TOWARD A COHERENT STATISTICAL FRAMEWORK FOR DENSE DEFORMABLE TEMPLATE ESTIMATION

S. ALLASSONNIÈRE, Y. AMIT, A. TROUVÉ

Abstract. The problem of estimating probabilistic deformable template models in the field of computer vision or of probabilistic atlases in the field of computational anatomy has not yet received a coherent statistical formulation and remains a challenge. In this paper, we provide a careful definition and analysis of a well defined statistical model based on dense deformable templates for gray level images of deformable objects. We propose a rigorous Bayesian framework for which we can derived an iterative algorithm for the effective estimation of the geometric and photometric parameters of the model in a small sample setting, together with an asymptotic consistency proof. The model is extended to mixtures of finite numbers of such components leading to a fine description of the photometric and geometric variations. We illustrate some of the ideas with images of handwritten digits, and apply the estimated models to classification through maximum likelihood.

1. Introduction

Modeling the geometric variability of object classes with deformable templates has proved to be a powerful tool in image analysis. Important applications can be found in general object detection and recognition problems in vision, ([6], [12], [1], [4]), where in addition to explicit modeling of geometric variability the deformable template framework facilitates the formulation of credible generative models for the data. Another important application is in medical imaging involving the quantitative analysis of anatomies [11, 19]. Here deformable models offer two important contributions: the possibility of generating digital anatomical atlases and the emphasis on deformation "costs" as the core of a quantitative analysis of shape variability. Over the last decade progress has been made in formulating a metric approach to shapes. Shape manifolds and their metrics have been properly defined using deformation costs and actions of infinite dimensional transformation groups [16, 18, 17, 5, 20]. Originating in Grenander's pattern theory this direction of research has produced a rich family of shape spaces, and the study of their properties and intrinsic geometries is now an active field of research.

On the other hand the probabilistic and statistical side of the deformable template framework has received less attention. In [3] a coherent probabilistic deformable model is proposed using Gaussian random vector fields to model the deformation process with i.i.d noise added to the deformed template at the observation pixels, but no proposals are offered for estimating the relevant parameters. In general very little can be found on estimating a

Date: March 28, 2006.

dense deformable model from a relatively small training set of images. Of interest are the estimation of the covariance structure of the random deformation field (i.e. the metric) as well as the template image. As we will see there are some non trivial conceptual problems that arise in a rigorous formulation of these issues which are often ignored or buried in ad doc approximations.

In terms of estimating the metric (covariance structure of the deformation field) one finds little beyond the wide spread use of PCA analysis: assume that the template is known, then build a PCA model from the optimal deformations computed between the template and each observation. This point of view suffers from two important drawbacks: in the small sample setting, the empirical covariance is degenerate leading to ad hoc smoothing and thresholding methods; furthermore, this cannot provide a coherent statistical scheme since the initial optimal matchings are no longer compatible with the updated covariance structure. This problem seems to be generally ignored despite the fact that these algorithms fail to be statistically *consistent* as the sample size grows.

Even less can be found on the problem of estimating the template. A trivial possibility, used quite often, is to choose one of the observed images. This option can be viewed as an 'apriori' guess with the drawback that the resulting template will be corrupted by noise and only defined on a discrete set of pixels. A more involved solution in the context of metric shape spaces is to compute a template as an 'average' image (Karcher mean) for the Riemannian metric on the space of images. Specifically one searches for the point on the image manifold that minimizes the distances to all samples.

It is important to emphasize that the space of templates is typically distinct from the space of observed images (noisy and discretized). Templates are smooth functions defined on continuous domains whereas images are defined on discrete domains, and are not necessarily smooth. Usually this problem is ignored and one considers the initial template as a smoothed ("denoised") and interpolated version of one of the observations. Other observations are "projected" on the orbit of this template by computing the optimal deformation given a noise model. The Karcher mean is then computed yielding a template. In some cases this process is iterated. Even if this provides an effective algorithm to compute a template, it is sensitive to the choice of the initial observation and to the noise present there.

An important step towards the statistical formulation of template estimation is developed in [10] through a maximum likelihood point of view leading to an approach analogous to generalized Procrustes analysis. The stochastic model for the observed image is

$$y(x + u(x)) = I_0(x) + \epsilon_x,\ \epsilon_x \sim \mathcal{N}(0, \sigma^2),\ x \in X$$

where X is the pixel grid, I_0 is the template, u is a random deformation field with a given distribution and $(\epsilon_x)_{x \in X}$ is white noise (a more complex Fourier-Von Mises image model is also provided but we work here with the simpler Gaussian model since it does not change the overall analysis). In this model, σ is fixed, u is a hidden variable, and the model

parameter I_0 is estimated from data as

$$\hat{I}_0 = \underset{I_0,u_1,\cdots,u_n}{\operatorname{argmax}} p_{I_0}(u_1,\cdots,u_n|y_1,\cdots,y_n) = \underset{I_0,u_1,\cdots,u_n}{\operatorname{argmax}} \sum_{i=1}^{n} \big(\log(p_{I_0}(y_i|u_i)) + \log(p(u_i))\big)$$

The above approach gives interesting results in several template estimation problems but is not entirely satisfactory if we want to define a coherent statistical framework (see also comments in subsequent discussion [15].) Indeed, since the image y is observed on a discrete grid of pixels, the variables $(y(x+u(x)))_{x\in X}$ are *not observed* for a generic displacement u. This problem is handled by interpolation. However this still does not provide a well defined statistical model for the observed data. Furthermore no proposal is provided for estimating the metric together with the template, nor is the noise parameter (σ) estimated. A similar approach is proposed in [14].

In [8, 9] a related problem is addressed in the context of object detection. Objects are modeled through a sparse representations consisting of an object specific probabilistic data model around interest points on the object and a generic background model everywhere else. A model for the geometric arrangement of the interest points is introduced as well. The Bayesian formulation enables the estimation of models from small data sets using variants of the EM algorithm. The sparse framework using interest points raises some complications in formulating well defined statistical models. It also requires the computation of matches between point sets leading to difficult combinatorial optimization problems in the estimation procedure.

In this paper our goal is to propose a coherent statistical framework for dense deformable templates both in terms of the probability model, and in terms of the estimation procedure of the template and of the deformation covariance structure. The framework is extended to *mixtures* of template models, which prove useful for modeling heterogeneous object classes. The observations are modeled on a fixed discrete grid but the template and the deformation field are defined on *continuous* domains. For simplicity we assume an additive Gaussian noise model, but the theoretical and algorithmic setting can be easily generalized to other forms of data models.

We do not parameterize the template through its values on the observation grid, rather as a finite linear combination of kernels defined on the continuum. The deformation field is defined in a similar form and the covariance structure reduces to a finite dimensional covariance matrix. Estimation is formulated in a Bayesian framework with priors both on the template parameters and on the covariance parameters. We show that such apriori smoothing is essential in small sample problems. Estimation is formulated as a well defined maximum a-posteriori problem, with missing data - the deformations. We show that with some mild assumptions this procedure is consistent.

An EM formulation is proposed for estimation with finite samples. The expectation (E) step cannot be computed analytically due to the complex nature of the conditional distribution on the missing variables. A simple approximation is proposed using the mode of this distribution. For a different form of approximation to the estimation procedure and alternative photometrically invariant data models see [4].

The paper is organized as follows. In section 2 we present the probabilistic model and define the priors on the various parameters. Some properties of the maximum posterior estimate are provided in section 3, where we state a consistency theorem which is proved in the appendix. In section 4 we formulate the EM estimation procedure and offer a simple approximation which allows for efficient computation, this is then generalized in section 5 to the mixture case. In section 6, to illustrate some of the issues raised in the paper, we present some experiments on estimating mixtures of deformable template models for hand written digits from small samples.

2. The observation model

We observe a gray level image sequence $y_1^n = (y_i)_{1\leq i\leq n}$ defined on a grid of pixels Λ. Assume that $\Lambda \hookrightarrow \mathbb{R}^2$ where for each $s \in \Lambda$, x_s is the location of pixel s in $\mathbb{R}^2$. The points x_s are all in some fixed domain D (typically the square $[-1,1]\times[-1,1]$.) The template is defined as a function $I_0 : \mathbb{R}^2 \to \mathbb{R}$, and for each observation y, we assume the existence of an unobserved deformation field $z : \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$y(s) = I_0(x_s - z(x_s)) + \sigma\epsilon(s)$$

where $\epsilon(s)$ are i.i.d $\mathcal{N}(0,1)$, independent of all other variables. We denote by zI_0 the vector of observations of the deformed template at the grid points:

$$zI_0(s) = I_0(x_s - z(x_s)), s \in \Lambda,$$

so that

$$y = zI_0 + \sigma\epsilon$$

2.1. **The template model.** The template $I_0 : \mathbb{R}^2 \to \mathbb{R}$ is assumed to belong to a reproducing kernel Hilbert space V_p with kernel K_p. We focus on a fixed finite dimensional sub-space determined by a set of landmark points $(p_k)_{1\leq k\leq k_p}$. These points will typically cover a domain D_p which contains D since the deformations at times require template values outside the observed domain. (Typically $D_p = [-1.5, 1.5]\times[-1.5, 1.5]$.) The template is defined as a linear combination of the kernels centered at the landmark points, and is therefore parameterized by the coefficients $\alpha \in \mathbb{R}^{k_p}$. We write

$$I_\alpha = \mathbf{K_p}\alpha, \text{ where } (\mathbf{K_p}\alpha)(x) = \sum_{k=1}^{k_p} K_p(x, p_k)\alpha(k) \ . \tag{1}$$

2.2. **The deformation model.** We use the same framework to describe the deformations. Let V_g be a reproducing kernel Hilbert space of vector fields with kernel K_g. Pick a fixed set of landmarks $(g_k)_{1\leq k\leq k_g} \in D$. For $\beta = (\beta^{(1)}, \beta^{(2)}) \in \mathbb{R}^{k_g}\times\mathbb{R}^{k_g}$ we define the deformation field

$$z_\beta(x) = (\mathbf{K_g}\beta)(x) = \sum_{k=1}^{k_g} K_g(x, g_k)(\beta^{(1)}(k), \beta^{(2)}(k)). \tag{2}$$

If we assume that the underlying deformation field is Gaussian it induces a Gaussian distribution on β.

In the experiments below both K_p and K_g will be radial Gaussian kernels but any smooth kernel vanishing at infinity could be used.

2.3. **Parameters and likelihood.** The parameters of interest are α - the coefficients which determine the template (equation (1)), σ - the standard deviation of the additive noise, and the covariance matrix Γ_g of the variables β which determine the deformation (equation (2)). Let $\theta_g = \Gamma_g$ and let $\theta_p = (\alpha, \sigma^2)$. We assume that $\theta = (\theta_g, \theta_p)$ belongs to the parameter space Θ defined as the open set

$$\Theta = \{\ \theta = (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p},\ \sigma^2 > 0,\ \Gamma_g \in \Sigma^+_{2k_g,*}(\mathbb{R})\ \}.$$

Here $\Sigma^+_{2k_g,*}(\mathbb{R})$ is the set of strictly positive symmetric matrices which is identified through its upper triangular part and hence is viewed as an open subset of $\mathbb{R}^{k_g(2k_g+1)}$, inheriting the standard Lebesgue measure.

Note that the likelihood of the observed data has the form of an integral over the unobserved deformation parameters:

$$q(y|\theta) = \int q(y|\beta, \theta_p) q(\beta|\theta_g) d\beta\ ,$$

where

$$\begin{aligned} q(\beta|\theta_g) &= \exp\left(-\beta^t \Gamma_g^{-1} \beta/2\right) (2\pi)^{-k_g} |\Gamma_g|^{-1/2} \\ q(y|\beta, \theta_p) &= \exp\left(-\tfrac{|y - z_\beta I_\alpha|^2}{2\sigma^2}\right) (2\pi\sigma^2)^{-|\Lambda|/2}\ . \end{aligned} \tag{3}$$

2.4. **The Bayesian model.** Even though the parameters are finite dimensional it is unreasonable to compute a maximum-likelihood estimator when the training sample is small. Our goal is to demonstrate that with the introduction of apriori distributions on the parameters, estimation with small samples is still possible even within the rather complex framework described here, yielding good results in some concrete examples. We use standard conjugate priors - an inverse-Wishart ν_g on Γ_g, a normal prior with *fixed* mean μ_p and covariance matrix Γ_p on α and an inverse-Wishart prior on σ^2 as well. All priors are assumed independent. More formally we have

$$\begin{cases} (\Gamma_g, \theta_p) \sim \nu_g \otimes \nu_p \text{ with } \theta_p = (\alpha, \sigma^2) \\ \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_g) \mid \Gamma_g, \theta_p \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(z_{\beta_i} I_\alpha, \sigma^2 \mathrm{Id}_\Lambda) \mid \beta_1^n, \theta_p, \Gamma_g \end{cases}$$

where

$$(4)\quad \begin{cases} \nu_g(d\Gamma_g) \propto \left(\exp(-\langle \Gamma_g^{-1}, \Gamma_g^0\rangle/2)\dfrac{1}{\sqrt{|\Gamma_g|}}\right)^{a_g} d\Gamma_g,\ a_g > 2k_g + 1 \\ \\ \nu_p(d\sigma^2, d\alpha) \propto \left(\exp\left(-\dfrac{\sigma_0^2}{2\sigma^2}\right)\dfrac{1}{\sqrt{\sigma^2}}\right)^{a_p} \exp\left((\alpha-\mu_p)^t(\Gamma_p)^{-1}(\alpha-\mu_p)\right) d\sigma^2 d\alpha\ . \end{cases}$$

Note that for two matrices A, B we have $\langle A, B\rangle \doteq tr(A^t B)$.

The model interpretation is simple. Generate (α, σ^2) from ν_p, form the template $I_\alpha = \mathbf{K_p}\alpha$, and independently draw a covariance matrix Γ_g from ν_g. Next draw $\beta_i, i = 1, \ldots, n$ independently from $\mathcal{N}(0, \Gamma_g)$. The variables β_i determine the deformations z_{β_i} through equation (2). Finally generate $z_{\beta_i} I_\alpha$ and add i.i.d Gaussian noise with variance σ^2 to form the observations y_i.

2.4.1. *Choice of Gaussian priors.* A natural choice for the apriori covariance matrices Γ_p and Γ_g^0 is to consider the matrices induced by the metric of the spaces V_p and V_g. Define the square matrices

$$(5)\quad \begin{array}{l} M_p(k,k') = K_p(p_k, p_{k'})\ \forall 1 \le k, k' \le k_p \\ M_g(k,k') = K_g(g_k, g_{k'})\ \forall 1 \le k, k' \le k_g \end{array}$$

Setting $\Gamma_p = M_p^{-1}$ and $\Gamma_g^0 = M_g^{-1}$ we see that the exponent in the distribution defined in equation (3) corresponds to the norm of the function $\mathbf{K_g}\beta$ in the space V_g and the exponent in the distribution defined in equation (4) corresponds to the norm of $\mathbf{K_p}\alpha$ in the space V_p. This has a more precise justification as the restriction of a random Gaussian linear functional on the space V_p (V_g) to the subspace spanned by $\mathbf{K_p}$ ($\mathbf{K_g}$). This has the advantage of giving a prior that is essentially independent of the number of landmarks k_p and k_g, and that only depends on the global choice made for the reproducing kernel Hilbert spaces V_p and V_g. In this context, the number of landmarks used determines a trade-off between accuracy of the approximations of functions in the respective spaces and the amount of required computation.

3. Estimation

The parameter estimates are obtained by maximizing the posterior density on θ conditional on y_1^n.

$$\hat{\theta}_n = \operatorname*{argmax}_\theta q(\theta | y_1^n).$$

We first show that for any finite sample the maximum posterior will lie in the parameter set Θ, this is non-trivial due to the somewhat complex relation between the parameters and the observations. We then state a consistency theorem which is proved in the appendix.

Theorem 1 (Existence of the MAP estimator)**.** *For any sample y_1^n, there exists $\hat{\theta}_n \in \Theta$ such that*

$$q(\hat{\theta}_n | y_1^n) = \sup_{\theta \in \Theta} q(\theta | y_1^n)\,.$$

Proof. From equation (3) we have that for any $\theta = (\theta_p, \theta_g) \in \Theta$

$$q(y|\beta, \theta_p)q(\beta|\theta_g) \leq (2\pi\sigma^2)^{-|\Lambda|/2}(2\pi)^{-k_g}|\Gamma_g|^{-1/2}$$

so that

$$\log(q(\theta|y_1^n)) \leq -\frac{a_g}{2}\langle R_g, \Gamma_g^0\rangle + \frac{n+a_g}{2}\log|R_g| - \frac{a_p\sigma_0^2}{2\sigma^2} - \frac{n|\Lambda|+a_p}{2}\log(\sigma^2) \\ - \frac{1}{2}(\alpha-\mu_p)^t\Gamma_p^{-1}(\alpha-\mu_p) + \text{ C}$$

where $R_g = \Gamma_g^{-1}$, and C does not depend on the parameters. If we denote η_g^0 the smallest eigenvalue of Γ_g^0 and $\|R_g\|$ the operator norm of R_g (which is also its largest eigenvalue), we get

$$\langle R_g, \Gamma_g^0\rangle \geq \eta_g^0\|R_g\| \text{ and } \log(|R_g|) \leq (2k_g - 1)\log\|R_g\| - \log\|\Gamma_g\|$$

so that

$$\lim_{\|R_g\|+\|\Gamma_g\|\to\infty} -\frac{a_g}{2}\langle R_g, \Gamma_g^0\rangle + \frac{n+a_g}{2}\log|R_g| = -\infty\,.$$

Similarly, we can show

$$\lim_{\sigma^2+\sigma^{-2}\to\infty} -\frac{a_p\sigma_0^2}{2\sigma^2} - \frac{n|\Lambda|+a_p}{2}\log(\sigma^2) = -\infty$$

and

$$\lim_{|\alpha|\to\infty} -\frac{1}{2}(\alpha-\mu_p)^t\Gamma_p^{-1}(\alpha-\mu_p) = -\infty\,.$$

Now considering the Alexandrov one-point compactification $\Theta \cup \{\infty\}$ of Θ, we have

$$\lim_{\theta\to\infty} \log(q(\theta|y_1^n)) \to -\infty\,.$$

Since $\theta \to \log(q(\theta|y_1^n))$ is smooth on Θ, we get the result. □

3.1. **Consistency.** We are interested in the consistency properties of the MAP estimator without making strong assumptions on the distribution of the observations y_1^n. In other words we do not assume that the observations are generated by the model described above. We denote the distribution governing the observations by P and seek to prove the convergence of the MAP estimator to the set Θ_* of model distributions 'closest' to P:

$$\Theta_* = \{\ \theta_* \in \Theta \mid E_P(\log q(y|\theta_*)) = \sup_{\theta\in\Theta} E_P(\log q(y|\theta))\}.$$

Theorem 2 (Consistency)**.** *Assume that Θ_* is non empty. Then, for any compact set $K \subset \Theta$,*

$$\lim_{n\to+\infty} P(\ \delta(\hat{\theta}_n, \Theta_*) \geq \epsilon\ \wedge \hat{\theta}_n \in K\ ) = 0\,,$$

where δ is any metric compatible with the usual topology on Θ.

Proof. The theorem is an immediate application of Wald's consistency Theorem (see Theorem 5.14, pg. 48 in [21]). We only need to verify that $y \to \log q(y|\theta)$ is P *a.s.* upper semi-continuous and that for any $\theta \in \Theta$, there exists an open set $U \ni \theta$ such that $E_P(\sup_{\theta' \in U} \log^+(q(y|\theta'))) < \infty$ (where $\log^+$ is the positive part of log). This is straightforward in our setting since $\theta \to \log(q(y|\theta))$ is smooth for any y. Furthermore for any $\theta \in \Theta$, there exists an open set $U \ni \theta$ such that $\sup_{y,\theta' \in U} \log(q(y|\theta') < \infty$. □

The previous result is not entirely satisfactory. One would like to show that Θ_* is non empty and that the map estimator does not escape to the boundary of Θ as $n \to \infty$. Interestingly, without specific assumptions on P, such unexpected behavior can indeed occur, especially if we do not assume the templates to be uniformly bounded. We propose below a reasonable framework in which the convergence towards the set Θ_* is guaranteed.

To this end we extend the previous model by introducing a baseline image $I_b : \mathbb{R}^2 \to \mathbb{R}$ and define

$$I_\alpha = \mathbf{K_p}\alpha + I_b \,. \tag{6}$$

In the previous framework $I_b \equiv 0$.

Let $\Sigma^+_{2k_g}(\mathbb{R})$ be the set of non-negative (possibly degenerate) symmetric matrices on $\mathbb{R}^{2k_p}$. For any $R > 0$ denote

$$\begin{cases} \Theta^R = \{\ \theta = (\alpha, \sigma^2, \Gamma) \mid \alpha \in \mathbb{R}^{k_p},\ |\alpha| \leq R,\ \sigma^2 \in \mathbb{R}^*_+,\ \Gamma \in \Sigma^+_{2k_g}(\mathbb{R})\ \} \\ \\ v(R) = \sup_{\theta \in \Theta^R} E_P(\log q(y|\theta)) \\ \\ \Theta^R_* = \{\ \theta \in \Theta^R \mid E_P(\log q(y|\theta)) = v(R)\ \} \end{cases} \tag{7}$$

Following the proof of Theorem 1, we conclude that for any $R > 0$ the set of MAP estimators is a subset of Θ^R. Let $\hat{\theta}^R_n$ denote any MAP estimator. Let

$$\dim_\beta = 2k_g,\ \dim_y = |\Lambda| \tag{8}$$

be the dimension of the β variables and of the observed images y respectively.

Theorem 3 (Consistency on bounded prototypes)**.** *Assume that $dim_\beta < dim_y$, that $P(dy) = p(y)dy$ where the density p is bounded with exponentially decaying tails and that the observations y_1^n are i.i.d under P. Assume also that the baseline I_b (see (6)) satisfies $|I_b(x)| > a|x| + b$ for some positive constant a. Then $\Theta^R_* \neq \emptyset$ and for any $\epsilon > 0$*

$$\lim_{n \to \infty} P(\delta(\hat{\theta}^R_n, \Theta^R_*) \geq \epsilon) = 0 \,,$$

where δ is any metric compatible with the topology on Θ^R.

Proof. See appendix. □

The condition $\dim_\beta < \dim_y$ implies that the dimension of the deformations is less than the number of observed image pixels. This condition is quite weak and fulfilled in our applications.

The condition on the baseline image is somewhat less natural but without this assumption very large deformations can occur at no cost in terms of the likelihood, and it may happen that the best model is not achieved inside Θ^R but with covariance $\Gamma \to \infty$. In practice such degenerate behavior has not been observed in any of our numerical examples in which we have used $I_b \equiv 0$.

4. Estimation with the EM algorithm

Since the deformation coefficients β_i are unobserved the natural approach is to use iterative algorithms such as EM ([7]) which we briefly summarize. Assume the conditional distribution on the unobserved variable u (in our case β) for any value of y and θ has a density with respect to some reference measure $\mu(du)$. We can write the log-marginal density on y as follows:

$$\log q(y|\theta) = \max_{\nu} \left[\int \log q(y,u|\theta)\nu(u)\mu(du) - \int \nu(u)\log\nu(u)\mu(du) \right], \tag{9}$$

where ν is any density over the variable u. The maximum is achieved for $\nu(u) = q(u|y,\theta)$. Thus maximizing the log-likelihood of the observed data with respect to the parameter becomes a double maximization

$$\max_{\theta}\max_{\nu} \left[\int \log q(y,u|\theta)\nu(u)\mu(du) - \int \nu(u)\log\nu(u)\mu(du) \right]. \tag{10}$$

The EM algorithm consists of iterating these two maximization steps. Given a current value θ_c of θ, the maximization with respect to the density ν is seen to yield $\nu_c(u) = q(u|\theta_c, y)$, or with multiple independent observations,

$$\nu_c(u_1^n) = \prod_{i=1}^{n} q(u_i|\theta_c, y_i).$$

This is often called the posterior density. Once ν_c is determined the second maximization - updating the parameters - involves only the first term in equation (10).

In the present context we assume here that $I_b \equiv 0$ (the introduction of a non vanishing baseline in the following computation is straightforward), we initialize the algorithm with the prior model θ_0 and we iterate the following two steps:

E step: Define the a-posteriori density:

$$\nu_l(\beta_1^n) = q(\beta_1^n|\theta_l, y_1^n).$$

Since the observations are independent ν_l is the product of the following densities

$$\nu_{l,i}(\beta) = \frac{q(y_i|\beta,\theta_{p,l})q(\beta|\theta_{g,l})}{\int q(y_i|\beta',\theta_{p,l})q(\beta'|\theta_{g,l})d\beta'}. \tag{11}$$

M step: Update the parameters:

$$\theta_{l+1} = \operatorname*{argmax}_{\theta} E_{\nu_l}\left(\log q(\beta_1^n, \theta|y_1^n)\right)) = \operatorname*{argmax}_{\theta} \int \log q(\beta_1^n, \theta|y_1^n)\nu_l(\beta_1^n)d\beta_1^n.$$

4.1. **Details of the maximization step.** The maximization step is divided in two, one involving geometric parameters and the other the photometric parameters. Writing the joint density we get

$$q(y_1^n, \beta_1^n, \theta) = q(\theta_g)q(\theta_p)q(\beta_1^n|\theta_g)q(y_1^n|\beta_1^n, \theta_p) \ .$$

So that,

$$\begin{aligned} \underset{\theta_g}{\operatorname{argmax}} E_{\nu_l}(\log q(y_1^n, \beta_1^n, \theta)) &= \underset{\theta_g}{\operatorname{argmax}} \{E_{\nu_l}(\log q(\beta_1^n|\theta_g)) + \log q(\theta_g)\} \ , \\ \underset{\theta_p}{\operatorname{argmax}} E_{\nu_l}(\log q(y_1^n, \beta_1^n, \theta)) &= \underset{\theta_p}{\operatorname{argmax}} \{E_{\nu_l}(\log q(y_1^n|\beta_1^n, \theta_p)) + \log q(\theta_p)\} \ . \end{aligned}$$

4.1.1. *Updating the geometric parameters.* The geometric parameter is simply the covariance matrix Γ_g of the Gaussian distribution of the β variables. For convenience we introduce the inverse covariance matrix $R_g = \Gamma_g^{-1}$. Plugging in the definitions in (4) and (3) we get

$$\begin{aligned} E_{\nu_l}(\log q(\beta_1^n|\theta_g)) + \log q(\theta_g) = \\ - \frac{n}{2}\langle R_g, [\beta\beta^t]_l\rangle + \frac{n}{2}\log|R_g| - \frac{a_g}{2}\langle R_g, \Gamma_g^0\rangle + \frac{a_g}{2}\log|R_g| + \text{Const}, \end{aligned}$$

where the empirical covariance matrix is defined as

$$[\beta\beta^t]_l = \frac{1}{n}\sum_{i=1}^{n}\int \beta\beta^t \nu_{l,i}(\beta)d\beta, \tag{12}$$

with $\nu_{l,i}(\beta)$ defined in (11). Setting the gradient with respect to R_g equal to 0 we get

$$\theta_{g,l+1} = \Gamma_{g,l+1} = \frac{1}{n + a_g}(n[\beta\beta^t]_l + a_g\Gamma_g^0). \tag{13}$$

4.1.2. *Updating the photometric parameters.* Here we are interested in the α variables that determine the template and in the noise parameter σ. Given a deformation field $z_\beta = \mathbf{K_g}\beta$ defined by the variables β, define the matrix $K_p^\beta \in \mathcal{M}_{|\Lambda|\times k_p}$ as:

$$K_p^\beta = z_\beta K_p(\cdot, p_k), \text{ i.e } K_p^\beta(s,k) = K_p(x_s - z_\beta(x_s), p_k), s \in \Lambda, 1 \leq k \leq k_p \ .$$

Thus when applying the deformation field z_β to I_α and evaluating at the image pixels we have,

$$z_\beta I_\alpha = K_p^\beta \alpha \ .$$

This yields

$$\log q(y_1^n|\beta_1^n, \theta_p) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}|y_i - K_p^{\beta_i}\alpha|^2 - \frac{n|\Lambda|}{2}\log(\sigma^2) - \frac{n|\Lambda|}{2}\log(2\pi) \ ;$$

and

$$E_{\nu_l}\left(\log q(y_1^n|\beta_1^n,\theta_p)+\log q(\theta_p)\right) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} E_{\nu_l}(|y_i-K_p^{\beta_i}\alpha|^2) - \frac{n|\Lambda|}{2}\log(\sigma^2) - \frac{a_p}{2}\left(\frac{\sigma_0^2}{\sigma^2}-\log(\sigma^2)\right) - \frac{1}{2}\left((\alpha-\mu_p)^t\Gamma_p^{-1}(\alpha-\mu_p)-\log(|\Gamma_p|)\right)+\text{Const.}$$

Define the following three statistics, where the last two involve integrating out the unobserved deformations based on the current posterior ν_l.

$$\left\{\begin{array}{rcl} [Y^tY]_l & = & \frac{1}{n}\sum_{i=1}^{n}|y_i|^2 \\ \left[\left(K_p^\beta\right)^t Y\right]_l & = & \frac{1}{n}\sum_{i=1}^{n}\int \left(K_p^\beta\right)^t y_i\nu_{l,i}(\beta)d\beta \\ \left[\left(K_p^\beta\right)^t K_p^\beta\right]_l & = & \frac{1}{n}\sum_{i=1}^{n}\int \left(K_p^\beta\right)^t K_p^\beta\nu_{l,i}(\beta)d\beta\ . \end{array}\right. \tag{14}$$

Now setting the derivatives in α and σ to zero we get

$$\left\{\begin{array}{l} \frac{n}{\sigma^2}\left(\left[\left(K_p^\beta\right)^t K_p^\beta\right]_l\alpha-\left[\left(K_p^\beta\right)^t Y\right]_l\right)+\Gamma_p^{-1}(\alpha-\mu_p)=0 \\ \\ \frac{1}{2\sigma^4}\left(n\left([Y^tY]_l+\alpha^t\left[\left(K_p^\beta\right)^t K_p^\beta\right]_l\alpha-2\alpha^t\left[\left(K_p^\beta\right)^t Y\right]_l\right)+a_p\sigma_0^2\right)-\frac{n|\Lambda|+a_p}{2\sigma^2}=0 \end{array}\right. \tag{15}$$

Solving for the unknowns in ((15)) yields

$$\left\{\begin{array}{rcl} \alpha & = & \left(n\left[\left(K_p^\beta\right)^t K_p^\beta\right]_l+\sigma^2\Gamma_p^{-1}\right)^{-1}\left(n\left[\left(K_p^\beta\right)^t Y\right]_l+\sigma^2\Gamma_p^{-1}\mu_p\right) \\ \\ \sigma^2 & = & \frac{1}{n|\Lambda|+a_p}\left(n\left([Y^tY]_l+\alpha^t\left[\left(K_p^\beta\right)^t K_p^\beta\right]_l\alpha-2\alpha^t\left[\left(K_p^\beta\right)^t Y\right]_l\right)+a_p\sigma_0^2\right)\ . \end{array}\right. \tag{16}$$

This system can be solved iteratively in α and σ initialized with the current values α_l, σ_l.

In the absence of training images, i.e. $n=0$ the updates of equations (13), (16) yield

$$\theta_g=\Gamma_0^g \text{ and } \theta_p=(\mu_p,\sigma_0^2).$$

4.2. **Fast approximation with modes.** The expressions in the M step require the computation of expectations with respect to ν_l which have no simple form. A classic approximation consists of replacing the distribution $\nu_{l,i}$ by the Dirac law $\nu_{l,i}^*(d\beta_i)=\delta_{\beta_i^*}$ with:

$$\beta_i^*=\arg\max_\beta \log q(\beta|\alpha_l,\sigma_l,\Gamma_{g,l},y_i)=\arg\min_\beta\left[\frac{1}{2}\beta^t\Gamma_{g,l}^{-1}\beta+\frac{1}{2\sigma_l^2}|y_i-K_p^\beta\alpha_l|^2\right]\ , \tag{17}$$

recalling that $K_p^\beta\alpha=(\mathbf{K_p}\beta)I_\alpha$. This is a standard template matching problem, with observations y_i, template $K_p^\beta\alpha_l$, noise level σ_l and covariance matrix $\Gamma_{g,l}$. The approximation can be formally interpreted as constraining the maximization over ν in equation (9) to Dirac delta functions. This is only formal since whatever distribution μ is chosen not all

point mass distributions on β will have a density with respect to μ, so that the entropy term becomes infinite. In any event this approximation reduces the EM algorithm to an iterative maximization of $q(y_1^n, \beta_1^n, \theta)$ in θ and β. It is important however to note that the matching problem is a difficult non-linear optimization problem and we can at most expect to obtain local minima.

As indicated in the introduction the proposed approximation is similar to the iterative maximization proposed in [10]. However as seen from the example below, the true EM iterations can provide better estimates, indeed it appears that in some cases the iterative maximization is not consistent and can yield wrong estimates of the template. Consider a simple 1d setting where the deformations are a discrete set of translations. Here the expectations with respect to the posterior can be computed explicitly.

Example: Let Λ be the set of integers between $[-L, L]$. Let $I_0(x) = 1_{[-K,K]}$ for some $K < L$ and consider a discrete set of translations $\tau \in [-\Delta, \Delta]$, with $\Delta < L - K$. The generative model involves sampling a random translation, shifting the template to $I_0(x+\tau)$ and adding i.i.d noise of variance σ^2 at every location between $[-L, L]$. In figure 1 we show the result of full EM vs. iterative maximization for a range of values of K, L and the variance σ^2 of the observation noise.

It emerges from this simple experiment that for high noise levels the template is poorly estimated especially near the boundaries between the two intensity levels. This seems to be due to the fact that high noise levels produce locations with high intensity gradients between neighboring pixels and the optimal shift tries to move the template to fit them. Consequently the values at the boundaries are over estimated for the higher values and under estimated for the lower values.

4.3. **Relation to PCA estimation.** Using the approximation with modes, $[\beta\beta^t]_l$ is simply the empirical covariance of the deformations $(\beta_i^*)_{1\leq i\leq n}$ estimated with respect to the observations. A common approach found in the literature is to perform a PCA analysis on this covariance matrix and generate a representation of the Gaussian distribution governing the deformations in terms of the components with non-zero eigenvalues. This corresponds to setting the prior weight $a_g = 0$. This approach suffers from several drawbacks.

- **Lack of regularization of the covariance estimates.** With small training samples the empirical covariance matrix is degenerate and many of the eigenvalues are very small and inaccurate. An ad-hoc solution which is often employed is to threshold the eigenvalues at some small positive value. However using a well defined prior one has better control on the direction of the non-degenerate correction of the degenerate estimate. For example in our case the correction is with Γ_g^0, which captures some apriori assumptions on the smoothness of the deformation fields.
- **Degenerate iterations.** Most algorithms do not go beyond the estimation of the empirical covariance. However, this cannot provide a consistent statistical scheme since the optimal β's have been computed with a wrong initial covariance structure. Note that even when the procedure is iterated interchanging the estimation of the template and of the geometric covariance matrix, in the absence of regularization the deformations are constrained to always lie in the subspace spanned by the

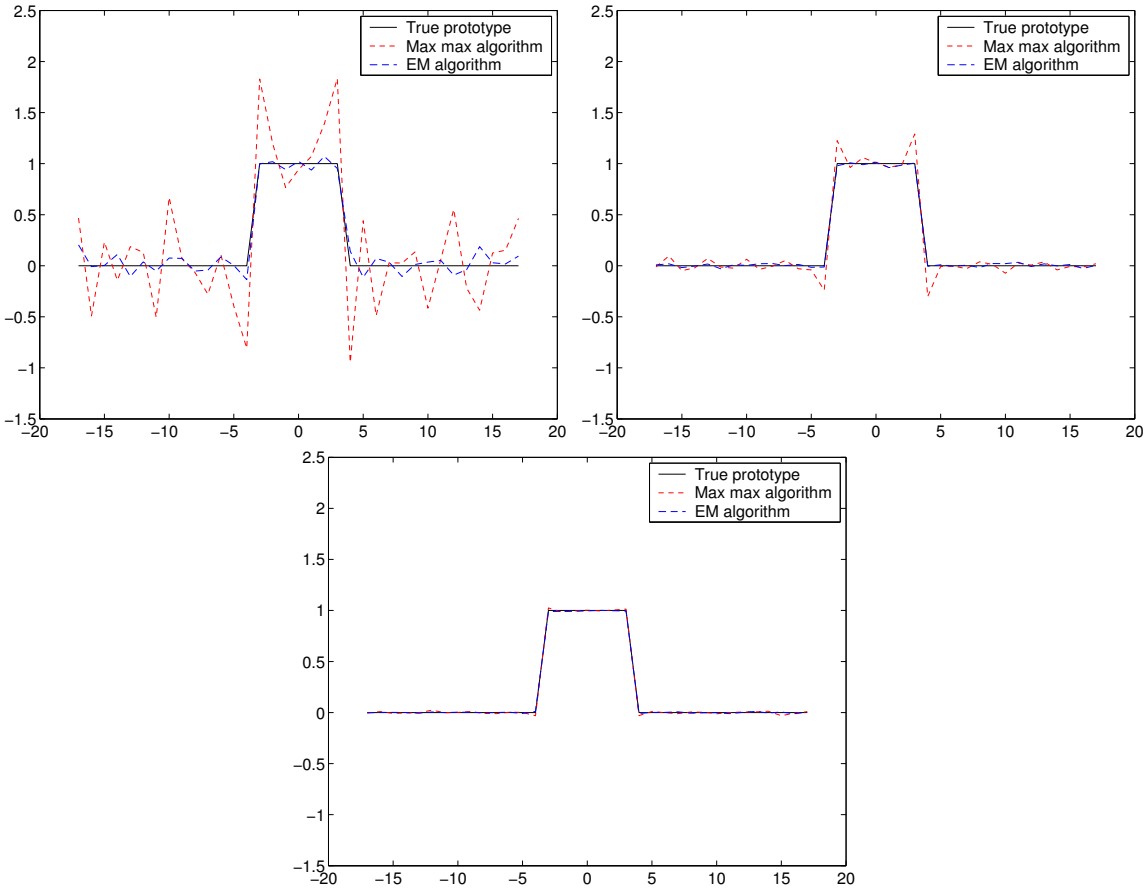

FIGURE 1. Comparison between the iterative maximization algorithm and the full EM algorithm. Top right $\sigma = 1$, top left $\sigma = .5$, bottom $\sigma = .1$. Red dashed line - Template estimated with iterative maximization. Blue dashed line - template estimated with EM algorithm. Solid line - true template.

initial covariance structure thus limiting the proper estimation of the photometric template.

5. Mixtures of deformable template models

In many situations object classes can not be described as smooth deformations of one template. Classes are often defined as a combination of structures with distinct topological characteristics. As a simple example consider hand written '2's with a loop in the base and without, or faces with or without glasses. It is therefore natural to extend the model framework to include mixtures of deformable templates and extend the EM framework to estimate the models of each mixture component as well as the weights of the different mixtures. We introduce the following notation:

$$\eta = (\theta, \rho) \text{ with } \theta = (\theta^\tau)_{1 \leq \tau \leq T} \text{ and } \rho = (\rho(\tau))_{1 \leq \tau \leq T} \ ,$$

where T denotes the number of model components, and $\rho = (\rho_\tau)_{1\leq\tau\leq T}$ are the mixture coefficients. As before each model θ^τ is composed of a photometric part θ_p^τ and a geometric part θ_g^τ.

For each observation y_i, we consider the pair $\xi_i = (\beta_i, \tau_i)$ of unobserved variables. Imposing a prior on the probability distribution ρ we define an extended generative model as follows. Draw ρ according to an *a priori* law ν_ρ and independently generate models (θ^τ) for each $1 \leq \tau \leq T$ as described in section 2.4. Then for each observation y_i a component τ_i is drawn from ρ. The remaining variables are drawn as described in section 2.4. This is summarized below:

$$\begin{cases} \rho \sim \nu_\rho \\ \theta = (\theta_g^\tau, \theta_p^\tau)_{1\leq\tau\leq T} \sim \otimes_{\tau=1}^T (\nu_g \otimes \nu_p) \mid \rho \\ \tau_1^n \sim \otimes_{i=1}^n \rho \mid \eta = (\theta, \rho) \\ \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_g^{\tau_i}) \mid \eta,\ \tau_1^n \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(z_{\beta_i} I_{\alpha_i}, \sigma_{\tau_i}^2 Id_\Lambda) \mid \beta_1^n,\ \eta,\ \tau_1^n \end{cases}$$

with $\theta_p^\tau = (\alpha_\tau, \sigma_\tau^2)$, $I_{\alpha_i} = \mathbf{K_p}\alpha_{\tau_i}$ and $z_{\beta_i} = \mathbf{K_g}\beta_i$ for all $1 \leq i \leq n$.

For the *a priori* law ν_ρ we choose the Dirichlet distribution with density

$$\mathcal{D}(a_\rho) : \nu_\rho(\rho) \propto \left(\prod_{\tau=1}^T \rho(\tau)\right)^{a_\rho},$$

with parameter a_ρ.

We seek the MAP estimate given the observed images - y_1^n

$$\hat{\eta} = \operatorname*{argmax}_\eta q(\eta | y_1^n)\ .$$

The EM update, $\eta_l \to \eta_{l+1}$, has the same form as before

E Step: Compute the posterior law on $(\beta_i, \tau_i), i = 1, \ldots, n$ as a product of the following distributions which have a density in β for each τ and are discrete in τ for each β:

$$\nu_{l,i}(\beta, \tau) = \frac{q(y_i|\beta, \alpha_{\tau,l}) q(\beta|\Gamma_{g,l}^\tau) \rho_l(\tau)}{\sum_{\tau'} \int q(y_i|\beta', \alpha_{\tau',l}) q(\beta'|\Gamma_{g,l}^{\tau'}) \rho_l(\tau') d\beta'}$$

M Step:

$$\eta_{l+1} = \operatorname*{argmax}_\eta E_{\nu_l(d\xi_1^n)}(\log q(\eta, \beta_1^n, \tau_1^n | y_1^n)).$$

5.1. **Details for the M step.** Let

$$\nu_{l,i}(\tau) = \int \nu_{l,i}(\beta, \tau) d\beta,$$

be the posterior marginal on component τ_i given the current parameter estimates. Letting $n_{\tau,l}$ denote the weight of component τ,

$$n_{\tau,l} = \sum_{i=1}^{n} \nu_{l,i}(\tau),$$

the update of the component parameters takes the standard form of EM updates for mixture models in the presence of a prior:

$$\rho_{l+1}(\tau) = \frac{n_{\tau,l} + a_\rho}{n + T a_\rho}.$$

Now define the empirical covariance matrix for each component:

$$[\beta\beta^t]_{l,\tau} = \frac{1}{n_{\tau,l}} \sum_{i=1}^{n} \int (\beta\beta^t) \nu_{l,i}(\beta, \tau) d\beta.$$

This corresponds to the definition in equation (12) where the sample y_i contributes $\nu_{l,i}(\tau)$ to component τ and then the unobserved deformation variables β are integrated out according to $\nu_{l,i}(\beta|\tau)$. With the prior regularization we obtain as in equation (13)

$$\theta^{\tau}_{g,l+1} = \frac{1}{n_{l,\tau} + a_g} (n_{l,\tau}[\beta\beta^t]_{l,\tau} + a_g \Gamma^0_g) \ .$$

5.1.1. *Photometry update.* With the same reasoning as above, adapting the definitions in equation (14) to each cluster we define

$$\begin{cases} [Y^t Y]_{l,\tau} & = & \frac{1}{n_{l,\tau}} \sum_{i=1}^{n} |y_i|^2 \nu_{l,i}(\tau) \\ \left[\left(K_p^\beta\right)^t Y \right]_{l,\tau} & = & \frac{1}{n_{l,\tau}} \sum_{i=1}^{n} \int \left(K_p^\beta\right)^t y_i \nu_{l,i}(\beta, \tau) d\beta \\ \left[\left(K_p^\beta\right)^t K_p^\beta \right]_{l,\tau} & = & \frac{1}{n_{l,\tau}} \sum_{i=1}^{n} \int \left(K_p^\beta\right)^t K_p^\beta \nu_{l,i}(\beta, \tau) d\beta \ . \end{cases}$$

The final update equations are the same as in (16)

$$\begin{cases} \alpha_\tau & = & \left(n_{l,\tau} \left[\left(K_p^\beta\right)^t \left(K_p^\beta\right) \right]_{l,\tau} + \sigma_\tau^2 (\Gamma_p)^{-1} \right)^{-1} \left(n_{l,\tau} [K_p^t Y]_{l,\tau} + \sigma_\tau^2 (\Gamma_p)^{-1} \mu_p \right) \\ \sigma_\tau^2 & = & \frac{1}{n_{l,\tau}|\Lambda| + a_p} \left(n_{l,\tau} \left([Y^t Y]_{l,\tau} + (\alpha_\tau)^t \left[\left(K_p^\beta\right)^t K_p^\beta \right]_{l,\tau} \alpha_\tau - 2(\alpha_\tau)^t \left[\left(K_p^\beta\right)^t Y \right]_{l,\tau} \right) + a_p \sigma_0^2 \right) \ , \end{cases}$$

which again can be solved iteratively for each cluster τ starting with the previous values $\alpha_{\tau,l}, \sigma^2_{\tau,l}$.

5.2. **Fast approximation with modes.** For efficiency in training we again consider an approximation of the probability distribution $\nu_{l,i}(\beta,\tau)$, using the mode over the β parameter. Specifically for each component τ let $\beta^*_{i,\tau}$ maximize the conditional distribution on β:

$$\beta^*_{i,\tau} = \arg\max_{\beta} \log q(\beta|\alpha_{\tau,l},\sigma_{\tau,l},\Gamma^\tau_{g,l},y_i) = \arg\min_{\beta}\left\{\frac{1}{2}\beta^t R^\tau_{g,l}\beta + \frac{1}{2\sigma^2_{l,\tau}}|y_i - K^\beta_p\alpha_{\tau,l}|^2\right\},$$

where $R^\tau_{g,l} = (\Gamma^\tau_{g,l})^{-1}$. We then approximate the joint posterior on (β_i,τ_i) as a discrete distribution concentrated at the T points $\beta^*_{i,\tau}$ with weights given by

$$w_l(\tau) = \frac{q(y_i|\beta^*_{i,\tau},\alpha_{\tau,l})q(\beta^*_{i,\tau}|\Gamma^\tau_{g,l})\rho_l(\tau)}{\sum_{\tau'} q(y_i|\beta^*_{i,\tau'},\alpha_{\tau',l})q(\beta^*_{i,\tau'}|\Gamma^{\tau'}_{g,l})\rho_l(\tau')}. \tag{18}$$

Thus we keep the weighting on the clusters after approximating each conditional distribution on the deformation parameters with a delta distribution at the mode. It is of interest to note that this procedure can also be viewed as the iterative maximization of a well defined function. Assume that the reference measure $\mu(d\beta,d\tau)$ in equation (10) is a product of Lebesgue measure in β and the counting measure on $\{1,\ldots,T\}$. Write the entropy term as

$$\int \nu(\beta,\tau)\log\nu(\beta,\tau)\mu(d\beta,d\tau) = \int \nu(\beta|\tau)\log\nu(\beta|\tau)\nu(\tau)d\mu(d\beta,d\tau) \tag{19}$$

$$+\int \nu(\tau)\log\nu(\tau)\mu(d\tau). \tag{20}$$

We formally restrict the maximization in ν to the set of distributions $\mathcal{P}$ of the form

$$\nu(d\beta,d\tau) = \sum_\tau w(\tau)\delta_\tau\otimes\delta_{\beta_\tau},$$

i.e. a weighted sum of Dirac delta functions at points β_τ. The first term on the right hand of equation (19) is infinite and is ignored. The second term is the nentropy

$$H(w) = -\sum_\tau w(\tau)\log w(\tau),$$

of the marginal on τ i.e. the discrete measure defined by the weights $w(\tau),\tau=1,\ldots,T$. Now the iterative maximization in equation (10) becomes

$$\max_\theta \max_{\nu\in\mathcal{P}} \int \log q(y,\beta,\tau|\theta)\nu(d\beta,d\tau) - H(w)$$

$$= \max_\theta \max_{\beta_1,\ldots,\beta_T,w} \sum_\tau \log q(y,\tau,\beta_\tau|\theta)w(\tau) - H(w). \tag{21}$$

Maximizing first in β_τ does not depend on the weights w, and is done separately for each term $q(y,\tau,\beta_\tau)$. This yields β^*_τ as defined above. Then maximizing in the weights yields

$$w_c(\tau) = q(\tau|\beta^*_\tau,y,\theta_c),$$

which is computed in equation (18). To summarize the fast EM iterations proposed here are equivalent to the iterative maximization of equation (21).

6. Experiments

In this section we illustrate some of the issues raised above in the simple context of images of handwritten digits. In this context it is possible to compare various model settings in terms of classification rates, although our goal here is not to obtain optimal results. The experiments are performed on the US-POSTAL data base which contains a training data set with about 7000 (16×16)-pixels handwritten digit images and a test set with about 2000 images. For information on a number of discriminative approaches to the classification of these digits see [13]. In figure 2 we present the images of 40 digits in each class which are used for training.

After estimating the parameters of a deformable template model for each class, classification should be performed by computing the maximum posterior on class given the image. The likelihood term involves an integral over the unobserved deformation variables which is difficult to compute, and is replaced again by the mode. Specifically for each class c let I_{α_c} be the estimated template, σ_c the estimated variance, and $R_{g,c}$ the estimated inverse geometric covariance matrix. Define

$$U_c(\beta) = -\frac{1}{2}|y - z_\beta I_{\alpha_c}|^2 - \frac{|\Lambda|}{2}\log(2\pi\sigma_c^2) + \frac{1}{2}\beta^t R_{g,c}\beta,$$

Assuming a flat prior on the 10 classes set

$$\hat{c} = \operatorname*{argmax}_c U_c(\beta_c^*), \text{ where } \beta_c^* = \operatorname*{argmax}_\beta U_c(\beta).$$

The justification of this form as an approximation of the marginal likelihood is provided in the appendix.

All images are rescaled to have intensities in the interval $[0, 2]$, and it is assumed that observation domain is the square $[-1, 1] \times [-1, 1]$. The template domain is infinite but we restrict the control points to the larger square $[-1.5, 1.5] \times [-1.5, 1.5]$. We use radial Gaussian kernels to represent the template and the deformation fields:

$$K_p(x, y) = \exp\left[\frac{-\|x - y\|^2}{2\sigma_p^2}\right], \; K_g(x, y) = \exp\left[\frac{-\|x - y\|^2}{2\sigma_g^2}\right].$$

The width of the kernels, i.e. σ_p, σ_g, depends on the overall smoothness we expect for the template and deformation fields respectively. The number of control points depends on the choice of the width parameters. At each iteration of the approximate EM algorithm the deformations β_i^* are computed using a straight-forward gradient descent algorithm on the cost function given in (17).

6.1. The estimated templates.

In figure 3 we show the templates of the 10 classes estimated with 20 images per class - the first 20 images of each row in figure 2. For the prior distribution on templates we choose a mean $\mu_p = 0$, which is constant at the

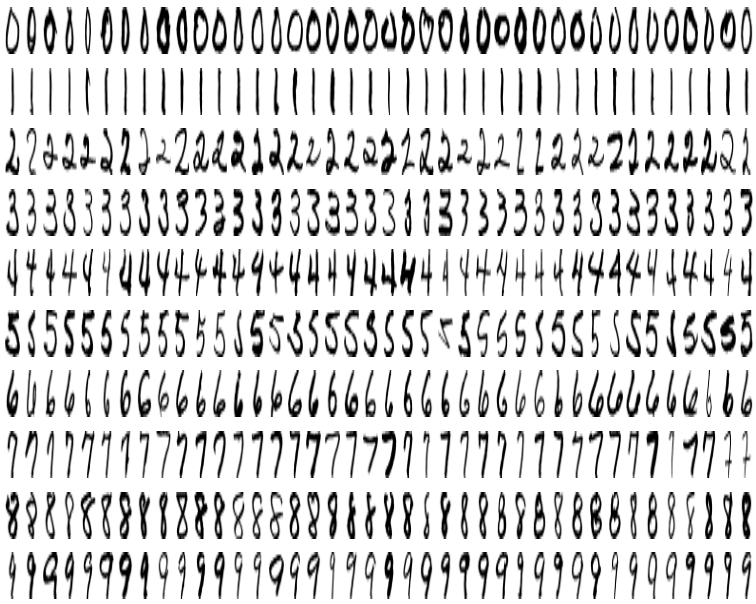

FIGURE 2. Training set : 40 images per class

'background' value, and a covariance Γ_p given in equation (5) and determined by the interpolating kernel K_p.

In this setting the first iteration of the EM algorithm yields deformations $\beta_i^* \equiv 0$ so that the resulting estimated template is the simple mean of the training images. These

FIGURE 3. Left: Simple average images, Middle and right: Estimated prototypes (20 images/class), $\sigma_g = 0.2$ (Middle), $\sigma_g = 0.3$ (Right)

means are shown in the left panel of figure 3. They are blurred because of the geometric variability within each class. As the iterations proceed the estimated prototypes present higher contrast thanks to the nonrigid registration which enables better fits. In the middle and right panels of figure 3 we present the estimated prototype for two different values of the width parameter σ_g, (0.2 and 0.3). Note that the templates seem similar to the initial means modulo some deblurring or contrast enhancement.

6.2. **The photometric variance.** The variance σ in the data model is estimated and evolves throughout the EM iterations. Initially the estimated variance is influenced both by

photometric variations but even more by the geometric variability, which is not accounted for. As the estimates evolve the variation in photometry at a given pixel is less and less a function of geometric variability and reduces to the inherent photometric noise in the data. Thus we can see in figure 4 how the estimated variance decreases with iteration for all classes. Those classes ('2','4')which are more heterogeneous and perhaps require more than one template exhibit higher final variance estimates.

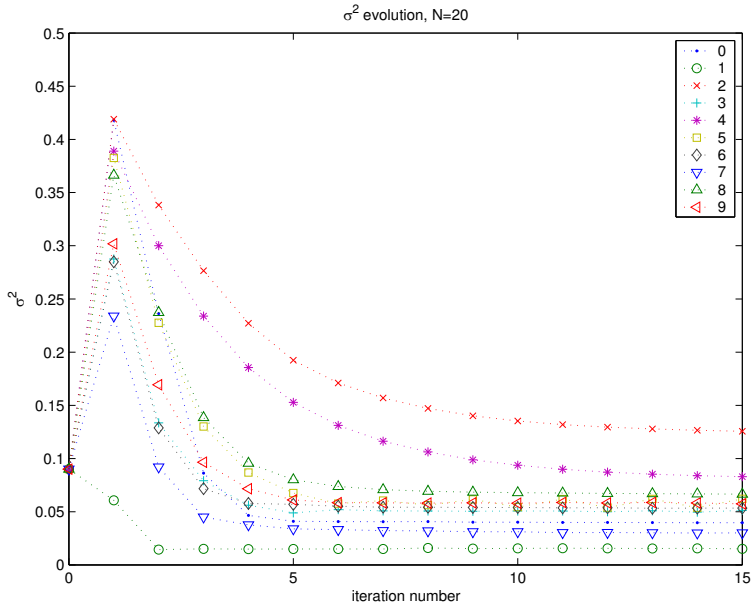

FIGURE 4. Evolution of estimates of σ for each class with iteration.

6.3. **The estimated geometric distribution.** Below are data pointing to the effect of the hyper-parameters of the geometric prior - a_g (the constant in the Wishart prior on covariance matrices (4)) and σ_g the width of the kernel K_g. Table 1 shows the error rate on a set of 1000 images randomly sampled from the training set outside the 200 used for training. [1] There is a clear decrease of performance as the weight a_g decreases. This may

a_g	0.01	0.1	1	5	10	∞
$\sigma_g = 0.3$	13.1	3.3	3.1	3.1	3	3.4
$\sigma_g = 0.2$	15.5	10.1	8.3	4.5	8.4	8.3

TABLE 1. Error rate for different values of σ_g (rows) and of the geometric *prior* weight a_g (columns). The training set contains 20 images per class, 20 iterations of EM were performed.

[1] It is well known that the test set of the USPOSTAL database contains more difficult images than the training set which explains the high classification rates we are able to achieve with a very small training sample set. Results on the full test set are presented below.

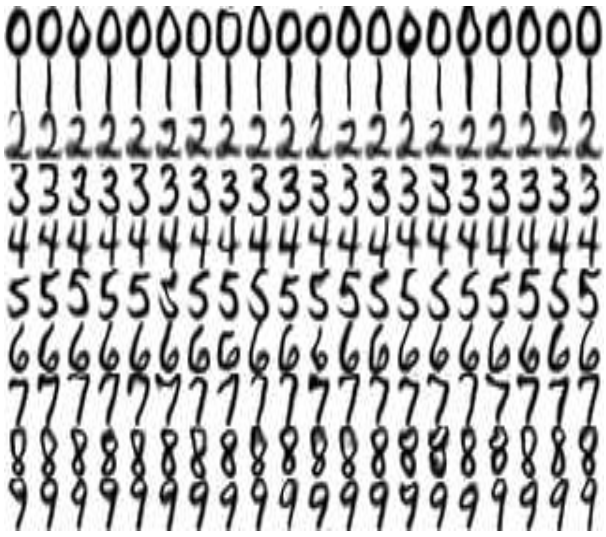

FIGURE 5. 20 examples of synthesized images from each class using estimated template and geometric covariance.

be due to the fact that with a small number of examples and a relatively flat prior, the estimated covariance matrix Γ_g is quasi-singular (see equation (13)). This restrains the possible displacements in a low dimensional subspace. On the other hand, as a_g increases there is essentially no estimation of the geometric covariance structure, i.e. $\Gamma_g = \Gamma_g^0$, the center of the covariance prior. Note that all values of a_g are admissible but as can be expected, the best performance is obtained for an intermediate value of this parameter. The effect of a_g can be important when the choice of other parameters is not optimal (see $\sigma_g = 0.2$).

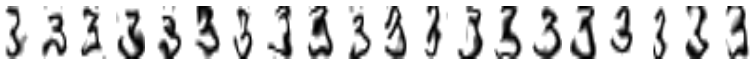

FIGURE 6. Synthesizing 3's with the estimated covariance class 2.

To illustrate the form of the geometric distribution estimated by the algorithm in figure 5 we show 20 synthesized examples of each class using the estimated photometric prototype and the estimated geometric covariance. By contrast, in figure 6, we show simulations using the estimated prototype of class 3 with the geometric distribution of class 2. This should be compared to the 3rd row of figure 5. The deformed 3's are not at all realistic implying that the estimated geometric covariances are non-trivial, and differ significantly from one class to the other.

6.4. **Classification rates and number of iterations.** As mentioned above, the EM algorithm for the computation of the MAP estimator of the photometric and geometric parameters leads to a natural and well defined iterative procedure. As the iterations

proceed the model fit to the data improves. This was shown earlier in terms of the reduction of the variance estimate (see figure 4) and is seen here in the decrease in error rates using the successive models produced at each iteration, (see table 2). The results are provided for a fixed value of all the hyper-parameters and a training set of 20 images per digit.

Nb of iter	1	2	3	4	5	10	15	20
Error rate	14.9	9.3	7.4	5.5	4.5	4.2	4	3.3

TABLE 2. Error rate while increasing the number of iterations, same test set as in table 1, 20 training examples per class, $a_g = 0.1$, $\sigma_g = 0.3$

6.5. **Mixture models.** Here we consider the computation of a mixture of deformable models for each class. In figure 7, we show the two components per class estimated with 40 training examples per class. It appears that for each class, the two chosen prototypes

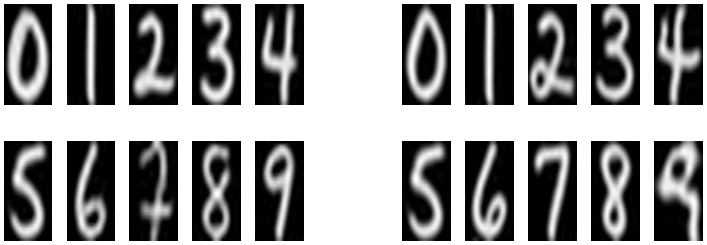

FIGURE 7. Templates of the 2 components (40 images per class, 20 iterations, 2 components per class).

correspond to a meaningful clustering of the training data (displayed in figure 2). Note in particular the case of class 2 with two topologically different versions (with and without loop), the European prototype appearing for class 7 or the 'broken' 8.

It is harder to visualize the geometric distribution estimated for each component. For two classes 0 and 7, and for each of the two components we display in figure 8 the evolution of the symmetrized Kullback distance between $N(0, \Gamma_{g,l})$ and $N(0, \Gamma_g^0)$, i.e. the value

$$d(\Gamma_g^0, \Gamma_g) \doteq (K(\mathcal{N}(0, \Gamma_g^0), \mathcal{N}(0, \Gamma_g))/2 + K(\mathcal{N}(0, \Gamma_g), \mathcal{N}(0, \Gamma_g^0))/2)^{1/2} .$$

The estimated covariance matrix Γ_g clearly moves away from the prior and the final distance is fairly different between the two components of each class. Note however that the first component of class seven shown in figure 7(the European seven) is based only on two sample points in the training set (coincidentally, the last two in the corresponding row of figure 2). In this small sample case the Bayesian estimate is strongly biased towards the initial Γ_g^0 as is seen by the plot on the lower left-hand panel.

This behavior is of particular interest in its ability to reveal two different geometric behaviors within a given population with a *homogeneous* photometric behavior. In other words within the framework of the mixture model we are able to identify a situation where the photometric template is unique but the distribution of the deformations is modeled as

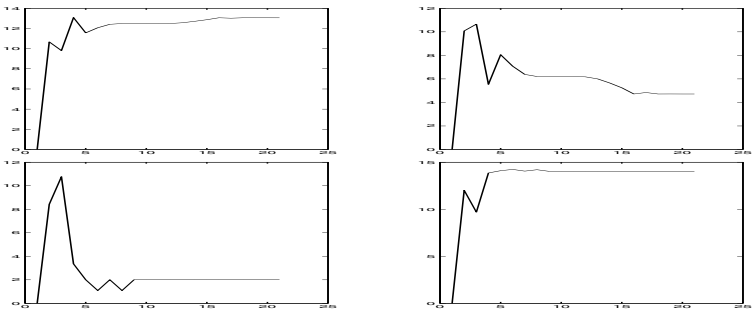

FIGURE 8. Evolution of the symmetrized Kullback distance between the current value of Γ_g and the prior center Γ_0. (40 observations per class, 2 components, 20 iterations). Top row: class 0, Bottom row: class 7. Left column: first component, Right column: second component.

a mixture of Gaussians as opposed to a simple Gaussian. Moreover since at times it can be difficult to make a clear separation between the photometric part and the geometric part of the variation, the ability to handle both simultaneously is an important feature.

FIGURE 9. Top: Synthesized 2's with template from second component of figure 7 and proper covariance. Bottom: Same template using covariance matrix of other 2 component.

In figure 9 we show synthesized images from class 2 with the template from the second component of figure 7 (with a loop at the bottom). The top shows samples with the correct geometric parameters and the bottom shows samples with the geometric parameters of the other component. Although the samples produced in the second row look like 2's they are not as natural looking as those of the first row, again indicating a non-trivial difference between the geometric covariance estimated for each of the components.

6.6. **Results on the full test set.** We present some error rates on the original test set as a function of the number of training images and the number of mixture components. Note that the largest training set we use has 100 digits per class and is a small fraction of the full training set of over 7000 images. The results are comparable to those obtained employing discriminative methods and trained with the full training set. Note that here classification is performed by simply choosing the most likely class, based on the estimated models, no decision boundaries are precomputed in training. The results are summarized in table 3, along with reported results for nearest neighbor and discriminative methods from [13]. The misclassified digits are shown in figure 10.

Nb. of components	1	2	3	5	10
20 per class	6.58	6.13	5.28	9.57	9.72
40 per class	6.43	5.18	5.38	5.23	7.075
100 per class	9.42	4.98	4.58	5.13	**4.136**

Neural network	Least square nearest Nbrs.	Tangent distance nearest nbrs.
4.9	5.5	2.6

TABLE 3. Top: Error rates for different numbers of components (column) and different numbers of training images (rows) per class. $a_g = 0.1$, $\sigma_g = 0.3$, 20 iterations. Bottom: Error rates using a neural network, nearest neighbors, and tangent distance nearest neighbors with full training set, as reported in [13].

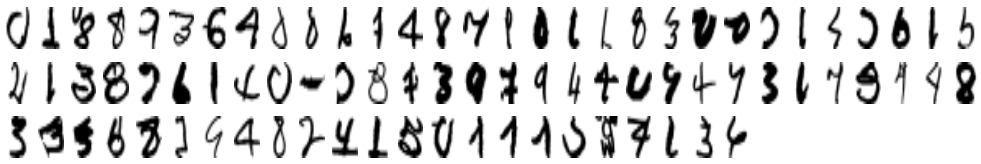

FIGURE 10. Misclassified digits from test set, error rate 4.136%. 10 components per class.

7. Discussion

We have provided a coherent statistical formulation for dense deformable template models and described an approximation to maximum-posterior estimation. The results on likelihood based classification of handwritten digits, with very small training sets, demonstrate that this formulation is of practical use. One major drawback of the current setting is the additive noise model. This model does not allow for photometric variations due to changes in lighting on an object or simply a change in the gray level map. Moreover it is not possible to use this type of data model to formulate a credible background model for observations off the object. In previous work [1], [2], [4] we have overcome this limitation using local photometrically invariant oriented edge features as a substitute for the original gray levels. The approach outlined here would extend easily to this alternative data model where instead of a template *image* we estimate template *probability maps* for each feature. Clearly some modifications would be required in the definition of the priors.

The deformation model employed here does not necessarily produce diffeomorphisms. This can create some difficulties such as behavior near the boundaries of the domain, and the need for a template defined on the entire plane. Using diffeomorphisms of the domain onto itself, as proposed in [19], would eliminate these problems perhaps yielding a more stable algorithm.

Another issue is the crude approximation to EM through the local modes of the posterior. As illustrated in the synthetic 1d example in section 4.2 this approximation can yield erroneous estimates. We are currently exploring the possibility of approximating the integrals with respect to the posterior on the deformations through various monte-carlo methods.

Appendix A. Proof of the consistency theorem for bounded prototypes

The proof of the theorem relies on several lemmas which control the expectation of the supremum over the parameter space of the positive part of the log-likelihood (Lemma 3), and the behavior of the likelihood at the boundary points of the parameter space (Lemma 4). We prove the theorem for the case with one deformable model. We emphasize that we are ignoring issues of identifiability by proving the convergence of the sequence of maximum-posterior estimates to the *set* Θ_*^R. (Refer to equation (7) for notation.)

Lemma 1. *Let $p < q$ be two integers and $F : \mathbb{R}^p \to \mathbb{R}^q$ be a C^1 mapping. Then for any compact subset C of $\mathbb{R}^p$, if $M = F(C)$ then*

$$\int \log^+(1/d(y, M))dy < \infty.$$

Proof. For any $\rho > 0$, denote $M_\rho = \{\ y \in \mathbb{R}^q \mid d(y, M) \leq \rho\ \}$. For any $\rho > 0$, there exists a finite set $\Lambda_\rho \subset C$ such that $|\Lambda_\rho| \leq K\rho^{-p}$ and $C \subset \cup_{x\in\Lambda_\rho} B(x, \rho)$, with K fixed. Denoting $\tau = \sup_C \|dF\|$, we get

$$M \subset \cup_{x\in\Lambda_\rho} B(F(x), \tau\rho) \text{ and } M_\rho \subset \cup_{x\in\Lambda_\rho} B(F(x), (\tau + 2)\rho).$$

Thus there exists a constant K' independent of ρ such that we have for $V(M_\rho) = \int_{M_\rho} dy$:

$$V(M_\rho) \leq K'\rho^{q-p}\,.$$

Let $0 < s < 1$ and $\rho_n = s^n$ for any $n \geq 0$. Then

$$\int \log^+(\frac{1}{d(y, M)})dy \leq \sum_{n=0}^{\infty} \log(\frac{1}{\rho_{n+1}})(V(M_{\rho_n}) - V(M_{\rho_{n+1}})) \leq \log(\frac{1}{\rho_1}) \sum_{n=0}^{\infty} V(M_{\rho_n}) < \infty$$

where the second inequality comes from the Abel transformation. □

Lemma 2. *Let p, q, F be as above and assume that (i) $\sup_{\mathbb{R}^p} \|dF\| < \infty$. (ii) there exist constants $a > 0, b$ such that $|F(x)| \geq a|x| + b, \forall x \in \mathbb{R}^p$. Let ν be a density on $\mathbb{R}^q$ with exponentially decaying tails (i.e. $\sup_{|y|\geq\lambda} \nu(y) \leq a_1 \exp(-a_2\lambda)$ for any $\lambda \geq 0$ and some $a_1, a_2 > 0$) and let $M = F(\mathbb{R}^p)$. Then*

$$\int \log^+ \frac{1}{d(y, M)} \nu(y)dy < \infty.$$

Proof. Let $\mathcal{T}$ be the integer lattice in $\mathbb{R}^p$, and let $B_t = \{x \in \mathbb{R}^p; |x-t|_\infty \leq 1/2\}$. Applying the previous Lemma and using (i) we have that

$$\int \log^+ \frac{1}{d(y, F(B_t))} dy < C_1,$$

for all $t \in \mathcal{T}$. Let $\overline{F(B_t)} = \{y : d(y, F(B_t)) \leq 1\}$. It is clear that

$$\int \log^+ \frac{1}{d(y, F(B_t))} \nu(y) dy < C_1 \max_{y \in \overline{F(B_t)}} \nu(y).$$

Let $\nu_\tau = \max_{|y|_\infty \geq \tau} \nu(y)$, then due to (ii) $\max_{\overline{F(B_t)}} \nu(y) \leq \nu_{a|t|_\infty + b'}$, for some constant b'. For any integer T, let $D_T = \cup_{|t| \leq T} B_t$. We have

$$1/d(y, F(D_T)) = \sup_{x \in D_T} 1/d(y, F(x)) \leq \sum_{|t| \leq T} \sup_{x \in B_t} 1/d(y, F(x)) = \sum_{|t| \leq T} 1/d(y, F(B_t)).$$

Consequently

$$\begin{aligned} \int \log^+ \frac{1}{d(y, F(D_T))} \nu(y) dy &\leq \sum_{j=1}^{T} \sum_{|t|_\infty = j} \int \log^+ \frac{1}{d(y, F(B_t))} \nu(y) dy \\ &\leq C_1 \sum_{j=1}^{T} j^p \nu_{aj+b'}. \end{aligned}$$

Due to the assumptions on ν this sum converges as $T \to \infty$. □

Lemma 3. *Assume that $dim_\beta < dim_y$ and let l be any integer sufficiently large for which $p \doteq l\, dim_\beta + dim_\alpha < q \doteq l\, dim_y$. Assume that $|I_b(x)| \geq a|x| + b$ for some positive constant a. Let $y_1^l = (y_1, \cdots, y_l)$ be i.i.d. under P. Assume $P(dy_1)$ has bounded density ν with exponentially decaying tails. Then for any $R > 0$*

$$E_{P(dy_1^l)} \left(\sup_{\theta \in \Theta^R} \left(\sum_{i=1}^{l} \log q(y_i|\theta) \right)^+ \right) < +\infty$$

Proof. Let $F : \mathbb{R}^p \to \mathbb{R}^q$ be defined by $F(\alpha, \beta_1^l) = (z_1 I_\alpha, \cdots, z_l I_\alpha)$ where $I_\alpha = \mathbf{K_p}\alpha + I_b$ and $z_i = \mathbf{K_g}\beta_i$. Since $\mathbf{K_g}$ and $\mathbf{K_p}$ are smooth we deduce that F is smooth. Define $\mathcal{I}_R^l = \{F(\alpha, \beta_1^l) \mid |\alpha| \leq R, \beta_i \in \mathbb{R}^{2k_g},\ \forall 1 \leq i \leq l\}$. We have

$$\sum_{i=1}^{l} \log q(y_i|\theta) \leq -\frac{l|\Lambda|}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} d(y_1^l, \mathcal{I}_R^l)^2. \tag{22}$$

where d denotes here the Euclidean distance on $(\mathbb{R}^{|\Lambda|})^l$. However, the right hand side is maximized for $\sigma^2 = d(y_1^l, \mathcal{I}_R^l)^2/(l|\Lambda|)$ so that there exists $K > 0$ such that

$$\sup_{\theta \in \Theta^R} \left(\sum_{i=1}^{l} \log q(y_i|\theta) \right)^+ \leq K + l|\Lambda| \log^+ \frac{1}{d(y_1^l, \mathcal{I}_R^l)}. \tag{23}$$

Now, since $\mathbf{K_g}$ is injective we have $|z| > \lambda|\beta|$ for some positive constant λ. It follows from the linear growth assumption on $I_b(x)$ that F satisfies condition (ii) of Lemma 2. Since the kernels K_g are bounded and have uniformly bounded derivatives it follows that dF is uniformly bounded over $\mathbb{R}^p$. We thus obtain the result from Lemma 2. □

Lemma 4. *Under the hypothesis of the previous lemma, we have*

(1) $P(dy_1^l)$ *almost surely,*

$$\lim_{m\to\infty} \sum_{i=1}^{l} \log q(y_i|\theta_m) = -\infty$$

for any sequence $\theta_m = (\alpha_m, \sigma_m^2, \Gamma_m)$, $m \in \mathbb{N}$, *such that* $\theta_m \in \Theta^R$, $\|\Gamma_m\| \to \infty$ *or* $\sigma_m^2 \to 0$ *or* $\sigma_m^2 \to +\infty$.

(2) *For any sequence* $\theta_m = (\alpha_m, \sigma_m^2, \Gamma_m)$, $m \in \mathbb{N}$, *such that* $\theta_m \in \Theta^R$, $\|\Gamma_m\| \to \infty$ *or* $\sigma_m^2 \to 0$ *or* $\sigma_m^2 \to \infty$ *we have*

$$\lim_{m\to\infty} E_{P(dy)}(\log q(y|\theta_m)) = -\infty$$

(3) *The mapping* $\theta \to E_{P(dy)} \log(q(y|\theta))$ *is continuous on* Θ^R *and* $\Theta_*^R \neq \emptyset$.

Proof. We prove the three points in order.

(1) Let $M_y = \max_i |y_i|$.

Case 1: $\|\Gamma_m\| \to \infty$. Due to the assumption on the growth of $I_b(x)$ and the fact that $\mathbf{K_g}$ is injective, there exist constants $A > 0, B$ such that $|(\mathbf{K_g}\beta)I_\alpha| \geq A|\beta| + B$. Given $w \geq 1$, we bound the marginal density on y_i given θ_m as follows:

$$\begin{aligned} q(y_i|\theta_m) &= \int q(y_i|\beta,\theta_m) q(\beta|\theta_m) d\beta \\ &\leq \frac{1}{(2\pi\sigma_m^2)^{|\Lambda|/2}} \int_{|\beta|>(w\cdot M_y - B)/A} \exp\left(-((w-1)M_y)^2/(2\sigma_m^2)\right) q(\beta|\theta_m) d\beta \\ &\quad + \max_{|\beta|\leq (w\cdot M_y - B)/A} q(\beta|\theta_m). \end{aligned}$$

As $\|\Gamma_m\| \to \infty$ the second integral goes to zero. As for the first integral, the integrand is maximized at $\sigma^2 = ((w-1)M_y)^2$, and hence it is bounded by $C \cdot (wM_y)^{-\Lambda}$. Thus

$$\limsup_m \sum_{i=1}^{l} \log q(y_i|\theta_m) \leq l(C - \Lambda \log w), i = 1, \ldots, n.$$

Since w can be arbitrarily large we obtain the result for the case $\|\Gamma_m\| \to \infty$.

Case 2: $\sigma_m \to 0$ or $\sigma_m \to \infty$. Fix an integer $M > 0$. There exists C_M such that for $|\beta| > C_M$, $|F(\alpha, \beta_1^l)| > 2M$ for any $|\alpha| \leq R$ with F as defined in Lemma 3. This implies that if $\sup_i |y_i| < M$, the distance of y_1^l to $\mathcal{I}_R^l$ is achieved at some $|\beta| < C_M$, i.e. $d(y_1^l, \mathcal{I}_R^l) = d(y_1^l, F(B))$, for some compact subset $B \subset \mathbb{R}^p$. Since $P(dy_1^l)$ has a continuous density and since $F(B)$ is compact and of dimension $p < q$,

$$P(|y_1^l|_\infty \leq M \text{ and } d(y_1^l, \mathcal{I}_R^l) > 0) = 1.$$

This is true for any integer M implying that $d(y_1^l, \mathcal{I}_R^l) > 0, a.s.$ Finally using the bound in equation (22), we get the required result whether $\sigma_m \to 0$ or $\sigma_m \to \infty$.

(2) Consider $f_m(y_1^l) = \inf_{n \geq m} \left(\sum_{i=1}^l \log q(y_i|\theta_n) \right)^-$. We deduce from (1) that a.s. $f_m(y_1^l)$ is a non decreasing and non negative sequence converging to $+\infty$. From the monotone convergence theorem we then have

$$E_{P(dy_1^l)}(f_m(y_1^l)) \to \infty \text{ and } E_{P(dy_1^l)} \left(\sum_{i=1}^l \log q(y_i|\theta_m) \right)^- \to \infty,$$

since $f_m(y_1^l) \leq \left(\sum_{i=1}^l \log q(y_i|\theta_m) \right)^-$. Concerning the positive part, if we denote $g_m(y_1^l) = \left(\sum_{i=1}^l \log q(y_i|\theta_m) \right)^+$, using the dominated convergence theorem, Lemma 3, and part 1), we get $E_{P(dy_1^l)} g_m(y_1^l) \to 0$. Finally, we have proved that $E_{P(dy_1^l)} \sum_{i=1}^l \log q(y_i|\theta_m) \to -\infty$ and point 2) follows immediately.

(3) The continuity statement is straightforward. If Θ_*^R is empty, any minimizing sequence θ_m satisfies (up to the extraction of a subsequence) $\theta_m \in \Theta^R$, $\|\Gamma_m\| \to \infty$ or $\sigma_m^2 \to 0$ or $\sigma_m^2 \to +\infty$ which is in contradiction with (2).

□

Proof of the consistency theorem for bounded prototypes. We follow the usual route of Wald's consistency proof, involving an adequate compactification of the parameter space Θ^R. Let $\overline{\Sigma_{2k_g}^+(\mathbb{R})} = \Sigma_{2k_g}^+(\mathbb{R}) \cup \{\infty\}$ be the one point Alexandrov compactification of $\Sigma_{2k_g}^+(\mathbb{R})$, $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$ and consider the compactification of Θ^R

$$\overline{\Theta}^R = \overline{B}_{\mathbb{R}^{k_p}}(0, R) \times \overline{\mathbb{R}}_+ \times \overline{\Sigma_{2k_g}^+(\mathbb{R})}$$

where $\overline{B}_{\mathbb{R}^{k_p}}(0, R)$ is the closed ball in $\mathbb{R}^{k_p}$ of radius R. Let l be as in Lemma 3. It is sufficient to check that for any point $\theta_\infty \in \overline{\Theta}^R$ for which $\delta(\theta_\infty, \Theta_*^R) \geq \epsilon$, there exists an open set U such that

$$\frac{1}{l} E_{P(dy_1^l)}(\sup_{\theta \in U \cap \Theta^R} \sum_{i=1}^l \log q(y_i|\theta)) < v(R)\,. \tag{24}$$

Let $(U_h)_{h \geq 0}$ be a non increasing sequence of open sets for which $\cap_{h \geq 0} U_h = \{\theta_\infty\}$, and define $f_h(y_1^l) = \frac{1}{l} \sup_{\theta \in U_h} \sum_{i=1}^l \log q(y_i|\theta)$, which is a non increasing sequence. If $\theta_\infty \in \Theta^R$, then from the continuity of $\theta \to \sum_{i=1}^l \log(q(y|\theta))$ for every $\theta \in \Theta^R$ and from Lemma 3, we deduce (using the monotone convergence theorem) that since $\theta_\infty \in \Theta^R \setminus \Theta_*^R$,

$$E_{P(dy_1^l)}(f_h(y_1^l)) \to \frac{1}{l} E_{P(dy_1^l)}(\sum_{i=1}^l \log q(y|\theta_\infty)) < v(R).$$

If $\theta_\infty \in \overline{\Theta}^R \setminus \Theta^R$, we can prove that P a.s. $f_h(y_1^l) \to -\infty$. Indeed, assume that there exists an event A such that $P(y_1^l \in A) > 0$ and $\inf f_h > -\infty$ on A. Then, for any $y_1^n \in A$, there

exists a sequence $(\theta_m)_{m\in\mathbb{N}}$ in Θ^R such that $\theta_m \to \theta_\infty$ and $\liminf_{m\in\mathbb{N}} \sum_{i=1}^{l} \log(q(y_i|\theta_m)) > -\infty$. However, since $\theta_\infty \in \overline{\Theta}^R \setminus \Theta^R$, $\theta_\infty = (\alpha, \sigma^2, \Gamma)$ with $\sigma^2 \in \{0, +\infty\}$ or $\Gamma = \infty$ in contradiction to Lemma 4 (1). Finally, using the monotone convergence theorem and Lemma 3, we get that

$$E_{P(dy_1^l)}(f_h(y_1^l)) \to -\infty < v(R),$$

and we have proved (24). Since $C = \{\ \theta \in \overline{\Theta^R} \mid \delta(\theta, \Theta_*^R) \geq \epsilon\}$ is compact, there exists a covering of C by a finite family of open sets $(U^j)_{1\leq j\leq r}$ satisfying (24). Thus, denoting $k_n = \lfloor n/l \rfloor - 1$ and $l_n = n - k_n l$, we get

$$\sup_{C\cap\Theta^R} \sum_{i=1}^{n} \log q(y_i|\theta) \leq \sup_{1\leq j\leq r} \left(\sum_{k=0}^{k_n-1} \sup_{\theta\in U^j\cap\Theta^R} \sum_{i=1}^{l} \log q(y_{kl+i}|\theta) + \sup_{\theta\in U^j\cap\Theta^R} \sum_{i=1}^{l+l_n} \log q_\theta(y_{k_n l+i}) \right),$$

so that we deduce from the strong law of large numbers and from (24) that

$$\limsup_{n\to\infty} \frac{1}{n} \sup_{C\cap\Theta^R} \sum_{i=1}^{n} \log q(y_i|\theta) < v(R)\,. \tag{25}$$

Note that all earlier results hold for l sufficiently large. Using $l + l_n$ in the second sum guarantees the applicability of these results.

Given any element $\theta_* \in \Theta_*^R$, we have $\frac{1}{n}\sum_{i=1}^{n} \log q(y_i|\theta^*) \to v(R)$ a.s. and $\frac{1}{n}\sum_{i=1}^{n} \log q_{\hat{\theta}_n^R}(y_i) \geq \frac{1}{n}\sum_{i=1}^{n} \log q(y_i|\theta^*) + \frac{1}{n}(\log q(\theta^*) - \log q(\hat{\theta}_n^R))$ where $q(\theta)$ denote the density of the prior distribution on the parameters. Since this $q(\theta)$ is upper bounded on Θ, we deduce that $\liminf \frac{1}{n}(\log q(\theta_*) - \log q(\hat{\theta}_n^R)) \geq 0$ and

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log q_{\hat{\theta}_n^R}(y_i) \geq v(R)\,. \tag{26}$$

The results follows from (25) and (26). □

Appendix B. Log likelihood approximations

Given the estimated parameters $\theta \in \Theta$, the computation of the log-likelihood of an image y requires integrating out the hidden variable β. This integration could be done using some form of Monte Carlo but for the sake of efficiency we employ a simple approximation detailed below:

Let $y \in \mathbb{R}^{|\Lambda|}$ and $\theta = (\alpha, \sigma^2, \Gamma)$ be fixed, and denote $h(\beta) \doteq -\frac{1}{2}|y - z_\beta I_\alpha|^2 - \frac{|\Lambda|}{2}\log(2\pi\sigma^2)$. We have

$$q(y|\theta) = \int \frac{e^{h(\beta) - \frac{1}{2}\beta^t R\beta} d\beta}{(2\pi)^{k_g} |\Gamma|^{1/2}}\,,$$

where $R = \Gamma^{-1}$. Write $U(\beta) = -h(\beta) + \frac{1}{2}\beta^t R\beta$. Expanding U around any β' we have

$$U(\beta) - U(\beta') - \langle \nabla U(\beta'), \beta - \beta' \rangle = h(\beta) - h(\beta') - \langle \nabla h(\beta'), \beta - \beta' \rangle - \frac{1}{2}(\beta - \beta')^t R(\beta - \beta').$$

Thus for β^* achieving the minimum of U, the third term on the left is zero and we have

$$U(\beta) - U(\beta^*) = U(\beta) - U(\beta^*) - \langle \nabla U(\beta^*), \beta - \beta_* \rangle = \epsilon_*(\beta) - \frac{1}{2}(\beta - \beta^*)^t R(\beta - \beta^*)\,,$$

where $\epsilon_*(\beta) = h(\beta) - h(\beta^*) - \langle \nabla h(\beta^*), \beta - \beta^* \rangle$. Hence

$$\log q(y|\theta)) = U(\beta^*) + \log \int e^{\epsilon^*(\beta)} \mathcal{N}(\beta^*, \Gamma) d\beta\,.$$

The simplest approximation is to assume the integrand is 1 yielding $\log q(y|\theta) \sim E(\beta^*)$.

References

[1] Y. Amit. *2d Object Detection and Recognition: Models, Algorithms and Networks*. MIT Press, Cambridge, Mass., 2002.

[2] Y. Amit, D. Geman, and X. Fan. A coarse-to-fine strategy for multi-class shape detection. *IEEE PAMI*, 26:1606–1621, 2004.

[3] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable template. *Journal of the American Statistical Association*, 86(414):376–387, 1991.

[4] Y. Amit and A. Trouvé. Pop: Patchwork of parts models for object recognition. Technical report, Department of Statistics, University of Chicago, 2004.

[5] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J. Comp. Vis.*, 61(2):139–157, 2005.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Actives appearance models. In H. Burkhards and B. Neumann, editors, *5th European Conference on Computer Vision, Berlin*, volume 2, pages 484–498. Springer, 1998.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1:1–22, 1977.

[8] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of objet categories. In *Proc. ICCV*, 2003.

[9] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unserpervised scale invariant learning. In *Proc. CVPR*, volume 2, pages 264–271, 2003.

[10] C. A. Glasbey and K. V. Mardia. A penalised likelihood approach to image warping. *Journal of the Royal Statistical Society, Series B*, 63:465–492, 2001.

[11] U. Grenander and M. I. Miller. Computational anatomy: an emerging discipline. *Quarterly of Applied Mathematics*, LVI(4):617–694, 1998.

[12] T. Hastie and P. Y. Simard. Metrics and models for handwritten character recognition. *Statistical Science*, 13(1), 1998.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning theory*. Springer, 2001.

[14] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23:S151–S160, 2004.

[15] D. B. Judd. Discussion of 'a penalised likelihood approach to image warping' by glasbey and mardia. *Journal of the Royal Statistical Society, Series B*, 63:465–492, 2001.

[16] E. Klassen, A. Srivastava, W. Mio, and S. Joshi. Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Pattern Analysis and Machine Intelligence*, 26(4):372–383, 2004.

[17] S. Marsland and C. Twining. Constructing diffeomorphic representations for the groupewise analysis of non-rigid registrations of medical images. *IEEE Transactions on Medical Imaging*, 23, 2004.

[18] P. Michor and D. Mumford. Riemannian geometries on spaces of planar curves. *J. Eur. Math. Soc.*, 2005. To appear.

[19] M. I. Miller, A. Trouvé, and L. Younes. On the metrics and euler-lagrange equations of computational anatomy. *Annual Review of biomedical Engineering*, 4:375–405, 2002.

[20] A. Trouvé and L. Younes. Metamorphoses through lie group action. *Foundations of Computational Mathematics*, 5(2):173–198, April 2005.

[21] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.