1]^2 \to \mathbb{R}$, $\ell_1, \ell_2 = 0, \ldots, J-1$ by taking tensor products the $\phi_{j,\ell}$'s i.e.

$$\phi_{j,\ell_1,\ell_2}(x_1, x_2) = \phi_{j,\ell_1}(x_1)\phi_{j,\ell_2}(x_2).$$

Then, we take $e_k = e_{j,\ell_1,\ell_2} = (\phi_{j,\ell_1,\ell_2}, \phi_{j,\ell_1,\ell_2}) : [0, 1]^2 \to \mathbb{R}^2$. This makes a total of $K = \sum_{j=0}^{J-1} 2^{2j} = \frac{2^{2J}-1}{3}$ basis functions.

The assumptions of Theorem 2 impose that the coefficients used to compute the vector field belong to a compact subset of $\mathbb{R}^{2K}$, and this is mainly made to simplify the proof of the theorem. One could choose to control the amplitude of the deformations by controlling the size of this compact set which would then be a way to incorporate some regularization. However, we prefer to leave the size of this set very large (in practice we do not use any size constraint), and the amplitude of the deformations is rather control by the penalty term $\lambda_1 \text{pen}_1(v)$ in (3.6). The parameters λ_1 can be used to prevent huge or not-very-smooth deformations when searching for an optimal matching. Finding a data-based choice for λ_1 is a challenge and to the best of our knowledge there does not exist an automatic method for choosing such regularizing parameter in image warping problems, but we plan to study this in a future work. Instead, we provide in our simulations various examples illustrating the influence of this parameter (see Sect. 6).

For the choice of λ_2, we took the so-called universal threshold (see e.g. [3])

$$\lambda_2 = 2\sigma\sqrt{2 * \log(N)},$$

where σ denotes some estimation of the standard deviation of the additive noise and N is the number of pixels. Univer-

sal thresholding is a standard choice in image denoising that has good theoretical and numerical properties, and σ can be easily derived from the wavelet coefficients of a noisy image at high frequencies resolution (see [27] for further details).

4.3 Further Refinements of the Model

Our matching criterion to compare the alignment of two images is based on the sum of the square difference between the pixels of the images, which corresponds somehow to a Gaussian prior for the additive noise ϵ. However, one can use other matching criterion to compare images. Indeed one can check that it is possible to adapt our proofs of consistency of the M-estimators, if one replaces the discretized norm over the pixels:

$$\left|I_{a,\varepsilon} - Z \circ \Phi_v^1\right|_{\mathcal{P}}^2 = \sum_{p=1}^{N} \left(I_{a,\varepsilon}(p) - Z \circ \Phi_v^1(p)\right)^2$$

by any criterion of the form $L(I_{a,\varepsilon}, Z \circ \Phi_v^1)$ where $L : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^+$ is a real function which satisfies appropriate smoothness and convexity conditions.

Moreover, a set images may also present intensity variations, but our model does not take this into account. A nice extension for future investigation would be to incorporate an amplitude parameter in the estimation procedure to account for possible intensity variations between images.

5 Practical Computation of the M-Estimator

5.1 Algorithm for Mean Pattern Estimation

We describe an iterative procedure to compute the penalized M-estimator (3.5) . Given n images $I_1, \ldots, I_n$, recall that we have to find an image $\hat{Z}_n = \sum_{\lambda \in \Lambda} \hat{\theta}_\lambda \psi_\lambda$, with

$$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^\Lambda} \frac{1}{n} \sum_{i=1}^{n} \min_{v_i \in \mathcal{V}_A} \left(\left|I_i - Z_\theta \circ \Phi_{v_i}^1\right|_{\mathcal{P}}^2 + \lambda_1 \mathrm{pen}_1(v_i) \right) + \lambda_2 \mathrm{pen}_2(\theta).$$

In order to handle the two minimization steps, we use an alternative iterative procedure that works as follows:

Initialization $m = 0$: start with an initial guess $Z^{(0)}$. The choice of $Z^{(0)}$ is discussed in Sect. 5.3.

Iteration $m \geqslant 1$: repeat the following steps:

- for $i = 1, \ldots, n$, compute an optimal deformation $\Phi_{\hat{a}_i^m}$ which corresponds to the vector field

$$v_{\hat{a}_i^m} = \arg\min_{v_i \in \mathcal{V}} \left|I_i - Z^{(m-1)} \circ \Phi_{v_i}^1\right|_{\mathcal{P}}^2 + \lambda_1 \mathrm{pen}_1(v_i). \quad (5.1)$$

One may wonder how to compute such a minimum. In what follows, we will provide a gradient descent algorithm to solve this issue (see Sect. 5.2)

- Then, compute the image $\tilde{Z}^{(m)}$ that minimizes:

$$\tilde{Z}^{(m)} = \arg\min_{Z \in \mathcal{Z}} \underbrace{\sum_{i=1}^{n} \left|I_i - Z \circ \Phi_{\hat{a}_i^m}\right|_{\mathcal{P}}^2}_{:=E_m}.$$

If one does not constrained the images Z to belong to a specific set, then $\tilde{Z}^{(m)}$ can be easily found using a change of variable since it can be remarked that

$$E_m \simeq \sum_{i=1}^{n} \int_{[0;1]^2} \left(I_i - \tilde{Z}^{(m)} \circ \Phi_{\hat{a}_i^m}\right)^2 (x) dx.$$

The last approximation is due to the fact that E_m is computed for the discrete measure on the pixels of the image, and not exactly on the whole set $[0; 1]^2$. Changes of variables in the last n integrals by $u = \Phi_{\hat{a}_i^m}(x)$ yield the expression:

$$\begin{aligned} E_m &\simeq \sum_{i=1}^{n} \int_{[0;1]^2} \left(I_i \circ \Phi_{\hat{a}_i^m} - \tilde{Z}^{(m)}\right)^2 (u) \\ &\quad \times |\det Jac(\Phi_{\hat{a}_i^m}^{-1})(u)| du \\ &\simeq \int_{[0;1]^2} \sum_{i=1}^{n} \left(I_i \circ \Phi_{\hat{a}_i^m} - \tilde{Z}^{(m)}\right)^2 w_i(u) du. \end{aligned}$$

The solution of this least square problem is the classical weighted average using the coefficients w_i. The value of the solution $Z^{(m)}$ at any pixel p, is thus given by

$$\tilde{Z}^{(m)}(p) = \frac{\sum_{i=1}^{n} w_i(p) I_i \circ \Phi_{\hat{a}_i^m}^{-1}(p)}{\sum_{i=1}^{n} w_i(p)}, \quad (5.2)$$

where $w_i(p) = |\det Jac(\Phi_{\hat{a}_i^m}^{-1})(p)|$.

Then, apply wavelet soft thresholding with universal threshold to $\tilde{Z}^{(m)}$ to finally obtain a denoised image $Z^{(m)}$.

5.2 A New Matching Algorithm between Two Images

The minimization step (5.1) is a crucial point in the above described algorithm. It consists of finding an optimal deformation between two images using a specific parametrization of a set of vector fields. Below, we describe a gradient descent algorithm with an adaptive step to perform the minimization (5.1) which yields a new matching algorithm between two images.

To simplify the presentation, we took in our simulations the identity matrix for Γ in the formulation of pen_1. Remark that this choice does not take into account the presence of

correlations between the element of the spline basis. Another choice would be $\Gamma = G^{-1}$ where G is the Gram matrix with entries given by inner products of the spline basis function e_k^i. This choice would correspond to a uniform prior on deformations.

Given two images I and Z, one thus needs to optimize the following term

$$\Delta_{I,Z} = \left| I - Z \circ \Phi^1_{v_a} \right|^2_{\mathcal{P}} + \lambda_1 \sum_{i=1}^{2} \sum_{k}^{K} |a_k^i|^2$$

with respect to $a = (a_k^i)_{k,i}$, $k = 1 \dots K$ and $i \in \{1, 2\}$. In the above expression, v_a is given as (2.3). To implement a gradient descent algorithm, one needs to compute

$$\frac{\partial \Delta_{I,Z}}{\partial a_k^i} = -2 \sum_{p=1}^{N} [I(p) - Z(\Phi^1_{v_a}(p))] \times \left\langle \nabla Z_{\Phi^1_{v_a}(p)}; \frac{\partial \Phi^1_{v_a}(p)}{\partial a_k^i} \right\rangle + 2\lambda_1 a_k^i, \tag{5.3}$$

for all $k = 1, \dots, K$ and $i = 1, 2$. Now, suppose without loss of generality that $i = 1$. Then for any pixel p:

$$\frac{\partial \Phi^1_{v_a}(p)}{\partial a_k^1} = \frac{\partial [\int_0^1 v_a(\Phi^t_{v_a}(p))dt + p]}{\partial a_k^1} = \int_0^1 \begin{pmatrix} e_k^1(\Phi^t_{v_a}(p)) + \sum_{\alpha=1}^{K} a_\alpha^1 \langle \nabla e^1_{\alpha_{\Phi^t_{v_a}(p)}}, \frac{\partial \Phi^t_{v_a}(p)}{\partial a_k^1} \rangle \\ \sum_{\alpha=1}^{K} a_\alpha^2 \langle \nabla e^2_{\alpha_{\Phi^t_{v_a}(p)}}, \frac{\partial \Phi^t_{v_a}(p)}{\partial a_k^1} \rangle \end{pmatrix} dt.$$

As $\frac{\partial \Phi^0_{v_a}(p)}{\partial a_k^1}$ vanishes, $\psi_{k,1,1}(p) = \frac{\partial \Phi^1_{v_a}(p)}{\partial a_k^1}$ is solution at time $t = 1$ of the following O.D.E.:

$$\frac{d\psi_{k,1,t}(p)}{dt} = \begin{pmatrix} e_k^1(\Phi^t_{v_a}(p)) + \sum_{\alpha=1}^{K} a_\alpha^1 \langle \nabla e^1_{\alpha_{\Phi^t_{v_a}(p)}}, \psi_{k,1,t}(p) \rangle \\ \sum_{\alpha=1}^{K} a_\alpha^2 \langle \nabla e^2_{\alpha_{\Phi^t_{v_a}(p)}}, \psi_{k,1,t}(p) \rangle \end{pmatrix}$$

with initial condition $\psi_{k,1,0}(p) = 0$. To get a gradient descent algorithm, one uses the above O.D.E. to evaluate the gradient (5.3). The computation of the optimal choice of the a_k^i's follows from a classical gradient descent algorithm with an adaptive step starting from $(a_k^i)_{k,i} = 0$.

This gradient descent may fall into a local minima since our criterion may not be convex. However, our hierarchical choice for the splines described in Sect. 4.2 induces a kind of multi-scale framework which gives an algorithm that performs well in practice. At last, we have used the stopping criterion of [18] to end the gradient descent algorithm.

5.3 Initialization of the Algorithm

The simplest to initialize our iterative algorithm is to take the naive estimate $Z^{(0)}_{naive} = \frac{I_1 + \dots + I_n}{n}$. However, this may give a very poor preliminary estimator which may considerably affect the quality of the mean pattern.

Alternatively, we have implemented a new matching criteria proposed by [14, 39] to find rigid transformations between a set of curves. In our setting, this criteria is a global measure of how well a set of images are aligned and can be written as matching function $M_n : \mathcal{A}^n \to \mathbb{R}^+$ given by

$$M_n(a^1, \dots, a^n) = \frac{1}{n} \sum_{i=1}^{n} \left| I_i \circ \Phi^1_{v_a^i} - \frac{1}{n} \sum_{i'=1}^{n} I_{i'} \circ \Phi^1_{v_a^{i'}} \right|^2_{\mathcal{P}} + \lambda_1 \sum_{i=1}^{n} \|a^i\|^2_{\mathbb{R}^{2K}},$$

where $\mathcal{A}$ is a subset of $\mathbb{R}^{2K}$ used to parametrize the vector fields. The above criterion M_n is closely related to Procrustes analysis which is classically used for the statistical analysis of shapes (see e.g. [12]) and the registration of a set of curves onto a common target function. However, here the common target function is directly given by the average of the registered images given a possible choice of deformation parameters $a^1, \dots, a^n$. An initial image can then be defined by searching

$$(\hat{a}^1, \dots, \hat{a}^n) = \arg \min_{(a^1, \dots, a^n) \in \mathcal{A}^n} M_n(a^1, \dots, a^n)$$

and then by taking

$$Z^{(0)}_* = \frac{1}{n} \sum_{i=1}^{n} I_i \circ \Phi^1_{v_{\hat{a}^i}}. \tag{5.4}$$

Surprisingly, our simulations show that this initial estimator $Z^{(0)}_*$ which will be referred to as the direct mean, already gives very accurate results. Note that the gradient of the criterion M_n can be computed as described in Sect. 5.2, and thus we have again chosen to compute the coefficients $(\hat{a}^1, \dots, \hat{a}^n)$ via a gradient descent algorithm with an adaptive step.

5.4 Convergence of the Numerical Scheme

The approximation (4.3) is used in [1] to simplify the M-step in the EM-algorithm used to compute numerically the minimizer of the incomplete log-likelihood $G_n(\theta) = \sum_{i=1}^{n} \log q(I_i|\theta)$ (this is referred to as fast approximation

Fig. 6 Naive mean (*lower left image*), direct mean $Z_{\star}^{(0)}$ (*lower middle image*) and mean pattern $Z^{(3)}$ (*lower right image*) based on 20 images of the digit "2" (*upper rows*)

with modes in [1]). This simplification yields a similar iterative algorithm to the one used in this paper. However, the fast approximation with modes used in [1] does not guarantee to obtain an iterative scheme which converges to a minimizer of $G_n(\theta)$. To overcome this problem, a stochastic EM algorithm is proposed in [4] yielding an iterative procedure which is shown to converge to the true MAP estimator. In our approach, we also use an alternative scheme to find a minimizer of the empirical contrast function $F_n(Z)$, but this iterative procedure follows directly from the formulation of our criteria via a double minimization. As we do not use any approximation of the functional $F_n(Z)$ to derive this alternative scheme, we believe that the sequence of images $Z^{(m)}$ (see Sect. 5) is likely to give a good approximation of $\hat{Z}_n$ as m grows to infinity although this remains to be proved rigorously. Moreover, in the next section we discuss a new matching criterion to initialize our iterative algorithm which gives surprisingly good results.

6 Numerical Results

Recall that in all our simulations, we used the hierarchical basis with $K = \frac{2^{2J}-1}{3} = 21$ using $s = 3$ and $J = 3$ as described in Sect. 4.2.

6.1 A Real Example (Mnist Database)

First we return to the example shown previously on handwritten digits (mnist database). As these images are not very noisy, the denoising step via wavelet thresholding does not improve the results. A value of $\lambda_1 = 10$ gave good results but more discussion on the influence of this parameter can be found in the next section of faces averaging.

In Fig. 6, we display the naive mean $Z_{naive}^{(0)}$ and the direct mean $Z_{*}^{(0)}$ the obtained from $n = 20$ images of the digits "2". Surprisingly the result obtained with $Z_{*}^{(0)}$ is very satisfactory and is a better representative of the typical shape of the digits "2" in this database. In Fig. 6, the image $Z^{(3)}$ obtained after 3 iterations of the algorithm is also displayed with $Z^{(0)} = Z_{*}^{(0)}$. We wee that the iterations slightly improves the initial result. Moreover, note that $Z^{(3)}$ has sharper edges than the naive mean which is very blurred.

In Fig. 7 we finally display the comparison between the naive mean, the direct mean and the mean pattern $Z^{(3)}$ (initialized with $Z^{(0)} = Z_{*}^{(0)}$), for all digits between 0 and 9 with 20 images for each digit. One can see that our approach yields significant improvements. In particular it gives mean digits with sharp edges.

6.2 Influence of the Gradient Descent and the Initialization

In Fig. 7, the second and third rows are almost identical, which validates our initialization using the direct mean, see (5.4), but not the rest of the framework. Indeed, one may wonder if the iterative process by gradient descent does not get stuck into a local minima and if $Z^{(n)}$ is really better than the initialization $Z^{(0)}$. To validate our framework, we display in Fig. 8 an example of the improvements by the iterative process when starting from an initialization with the naive mean instead of the direct mean (5.4) for digits "8" and "9".

6.3 Influence of the Choice of λ_1 (Olivetti Database)

Influence of λ_1 We illustrate the role of the parameter λ_1 which controls the amount of deformation with a problem a faces alignment. Figure 9 represents two images of the same subject with varying lighting and facial expression. These images are taken from the Olivetti face database [31] and their size is $N_1 = 98$ and $N_2 = 112$. The results of the gradient descent algorithm with various values for λ_1 are given

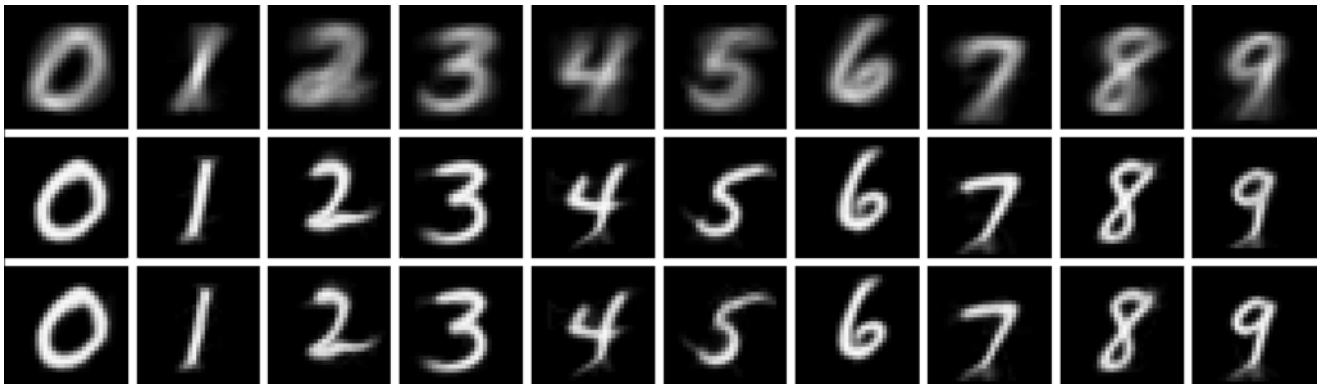

Fig. 7 Naive mean (*first row*), direct mean (*second row*) and mean pattern $Z^{(3)}$ (*last row*) based on 20 images on the mnist database

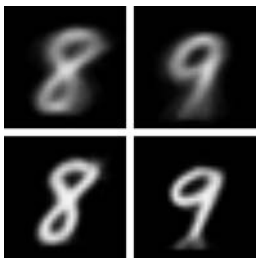

Fig. 8 *First row*: naive mean for digits "8" and "9", *second row*: $Z^{(5)}$ obtained by starting from an initialization $Z^{(0)}$ by the naive mean (*images of the first row*)

in the second row of Fig. 9. As expected large values of λ_1 yield small deformations while a small value allows much more flexible diffeomorphic warping.

Mean Images on Olivetti Database For each subject of the Olivetti database, $n = 9$ images have been taken with various facial expression. Figure 10 shows the faces used in our simulations.

In Fig. 11 we present some mean pattern obtained with an iterative algorithm with $Z^{(0)} = Z_*^{(0)}$, $\lambda_1 = 1000$, and compare them with the corresponding naive mean. Obviously our method clearly improves the naive estimate, and yields satisfactory average faces especially in the middle of the images. However, some parts along the image boundaries in the second row of Fig. 11 are still slightly blurred. This is due to the fact that the basis functions that we have chosen are vanishing along image boundaries (see Fig. 5). This can be improved by incorporating other basis functions to allow more flexible warping along image boundaries, but we prefer to leave this example to illustrate the influence of the choice of the basis functions.

6.4 A Simulated Example

In this section, we generate some simulated noisy images to judge the quality of the method when the true image to recover is known. The reference image I^* is the Shepp-Logan phantom image (see [21]) of size $N_1 \times N_2$ with $N_1 = N_2 = 128$ shown in Fig. 12. We have then simulated $n = 20$ noisy and randomly warped images from $I^\star$. However, the random deformations are generated via homogeneous vector fields that *are not expressed* in the basis e^k, $k = 1, \ldots, K$ to illustrate the robustness of the method via a kind of mis-specification of the model. These vector fields are generated by a finite linear combination of Gaussian kernels with random amplitudes and random locations following a uniform distribution on a subset of $[0; 1]^2$.

In Fig. 13, we display the direct mean $Z_*^{(0)}$ followed by wavelet thresholding obtained from these 20 images with various values of λ_1. Again, these initial estimates are very accurate estimate of the original template shown in Fig. 12. In this example running the iterative algorithm does not improve the results, and this can be explained by the fact the initial estimate is already very good. These simulated data tend thus to show that our method is also somehow robust to mis-specification of the model since we recall that the random vector fields used for the simulations have been not constructed from the multi-scale B-spline basis described previously.

6.5 Application to Image Clustering and Classification

Clustering We finally end this section on numerical experiments by showing an example of clustering using the k-means algorithm (see e.g. [26]). To cluster a set of images by the k-means algorithm one must choose a proper distance to compare images and a way of calculating the mean of a cluster. Given two images I_1 and I_2 we define a "distance" between them using diffeomorphic warping as follows (with $\lambda_1 = 10$):

$$d(I_1, I_2) = \min_{v_a \in \mathcal{V}} \left| I_1 \circ \Phi_{v_a}^1 - I_2 \right|_{\mathcal{P}}^2 + \lambda_1 \|a\|_{\mathbb{R}^{2K}}^2 .$$

Then, for a set images belonging to the same cluster, the mean is defined as $Z^{(4)}$ with initialization by direct mean. In Fig. 14, we give an example of k-means clustering with two classes for the digit "2" of the images of the training

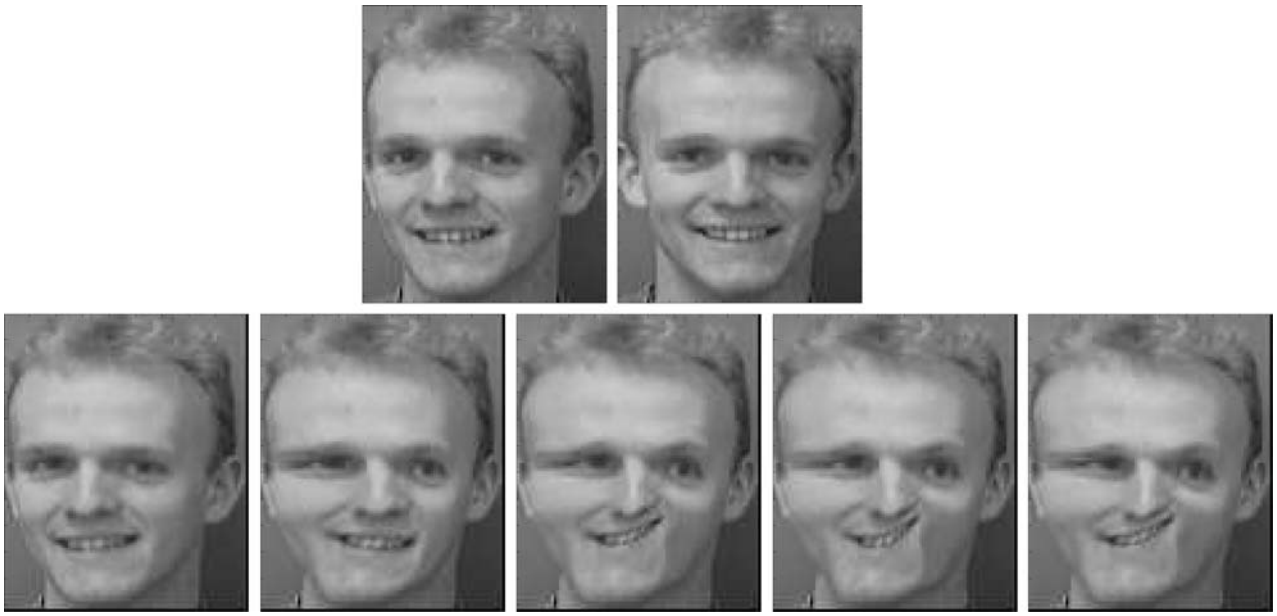

Fig. 9 *First row*: two images of the same subject taken from the Olivetti database of faces. *Second row*: warping of the *left image* onto *the right image* with (*from left to right*) varying values of $\lambda_1 = 10000, 1000, 100, 10, 1$

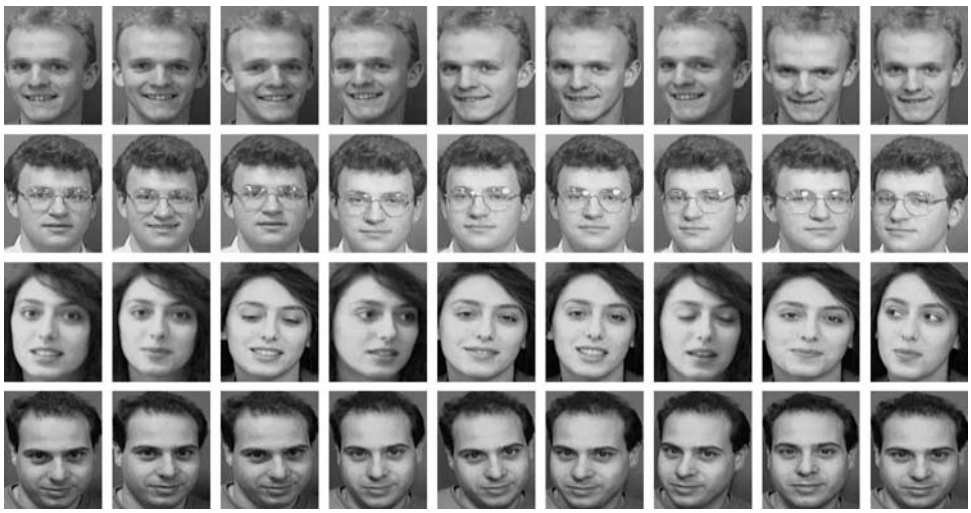

Fig. 10 9 samples of the Olivetti database for 4 subjects

set. One can see that the algorithm gives two different mean clusters $Z^{(m)}$ which correspond to digits "2" with or without a loop. Again the results are visually very good. Finally, we display in Figs. 15,16 and 17 the clusters for the images of the digit "2", "3" and "5" of the training set. In all Figures the upper left image is the mean $Z^{(4)}$ of the cluster. One can see that the images are classified according to their vertical orientation.

Classification Even if our goal is not to implement a new classification method for image recognition, one can easily adapt our method to reach an automatic supervised classification procedure. We consider the 10 classes of the Mnist database and we compute a clustering of two subsets of each class. On each cluster, the mean patterns are computed and we use them to classify images belonging to a test set consisting of 100 images of digits between 0 and 9 which makes on overall set of 1000 images. Then, a simple criterion based on the norm $|.|_{\mathcal{P}}$ is used to classify these data. The decision rule for any image I in the test set follows naturally from our minimization algorithm:

$$d(I) = \arg \min_{i=1\ldots10} \min_{v_a \in \mathcal{V}_A} \left| I \circ \Phi^1_{v_a} - \hat{I}_i \right|^2_{\mathcal{P}} + \lambda_1 \|a\|^2_{\mathbb{R}^{2K}}.$$

We use here $\lambda_1 = 10$ as it performed well in our simulations. Here, $d(I)$ denotes the predicted class for I in the test set. The computation of $d(I)$ simply consists in warping the image I to the closest image among $\hat{I}_1, \ldots, \hat{I}_q$. The

Fig. 11 Example of face averaging for 4 subjects from the Olivetti database. *First row*: naive mean, *second row*: mean pattern $Z^{(7)}$

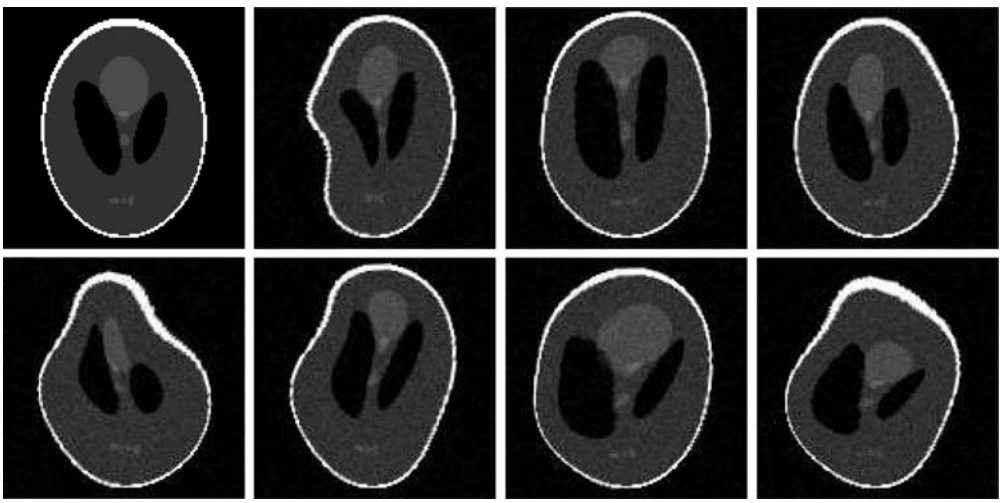

Fig. 12 Simulated example: seven deformed and noisy images of the Shepp-Logan phantom (out of a sample of 20 images). The *upper left image* is the unknown template $I^\star$

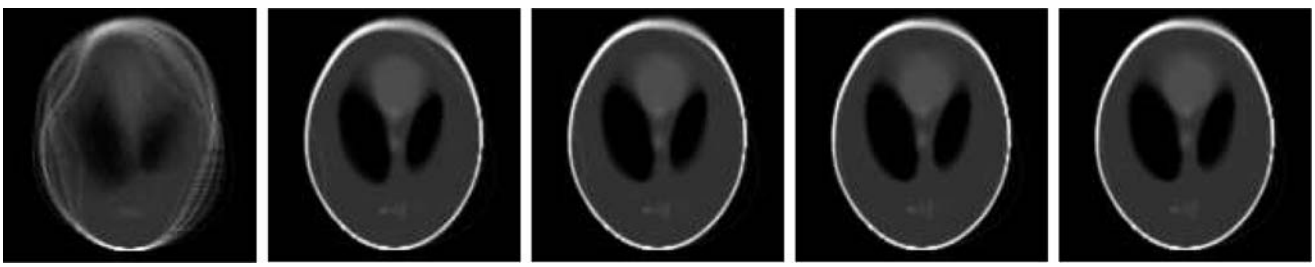

Fig. 13 Naive mean (*right image*), and direct mean $Z_*^{(0)}$ followed by wavelet thresholding with (*from left to right*) $\lambda_1 = 1000, 500, 100, 10$

rule $d(I)$ will be referred to as classification with warping in what follows.

The computational cost of the decision rule is low since the ten *mean images* $\hat{I}_i, i \in \{0, \ldots, 9\}$ of the ten classes are computed off-line with the training set. Indeed, computing the decision $d(I)$ is equivalent to run 10 matching algorithms with our gradient method.

To evaluate the performances of this classification rule, we have compared its misclassification rate with those of two other approaches:

- Naive classification: simply take the naive mean for each class as a typical representative of the images within a class. Then, for a new image I of the training set, take the following classification rule simply based on the norm $|.|_{\mathcal{P}}$ (without any warping)

$$d^{naive}(I) = \arg\min_{i=1\ldots q} \left| I - \hat{I}_i^{naive} \right|_{\mathcal{P}}^2 .$$

- Support vector machine (SVM) classification: we have a multi-class classification problem. Basically, SVM classifiers can only solve binary classification problems (see e.g. [32, 35]). To allow for multi-class classification, we have used the algorithm implemented in the R library `e1071` [10] that uses the one-against-one technique by fitting all binary subclassifiers and finding the correct class by a voting mechanism (see also [19] for gentle introduction to SVM classification). Note that in the case of SVM classification, the images are simply considered as vectors in $\mathbb{R}^N$ and that the spatial dependency of the pixels is thus not taken into account.

  The parameters of the SVM have an important influence on the accuracy of the prediction. They have been set as follows: we use a Gaussian kernel (RBF) as it performs generally better than polynomial kernels. The several parameters (margin parameter C and variance parameter σ^2) has been set using a tuning step of cross validation to obtain the best performance as possible. This

Fig. 14 K-means clustering for the 20 images for the class of digit 2 of the training set

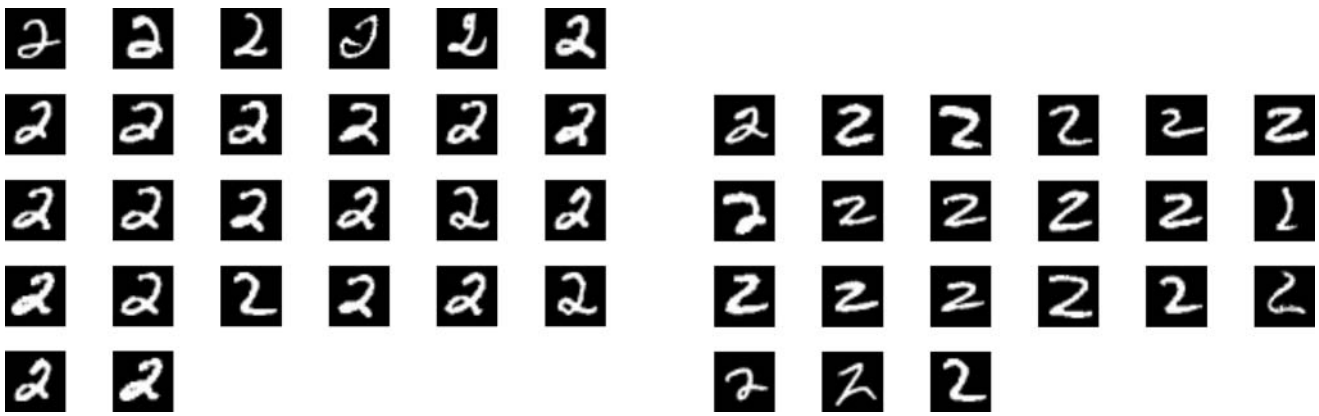

Fig. 15 Two clusters obtained by K-means clustering for the 20 images for the class of digit "2" of the training set

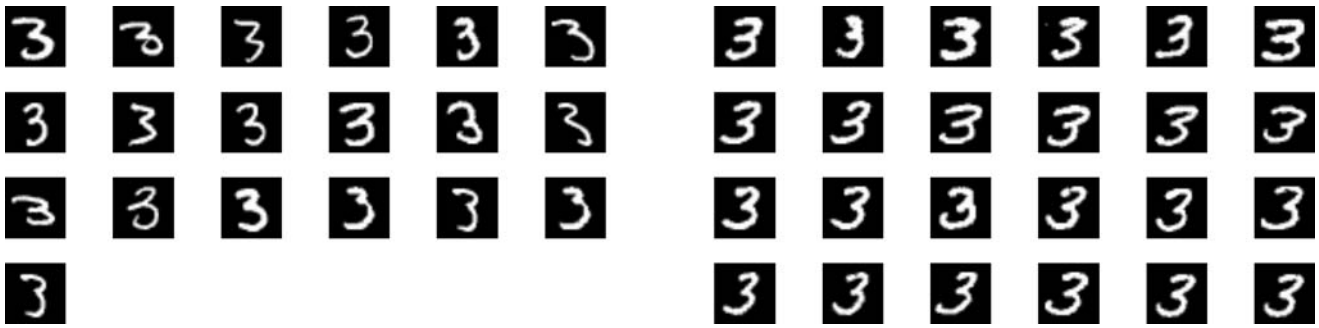

Fig. 16 Two clusters obtained by K-means clustering for the 20 images for the class of digit "3" of the training set

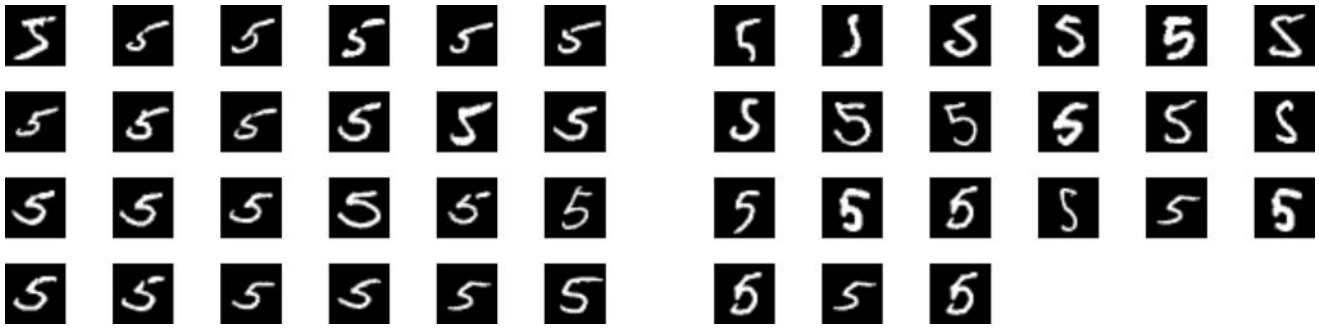

Fig. 17 Two clusters obtained by K-means clustering for the 20 images for the class of digit "5" of the training set

Table 1 Classification error rate on the test sample for the mnist dataset

Naive classification	Classification with warping	Classification with warping after clustering	SVM
30.2%	15.3%	8.6%	21.3%

can be easily performed with the tune function of the R library `e1071`.

In Table 1, we give the mis-classification rate over the 1000 images of the test samples for the two classification methods described above and our method based on warping before and after clustering with K-means. The classification with warping clearly gives the best result. This seems natural as this rule is the only one which takes into account the spatial local deformations that may exist between similar images. One may argue that a classification rate of 15.3% is not very satisfactory and that much better rates of classification have been obtained for this database (see e.g. [24]). However, remark first that we have only used 20 images per class for the training set which is very small. Secondly, we only want to show that taking into account the spatial variability due to the presence of local deformations between images may improve standard classification rules. At last, we can largely improve this performance using several clusters to describe each class as pointed in third column of Table 1 (8.6% classification error rate).

Finally, note that classifying images using the distances to the orbit generated by the deformation on the learned templates for each class is questionable, and seems to give not optimal results when compared to the performances obtained by [5] with small training sets of the MNIST database. Some further work is certainly needed to improve these results by using for example non-linear edge detectors features as in [5].

7 Conclusion and Perspectives

We end this paper by discussing several theoretical and computational aspects of our approach. First remark that we have built a very general model of random diffeomorphisms to warp images. This construction relies mainly on the choice of the basis functions e_k for generating the deformations. The choice of the e_k's is relatively large since one is only restricted to take functions with a sufficient number of derivatives that vanish at the boundaries of $[0,1]^2$. Moreover, our estimation procedure does not require the choice of a priori distributions for the random coefficients a_k^i. Hence, this model is very flexible as many parameterizations can be chosen.

Nevertheless, some difficult problems remain to be studied. We have discussed many different ways for incorporating some regularization in our estimation procedure. However, all these regularization methods depends on some hyperparameters that have to be carefully calibrated, and a challenging problem is to find data-based choices for these parameters. Moreover, we have only focused on the estimation of the mean pattern of a set of images, but one would like to build other statistics like principal modes of variations of the learned distribution of the images or the deformations. Building statistics going beyond the simple mean of set of images within the setting of our model is very challenging for future investigation.

Acknowledgements We are very much indebted to the referees and the Editor for their constructive criticism, comments and remarks that resulted in a significant improvement of the original manuscript.

Appendix A

A.1 Proof of Theorem 1

To obtain the asymptotic convergence of (3.3) toward (3.4) we use the following proposition whose proof follows from Theorem 6.3 in [6]:

Proposition 1 *Assume that the following two conditions hold*

(C1) *the set* $\{f(\cdot,\cdot,Z) : Z \in \mathcal{Z}\}$ *is an equicontinuous family of functions at each point of* $\mathcal{X} = [-A; A]^{2K} \times \mathbb{R}^N$.

(C2) *there is a continuous function* $\phi : \mathcal{X} \to \mathbb{R}^+$ *such that* $\int_{\mathcal{X}} \phi(a,\varepsilon) dP(a,\varepsilon) < +\infty$, *and for all* $(a,\varepsilon) \in \mathcal{X}$ *and* $Z \in \mathcal{Z}$, $|f(a,\varepsilon,Z)| \leqslant \phi(a,\varepsilon)$.

Then

$$\hat{Q}_\infty \subset Q_0 \quad a.s., \tag{A.1}$$

where $\hat{Q}_\infty$ *is defined as the set of accumulation points of the* $\hat{Z}_n$, *i.e. the limits of convergent subsequences* $\hat{Z}_{n_k}$ *of minimizers* $\hat{Z}_n \in \hat{Q}_n$.

In what follows, we establish assumptions (C1) and (C2) which proves Theorem 1.

Let us denote by $\langle I_1, I_2\rangle = \sum_{p=1}^N I_1(p) I_2(p)$ the "inner product" on the pixels p and by $|I_1|_{\mathcal{P}}$ the empirical "norm" associated to this inner product, where I_1, I_2 denotes two images observed at N pixels (and can thus be viewed as vectors in $\mathbb{R}^N$). We start with establishing a result on the regularity of F and F_n.

Lemma 1 *F and F_n are continuous over $\mathcal{Z}$ with respect to the supremum norm $\|.\|_\infty$ on $[0;1]^2$.*

Proof We first study the map $Z \to f(a,\varepsilon,Z)$. Consider $(Z_1, Z_2) \in \mathcal{Z}^2$ and fix any parameters of the deformations a and noise ε. Remark that for $Z \in \mathcal{Z}$, one can find $v_Z \in \mathcal{V}_A$ such that

$$v_{a,\varepsilon,Z} = \arg\min_{v \in \mathcal{V}_A} f(a,\varepsilon,Z),$$

where $f(a,\varepsilon,Z) = |I_a + \varepsilon - Z \circ \Phi_v^1|_{\mathcal{P}}^2$. This minimum is reached in $\mathcal{V}_A$ since $\mathcal{V}_A$ is here described by a bounded and closed finite dimensional space which is thus compact.

Using the mere definition of $v_{Z_1} = v_{a,\varepsilon,Z_1}$ and $v_{Z_2} = v_{a,\varepsilon,Z_2}$, we get

$$\begin{aligned}
&|I^\star \circ \Phi_a^1 + \varepsilon - Z_1 \circ \Phi_{v_{Z_1}}^1|_{\mathcal{P}}^2 \\
&\quad \leqslant |I^\star \circ \Phi_a^1 + \varepsilon - Z_1 \circ \Phi_{v_{Z_2}}^1|_{\mathcal{P}}^2 \\
&\quad \leqslant 2|I^\star \circ \Phi_a^1 + \varepsilon - Z_2 \circ \Phi_{v_{Z_2}}^1|_{\mathcal{P}}^2 \\
&\qquad + 2|(Z_1 - Z_2) \circ \Phi_{v_{Z_2}}^1|_{\mathcal{P}}^2.
\end{aligned}$$

Using the coarse following upper bound

$$|(Z_1 - Z_2) \circ \Phi_{v_2}^1|_{\mathcal{P}}^2 \leqslant N \|Z_2 - Z_1\|_\infty^2,$$

leads to

$$f(a,\varepsilon,Z_1) \leqslant f(a,\varepsilon,Z_2) + N\|Z_2 - Z_1\|_\infty^2.$$

Finally, this implies that

$$|f(a,\varepsilon,Z_1) - f(a,\varepsilon,Z_2)|^2 \leqslant N\|Z_2 - Z_1\|_\infty^2$$

proving the continuity of the function $Z \to f(a,\varepsilon,Z)$. We now return to the functions F and F_n, we have

$$|f(a,\varepsilon,Z)| \leqslant 2|I^\star \circ \Phi_a^1 + \varepsilon|_{\mathcal{P}}^2 + \underbrace{2|Z \circ \Phi_{v_{Z_1}}^1|_{\mathcal{P}}^2}_{\leq M}$$

since $\|Z\|_\infty$ is bounded by some constant M independent of a and ε. Then we get from assumptions A1 and A2 that

$$\int_{[-A;A]^{2K} \times \mathbb{R}^N} \left[|I^\star \circ \Phi_a^1 + \varepsilon|_{\mathcal{P}}^2 + M \right] dP(a,\varepsilon) < +\infty,$$

$I^\star$ being bounded since it is a Lipschitz on a $[0;1]^2$.

Hence $Z \to \int f(a,\varepsilon,Z) dP(a,\varepsilon) = F(Z)$ is continuous using the dominated convergence theorem. By the same argument, F_n is also continuous, which completes the proof. □

We next establish the existence of Q_0 and $\hat{Q}_n$. From the definition of the sets of minimizers, $\hat{Q}_n$ stands for candidates of the estimate of the mean image and Q_0 candidates for the mean image. Using the continuity of F and F_n (Lemma 1) and since $\mathcal{Z}$ is compact, we deduce the next result:

Lemma 2 *Q_0 and $\hat{Q}_n$ are well defined and non empty for all integer $n \in \mathbb{N}$.*

We now establish the conditions (C1) and (C2). We study first the family of functions indexed by $Z \in \mathcal{Z}$: $\{f(.,.,z), z \in \mathcal{Z}\}$.

Proposition 2 *For any compact set $\mathcal{Z}$, $\{f(.,.,z),\ z \in \mathcal{Z}\}$ is an equicontinuous family of functions of variables (a,ε).*

Proof Let $a_1, a_2, \varepsilon_1, \varepsilon_2$ be such that (for the standard euclidean norm on $[-A;A]^{2K} \times \mathbb{R}^N$)

$$\|(a_1,\varepsilon_1) - (a_2,\varepsilon_2)\| \leqslant \delta,$$

and note v_{Z_i} the optimal vector field obtained to match Z_i on I_{a^i,ε^i}. Hence, for any $Z \in \mathcal{Z}$, one have

$$\begin{aligned}
f(a_1,\varepsilon_1,Z) &= |I^\star \circ \Phi_{a_1}^1 + \varepsilon_1 - Z \circ \Phi_{v_{a_1,\varepsilon_1,Z}}^1|_{\mathcal{P}}^2 \\
&\leqslant |\, I^\star \circ \Phi_{a_1}^1 + \varepsilon_1 - Z \circ \Phi_{v_{a_2,\varepsilon_2,Z}}^1|_{\mathcal{P}}^2 \\
&\leqslant |\, I^\star \circ \Phi_{a_2}^1 + \varepsilon_2 - Z \circ \Phi_{v_{a_2,\varepsilon_2,Z}}^1|_{\mathcal{P}}^2 \\
&\qquad + |\varepsilon_1 - \varepsilon_2 + I^\star \circ \Phi_{a_1}^1 - I^\star \circ \Phi_{a_2}^1|_{\mathcal{P}}^2 \\
&\qquad + 2\langle I^\star \circ \Phi_{a_2}^1 + \varepsilon_2 - Z \circ \Phi_{v_{a_2,\varepsilon_2,Z}}^1, \\
&\qquad\qquad \varepsilon_1 - \varepsilon_2 + I^\star \circ \Phi_{a_1}^1 - I^\star \circ \Phi_{a_2}^1 \rangle.
\end{aligned}$$

Then, using the fact that the noise is bounded and that the images in $\mathcal{Z}$ are uniformly bounded, we obtain that there is a constant Λ such that

$$\begin{aligned} &f(a_1,\varepsilon_1,Z)\\ &\quad\leqslant f(a_2,\varepsilon_2,Z)+2|I^\star\circ\Phi^1_{a_1}-I^\star\circ\Phi^1_{a_2}|^2_{\mathcal{P}}+2|\varepsilon_2-\varepsilon_1|^2_{\mathcal{P}}\\ &\qquad+\Lambda\left(|I^\star\circ\Phi^1_{v_{a_1,\varepsilon_1,Z}}-I^\star\circ\Phi^1_{v_{a_2,\varepsilon_2,Z}}|_{\mathcal{P}}+|\varepsilon_2-\varepsilon_1|_{\mathcal{P}}\right), \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz and the triangular inequalities. Under assumption A2, we get

$$\begin{aligned} &f(a_1,\varepsilon_1,Z)-f(a_2,\varepsilon_2,Z)\\ &\quad\leqslant 2L^2\|\Phi^1_{a_2}-\Phi^1_{a_1}\|^2+2|\varepsilon_2-\varepsilon_1|^2_{\mathcal{P}}\\ &\qquad+\Lambda\left(L\|\Phi^1_{a_2}-\Phi^1_{a_1}\|+|\varepsilon_2-\varepsilon_1|_{\mathcal{P}}\right)\\ &\quad\leqslant 2L^2N\|\Phi^1_{a_2}-\Phi^1_{a_1}\|^2_\infty+2|\varepsilon_2-\varepsilon_1|^2_{\mathcal{P}}\\ &\qquad+\Lambda\left(L\sqrt{N}\|\Phi^1_{a_2}-\Phi^1_{a_1}\|_\infty+|\varepsilon_2-\varepsilon_1|_{\mathcal{P}}\right) \end{aligned}$$

Using results in [42], $(v,\|.\|_\infty)\to(\Phi^1_v,\|.\|_\infty)$ is continuous. Hence under an appropriate choice of δ_1 and δ_2 such that

$$\|a_1-a_2\|\leqslant\delta_1\qquad|\varepsilon_1-\varepsilon_2|_{\mathcal{P}}\leqslant\delta_2,$$

then

$$|f(a_1,\varepsilon_1,Z)-f(a_2,\varepsilon_2,Z)|\leqslant\eta,$$

which proves the equicontinuity of $\{f(.,.,Z),\ Z\in\mathcal{Z}\}$, and completes the proof. □

Thus assumption (C1) is proved. The proof of assumption (C2) follows from the proof of Lemma 1.

A.2 Proof of Theorem 2

We provide here a proof of consistency of the M-estimator defined in Theorem 2. Recall that we consider now the more general case where the images I_i are i.d.d. observations derived from an *unknown* distribution P on $\mathbb{R}^N$.

First remark that from assumption (3.8) and since Λ is finite, the supremum norm $\|\cdot\|_\infty$ for functions Z_θ on $[0,1]^2$ (with $\theta\in\mathbb{R}^\Lambda$ is equivalent to the supremum norm on $\mathbb{R}^\Lambda$. Therefore, by equivalence of norms, any function defined on the set of images $\mathcal{Z}=\{Z_\theta,\theta\in\Theta\}$ that is continuous with respect to the supremum norm $\|\cot\|_\infty$ for functions Z_θ on $[0,1]^2$ is also a continuous function on $\mathbb{R}^\Lambda$.

To derive the result of Theorem 2, one can then simply apply Theorem 5.10 of [38] which provides sufficient conditions for the consistency of M-estimator in general cases.

Recall that for our purpose, we have set

$$\mathrm{pen}_1(v)=\sum_{i=1}^{2}\sum_{k,k'=1}^{K}a^i_k\Gamma_{k,k'}a^i_{k'},$$

and

$$\mathrm{pen}_2(\theta)=\sum_{\lambda\in\Lambda}|\theta_\lambda|.$$

With our notations, this theorem ensures that

$$\lim_{n\to\infty}\|\hat{Z}_n-Z_{\theta^\star}\|_\infty=0\quad a.s.,$$

under the conditions

(B1) $\{f(.,Z_\theta),\theta\in\Theta\}$ is a Glivenko-Cantelli class,
(B2) $F(Z_\theta)$ has a unique minimum at $Z_{\theta^\star}$ for $\theta\in\Theta$.

Condition (B2) is a mere assumption of Theorem 2. The condition (B1) is somewhat more complicated to establish and rely on the theory of empirical processes. We proceed as in Lemma 1 using the compactness assumption for the parameters a that define the vector fields v_a. For any Z_{θ_1} and Z_{θ_2} in $\mathcal{Z}$, and any image $I\in\mathbb{R}^N$, we denote by $v_1(I)$ and $v_2(I)$ the vector fields which yield $f(I,Z_{\theta_1})$ and $f(I,Z_{\theta_2})$ i.e.

$$v_k(I)=\arg\min_{v\in\mathcal{V}}\left[\|I-Z_{\theta_k}\circ\Phi^1_v\|^2_{\mathcal{P}}+\lambda_1\mathrm{pen}_1(v)\right],\quad k=1,2.$$

If we denote by $\tilde{f}(I,Z_\theta)$ the map $f(I,Z)-\lambda_2\mathrm{pen}_2(\theta)$, we have

$$\begin{aligned} \tilde{f}(I,Z_{\theta_1})&=\|I-Z_{\theta_1}\circ\Phi^1_{v_1(I)}\|^2_{\mathcal{P}}+\lambda_1\mathrm{pen}_1(v_1(I)) \qquad (A.2)\\ &\leqslant\|I-Z_{\theta_1}\circ\Phi^1_{v_2(I)}\|^2_{\mathcal{P}}+\lambda_1\mathrm{pen}_1(v_2(I))\\ &\leqslant N\|Z_{\theta_1}-Z_{\theta_2}\|^2_\infty\\ &\quad+\|I-Z_{\theta_2}\circ\Phi^1_{v_2(I)}\|^2_{\mathcal{P}}+\lambda_1\mathrm{pen}_1(v_2(I))\\ &\leqslant N\|Z_{\theta_1}-Z_{\theta_2}\|^2_\infty+\tilde{f}(I,Z_{\theta_2}). \qquad (A.3) \end{aligned}$$

The above inequality immediately imply the continuity of $Z\mapsto\tilde{f}(I,Z)$ and of course of $Z\mapsto f(I,Z)$ for any fixed image I with respect to the norm $\|.\|_\infty$ on $\mathcal{Z}$ which establishes that $Z\mapsto f(I,Z)$ is continuous, for any image I.

Then the compactness assumption on the set $\mathcal{V}$ of vector fields, and the continuity of pen_1, imply that $\mathrm{pen}_1(v)$ is uniformly bounded by a constant C_1 for $v\in\mathcal{V}$. Also, since $\mathrm{pen}_2(\theta)$ is a continuous function of Z_θ, one has that for any fixed $Z_{\theta_0}\in\mathcal{Z}$ and for any $\delta>0$, $\mathrm{pen}_2(\theta)-\mathrm{pen}_2(\theta_0)$ is uniformly bounded by a constant C_2 when $\mathcal{Z}_\theta\in B(\mathcal{Z}_0,\delta)$, and this bound is independent of I. Therefore, from the inequality (A.3), we derive that

$$\sup_{Z/\|Z-Z_0\|_\infty\leqslant\delta}|f(I,Z)|\leqslant N\delta^2+N\|I-Z_0\|^2_{\infty,N}+\lambda_1C_1+\lambda_2C_2,$$

which is dominated by a function of I. Since it is assumed that

$$\int \|I\|^2_{\infty,N} dP(I) < \infty,$$

hence, on any neighborhood B of an image $Z_0 \in \mathcal{Z}$, $\sup_{Z\in B} |f(.,Z)|$ is uniformly bounded by an integrable function (with respect to $dP(I)$) depending only on $I \in \mathbb{R}^N$.

For any $\theta \in \Theta$, let Θ_m be a decreasing sequence of neighborhoods such that $\bigcap_m \Theta_m = \{\theta\}$. Define $f_{u,m}(.)$ respectively $f_{l,m}(.)$ the supremum, resp. the infimum of $f(.,Z_\theta)$ over $\theta \in \Theta_m$:

$$f_{l,m}(I) = \inf_{\theta\in\Theta_m} f(I,Z_\theta) \quad \text{and} \quad f_{u,m}(I) = \sup_{\theta\in\Theta_m} f(I,Z_\theta).$$

Continuity implies that $\lim_{m\to+\infty}(f_{u,m} - f_{l,m}) = f(.,Z_\theta) - f(.,Z_\theta) = 0$. Dominated Convergence yields that $\lim_m \int (f_{u,m}(I) - f_{l,m}(I))dP(I) = 0$. Finally, for any $\theta \in \Theta$ and $\epsilon > 0$, there exists a neighborhood $B = B(\theta)$ and two functions $f_{u,B}$ and $f_{l,B}$ such that $\int (f_{u,B}(I) - f_{l,B}(I))dP(I) \leqslant \epsilon$. Compacity of Θ implies that there is a subcollection of such neighborhoods B, which covers Θ, resulting in a finite number of couple of functions $(f_{u,B}, f_{l,B})$. Hence for all $\theta \in \Theta$, write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n f_{l,B}(I_i) - \epsilon &\leqslant \frac{1}{n}\sum_{i=1}^n f(I_i,Z_\theta) - \int f(I,Z_\theta)dP(I)\\ &\leqslant \frac{1}{n}\sum_{i=1}^n f_{u,B}(I_i) + \epsilon.\end{aligned}$$

Since the set of functions $f_{u,B}$ and $f_{l,B}$ is finite, we have

$$\sup_B \left|\frac{1}{n}\sum_{i=1}^n f_{u,B}(I_i) - \int f_{u,B}(I)dP(I)\right| \leqslant \epsilon,$$

$$\sup_B \left|\frac{1}{n}\sum_{i=1}^n f_{l,B}(I_i) - \int f_{l,B}(I)dP(I)\right| \leqslant \epsilon,$$

hence

$$\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^n f(I_i,Z_\theta) - \int f(I,Z_\theta)dP(I)\right| \leqslant 2\epsilon. \qquad \text{(A.4)}$$

From (A.4), $\{f(.,Z) : Z \in \mathcal{Z}\}$ is thus a Glivenko-Cantelli class which shows that (B1) is true, completing the proof of Theorem 2.

References

1. Allassonière, S., Amit, Y., Trouvé, A.: Toward a coherent statistical framework for dense deformable template estimation. J. Stat. R. Soc. B **69**, 3–29 (2007)
2. Amit, Y., Grenander, U., Piccioni, M.: Structural image restoration through deformable template. J. Am. Stat. Assoc. **86**, 376–387 (1991)
3. Antoniadis, A., Fan, J.: Regularization of wavelet approximations. J. Am. Stat. Assoc. **96**, 939–967 (2001)
4. Allassonière, S., Kuhn, E., Trouvé, A.: Bayesian deformable models building via stochastic approximation algorithm: a convergence study. Technical Report (2007)
5. Amit, Y., Trouvé, A.: Pop: patchwork of parts models for object recognition. Int. J. Comput. Vis. **75**, 267–282 (2007)
6. Biscay, R.J., Mora, C.M.: Metric sample spaces of continuous geometric curves and estimation of their centroids. Math. Nachr. **229**, 15–49 (2001)
7. Candès, E., Donoho, D.: Recovering edges in ill-posed inverse problems: optimality of curvelet frames. Ann. Stat. **30**, 784–842 (2000)
8. Charpiat, G., Faugeras, O., Keriven, R.: Approximations of shape metrics and application to shape warping and empirical shape statistics. Found. Comput. Math. **5**, 1–58 (2005)
9. Charpiat, G., Faugeras, O.D., Keriven, R.: Image statistics based on diffeomorphic matching. In: ICCV, pp. 852–857. IEEE Computer Society, Los Alamitos (2005)
10. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
11. De Boor, C.: A Practical Guide to Splines. Applied Mathematical Sciences, vol. 27. Springer, New York (1978)
12. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, New York (1998)
13. Faugeras, O., Hermosillo, G.: Well-posedness of eight problems of multi-modal statistical image-matching. Biomed. Imaging **15**(23), 64 (2002)
14. Gamboa, F., Loubes, J.-M., Maza, E.: Semi-parametric estimation of shifts. Electron. J. Stat. **1**, 616–640 (2007)
15. Grenander, U., Miller, M.I.: Computational anatomy: an emerging discipline. Q. Appl. Math. **56**(4), 617–694 (1998). Current and future challenges in the applications of mathematics (Providence, RI, 1997)
16. Glasbey, C.A., Mardia, K.V.: A penalized likelihood approach to image warping. J. R. Stat. Soc. Ser. B Stat. Methodol. **63**(3), 465–514 (2001)
17. Grenander, U.: General Pattern Theory a Mathematical Study of Regular Structures. Oxford University Press, New, York (1994)
18. Glaunès, J., Vaillant, M., Miller, M.I.: Landmark matching via large deformation diffeomorphisms on the sphere. J. Math. Imaging Vis. **20**(1–2), 179–200 (2004). Special issue on mathematics and image analysis
19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Berlin (2003), Chap. 10
20. Huilling, L.: On the consistency of procrustean mean shapes. Adv. Appl. Probab. **30**, 53–63 (1998)
21. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice-Hall, Upper Saddle River (1989)
22. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley Series in Probability and Statistics. Wiley, Chichester (1999)
23. Korostelëv, A.P., Tsybakov, A.B.: Minimax Theory of Image Reconstruction. Lecture Notes in Statistics, vol. 82. Springer, New York (1993)
24. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
25. Loubes, J.-M., van de Geer, S.: Adaptive estimation with soft thresholding penalties. Stat. Neerl. **56**(4), 454–479 (2002)
26. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (eds.)

Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
27. Mallat, S.: A Wavelet Tour of Signal Processing. AP Professional, 2nd edn. Academic Press, San Diego (1998)
28. Markussen, B.: A statistical approach to large deformation diffeomorphisms. In: CVPRW'04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04), vol. 12, p. 181. IEEE Computer Society, Los Alamitos (2004)
29. Markussen, B.: Large deformation diffeomorphisms with application to optic flow. Comput. Vis. Image Underst. **106**(1), 97–105 (2007)
30. Pennec, X.: Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. J. Math. Imaging Vis. **25**(1), 127–154 (2006)
31. Samaria, F.S., Harter, A.: Parameterisation of a stochastic model for human face identification, pp. 138–142 (1994)
32. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
33. Trouvé, A., Younes, L.: Local geometry of deformable templates. SIAM J. Math. Anal. **37**(1), 17–59 (2005) (electronic)
34. Trouvé, A., Younes, L.: Metamorphoses through lie group action. Found. Comput. Math. **5**(2), 173–198 (2005)
35. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
36. van de Geer, S.A.: Applications of Empirical Process Theory. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 6. Cambridge University Press, Cambridge (2000)
37. Van der Waart, A.: Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 27. Cambridge Univ. Press, New York (1998)
38. Van der Waart, A.: Semiparametric Statistics. Lectures on Probability Theory, Ecole d'Ete de Probabilites de St. Flour XXIX-1999. Springer, Berlin (2002)
39. Vimond, M.: Efficient estimation in homothetic shifted in regression models. Ann. Stat. (2006, in press)
40. Vaillant, M., Miller, M.I., Trouvé, A., Younes, L.: Statistics on diffeomorphisms via tangent space representations. NeuroImage **23**, 161–169 (2004)
41. Younes, L.: Deformation analysis for shape and image processing. Lecture Notes available at: http://cis.jhu.edu/~younes/LectureNotes/deformationAnalysis.pdf
42. Younes, L.: Invariance, Déformations et Reconnaissance de Formes Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 44. Springer, Berlin (2004)

Jérémie Bigot received his PhD, Functional analysis of variance with wavelets, in Applied Mathematics in 2003 at the University Joseph Fourier, Grenoble, France. From 2003 to 2004 he was a postdoc researcher at the Institute of Statistics, UCL, Belgium. Since 2004, he is an Assistant Professor at Institut de Mathématiques de Toulouse. His research interest deals mainly with nonparametric statistics and warping models for signal and image processing.

Sébastien Gadat studied statistics and signal processing at the Ecole Normale Supérieure in Cachan. He received the PhD degree in mathematics in 2004 from ENS-Cachan and is currently an Associate Professor at the University of Toulouse. His research activities are mainly focused on asymptotic statistics and statistical analysis of signal deformation.

Jean-Michel Loubes received his PhD, M-estimation and empirical processes, in applied Mathematics in 2001 at the University of Toulouse 3. From 2002 to 2007 he was Investigator at CNRS in first Orsay University and then Montpellier 3 University. Since 2007, he is Professor at Institut de Mathématiques de Toulouse. His research interest deals mainly with asymptotic statistics.