Sparse Bayesian Registration of Medical Images for Self-Tuning of Parameters and Spatially Adaptive Parametrization of Displacements

Loïc Le Folgoc[a,b], Hervé Delingette[a], Antonio Criminisi[c], Nicholas Ayache[a]

[a] *Asclepios Research Project, Inria Sophia Antipolis, France*
[b] *Microsoft Research – Inria Joint Centre, France*
[c] *Machine Learning and Perception Group, Microsoft Research Cambridge, UK*

Abstract

The tools of Bayesian inference have recently gained interest in tasks of non-rigid registration of medical images, as seminal work demonstrated their potency towards addressing open problems such as the automatic determination of adequate regularization levels or the quantification of confidence in registration outputs. In this paper, we extend the Bayesian modeling of registration to allow for a data-driven, multiscale, spatially adaptive parametrization of deformations. Finer bases get introduced only in the presence of coherent image information and motion, while coarser bases ensure better extrapolation of the motion to textureless, uninformative regions. Adaptive parametrizations have been used with success in the literature to promote both the regularity and accuracy of registration schemes, but so far on non-probabilistic grounds – either as part of multiscale heuristics, or on the basis of sparse optimization. We provide a principled probabilistic approach to find an optimal parametrization of deformations among any preset, widely overcomplete range of basis functions. It thus retains the benefits of the Bayesian formalism, including the estimation of regularization and noise parameters. We further experiment with a richer, more generic model of data that proves to be more faithful for a variety of image modalities than the sum-of-squared differences. We demonstrate the feasibility and performance of our approach on time series of magnetic resonance (cine SSFP and tagged) and echocardiographic cardiac images, and show that the proposed quasi-automatic framework can match or outperform the state-of-the-art on benchmark datasets evaluating accuracy of motion and strain.

Keywords: Registration, Bayesian modelling, Sparse, Automatic Relevance Determination, Cardiac Imaging

1. Introduction

Non-rigid image registration is the ill-posed task of inferring a deformation Ψ from a pair of observed (albeit typically noisy), related images I and J. Classical approaches propose to minimize a functional which weighs an image similarity criterion $\mathcal{D}$ against a regularizing (penalty) term $\mathcal{R}$:

$$\arg\min_{\Psi} \mathcal{E}(\Psi) = \beta \cdot \mathcal{D}(I, J, \Psi) + \lambda \cdot \mathcal{R}(\Psi) \qquad (1)$$

Prior knowledge to precisely model the space of plausible deformations or the regularizing energy is generally unavailable. The optimal trade-off between the image similarity term and the regularization prior is itself difficult to find. Typically the user would manually adjust the ratio λ/β until a qualitatively good fit is achieved, which is time consuming and calls for some degree of expertise. Alternatively if quantitative benchmarks are available on a similar set of images, they can serve as a metric of reference on which to optimize parameters, under the assumption that the value that achieves optimality is constant across the dataset – say, for images of the same modality or among a given sequence. Unfortunately, this assumption generally does not hold. A major feat in the recent literature (Richard et al., 2009; Simpson et al., 2012; Risholm et al., 2013) was to realize that this issue can be tackled automatically by reinterpreting registration in a probabilistic setting. Gee and Bajcsy (1998) first noted that, in a Bayesian paradigm, the two terms in Eq. (1) relate respectively to a likelihood and prior on the latent transformation Ψ. In fact the parameters λ and β themselves can be treated as hidden random variables, equipped with broad prior distributions, and jointly inferred with Ψ or integrated out. In practice, analytical inference is precluded and various strategies

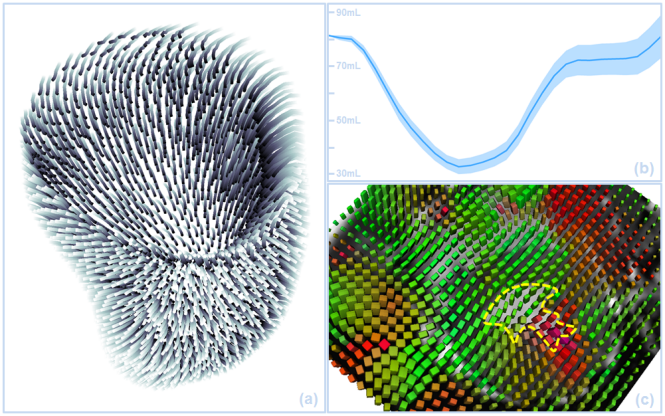


Figure 1: (a) Trajectories of points on the endocardium, following the registration of a time series of cardiac MR images by the proposed approach. (b) LV volume over time and 99.7% confidence interval. (c) Tensor visualization of directional uncertainty at end-systole, rasterized at voxel centers of a 2D slice. For a thorough description, please refer to the discussion in section 5.

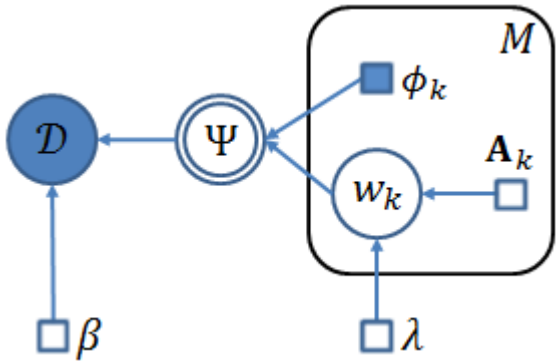


Figure 2: Graphical representation of the probabilistic registration model. The data pair D is put in relation via a transformation Ψ of space, up to some noise (β). The transformation is parameterized by weights w_k on a predefined over-complete set of basis functions $\{\phi_k, k = 1 \cdots M\}$. Priors on the transformation smoothness and on the relevance of individual bases introduce additional parameters (λ and $\mathbf{A}_k$ respectively). Random variables are circled, whereas deterministic parameters are in squares. Arrows capture conditional dependencies. Shaded nodes are observed. The doubly circled node indicates that the transformation Ψ is fully determined by its parent nodes (the ϕ_k and $\boldsymbol{w}_k$). The plate stands for M (replicated) groups of nodes, of which only a single one is shown explicitely.

are devised for approximate inference. Risholm et al. (2013) characterize the distributions of interest from MCMC samples. This is a most principled, generic and accurate approach provided that enough samples can be drawn within the available computational budget. Aside from monitoring the progress of the scheme, two difficulties arise: crafting an efficient proposal distribution over Ψ and computing the acceptance probability of the proposed sample. To circumvent this latter issue, the authors sample from an approximate posterior distribution derived in a variational free-energy framework. Alternatively, the *full* inference can be tackled in a variational framework. In this spirit, Simpson et al. (2012) derive a fast Bayesian approach using variational Bayes Expectation Maximization. This offers an appealing compromise between the computational burden and the quality of the estimates, depending on the chosen family of variational (approximate) posterior distributions.

Despite the incurred computational cost, this probabilistic view of registration offers substantial benefits. In this article, we demonstrate how such a framework can be extended so as to automatically select the scale and location of bases used to parameterize the transformation Ψ. The complexity of the mapping adapts to the underlying dataset, yielding a reduced set of relevant degrees of freedom: finer bases get introduced only in the presence of coherent image information and motion, while coarser bases ensure better extrapolation of the motion to textureless, uninformative regions. Spatial refinement of the parametrization was previously handled heuristically (Rohde et al., 2003); or it led to alternative formulations of registration via spatially anisotropic filtering (Stefanescu et al., 2004). Here and to our knowledge, for the first time, it is approached on principled grounds within a probabilistic framework. We develop a statistical model of registration in which a reduced parametrization of the transformation is automatically inferred from the data jointly with all model parameters; and propose an efficient algorithmic scheme that renders inference over this model tractable for real scale registration tasks. In particular we extend the scope of a state-of-the-art tool for sparse regression and classification Tipping (2001); Tipping et al. (2003). To increase its potency in the context of registration, we lift a strong assumption of independency on hidden variables, so that it now handles generic quadratic regularizing priors at no cost in algorithmic complexity. We also generalize it to multi*valued* regression (regression of vector fields as opposed to scalar fields), so as to preserve the natural invariance to a change of coordinate system.

This article expands on earlier work of the authors (Le Folgoc et al., 2014) in several ways. Firstly, we propose a different approximation of the likelihood term, effectively removing a computational bottleneck – specifically, the voxelwise, local optimization of the image similarity *via* dense block-matching. Instead, a step of global optimization of the posterior distribution w.r.t. the reduced parametrization (typically, a hundred degrees of freedom) is performed. Furthermore we introduce a flexible noise model that can account robustly for acquisition noise and artefacts, and seamlessly adapts to a range of image modalities. We demonstrate our approach on tasks of motion tracking on real cardiac data, specifically time sequences of 3D cine or tagged MR images and echocardiographic images.

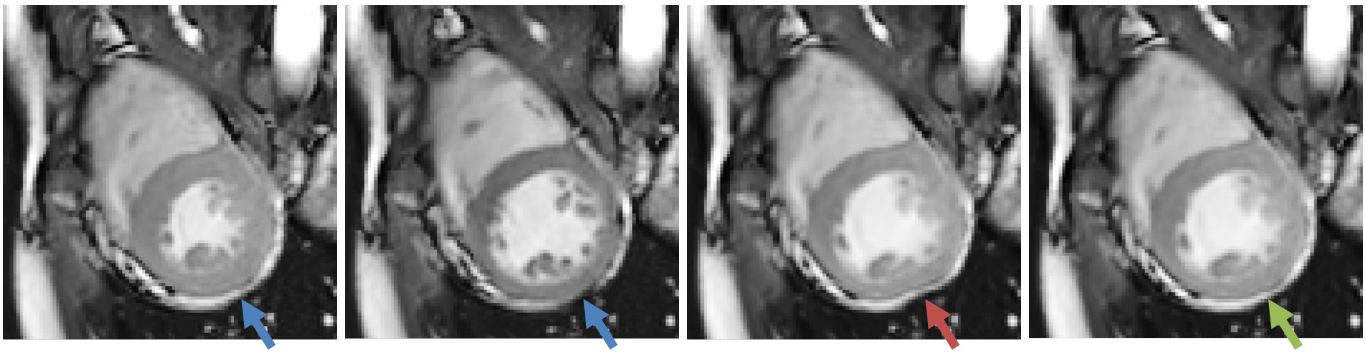
2. Statistical Model of Registration

Registration assumes that the dataset of interest describes objects related *via* some transformation of space. In a medical context, this occurs for instance when the motion of organs is followed over time in a sequence of images. Registration then aims at recovering the underlying transformation of space from the observed data. This can be formally regarded as an *inference* problem and handled as such. We start by expliciting our statistical model of pairwise registration. Fig. 2 provides a graphical depiction thereof. We specify a generative model of the data (*e.g.* images, landmarks) $D = \{D_1, D_2\}$ given the transformation Ψ, along with a prior over the admissible set of transformations. The general strategy to infer the parameters of this model is exposed at the end of the section.

The abstract graphical model depicted in Fig. 2 bears a strong resemblance to that of the Relevance Voxel Machine of Sabuncu and Van Leemput (2011), developed independently by the authors for regression and classification tasks. In Section 3 we propose alternative inference schemes with significant gains in algorithmic complexity, very much in the spirit of how the later work of Tipping et al. (2003) improved upon the original Relevance Vector Machine (Tipping, 2001). This effectively renders the approach applicable to non-rigid registration.

2.1. Data Likelihood

A good transformation Ψ should adequately map the datasets $D = \{D_1, D_2\}$, up to some misalignment and residual error attributable to the data formation process. Our knowledge of this process is captured in a *likelihood* model, which assigns a probability $p(D|\Psi)$ for the data D to be observed under some transformation Ψ. The data likelihood is commonly related to the



a. Moving image b. Fixed image c. SSD loss d. Robust loss

Figure 3: We illustrate the appeal of a robust variant of the SSD image loss based on a mixture-of-Gaussians model (GMM). Images (c,d) display the output warped images obtained after registering images (a) and (b), using respectively the SSD-based likelihood or the GMM-based likelihood (section 2.1). The arrows point towards a specific region that highlights the limitations of the SSD: the subset of hypo-intense voxels bordering the myocardium in the fixed image has no evident counterpart in the moving image. The SSD still drives the motion towards the best matches intensity-wise, which induces implausible tangential stretch of the myocardium. Using a variational argument it can be shown that the GMM, on the other hand, incorporates a natural mechanism to downweight regions that cannot be reliably paired from image to image based on intensity values. This leads to a locally conservative registration that is qualitatively closer to our expectations.

data-matching energy that appears in the classic optimization framework of Eq. (1) by:

$$\mathcal{D}(D,\Psi) = -\ln p(D|\Psi) + \text{const}\,. \tag{2}$$

For landmark registration, $\mathcal{D}(D,\Psi)$ is generally chosen to be the sum of squared distances between pairings. Alternatively the quadratic loss can be replaced by robust losses such as the L_1 norm, or other loss functions derived from the heavy tailed family of Student-t distributions (Tipping and Lawrence, 2005).

For pairwise registration of images, various data-matching terms were introduced in the literature. The simplest and most common image similarity term is the sum of squared difference (SSD) of voxel intensities, which can be improved upon by modeling spatially varying noise levels (Simpson et al., 2013) and artefacts (Hachama et al., 2012), or by relaxing assumptions over the intensity mapping between images – *e.g.* to a piecewise constant mapping (Richard et al., 2009), to a locally affine mapping (Cachier et al., 2003) or to a more complex, non-linear (Parzen-window type) intensity mapping (Janoos et al., 2012). Mutual information is another popular image similarity, especially in the context of registering images of different modalities (Wells III et al., 1996), and has been successfully applied to the registration of cardiac images (Chandrashekara et al., 2004).

SSD is a simple yet efficient image similarity term for registration of monomodal cardiac images. It naturally lends itself to a probabilistic interpretation and eases mathematical derivations. The target image J is modeled as the warped source image $I \circ \Psi^{-1}$ further corrupted by additive, independent identically distributed (i.i.d.) noise $e_i \sim \mathcal{N}(0,\beta)$ at each voxel $i = 1 \cdots N$:

$$J = I \circ \Psi^{-1} + \boldsymbol{e} \tag{3}$$

where $\boldsymbol{e} \sim \mathcal{N}(0,\beta\mathbf{I})$, $\mathbf{I}$ the $N \times N$ identity matrix. β is a global scaling parameter: it denotes the precision (or inverse variance) of the noise across the image. The SSD model can be described in a more familiar manner by the energy of Eq. (4), where $\{c_i\}_{i=1}^N$ is the list of voxel centers in the fixed image and $C_i = \Psi^{-1}(\mathrm{c_i})$ are the paired coordinates in the moving image.

$$\mathcal{D}_\beta(I,J;\Psi) = \frac{\beta}{2}\sum_{i=1}^{N}\left(J[\mathrm{c_i}] - \mathrm{I}[\mathrm{C_i}]\right)^2 \tag{4}$$

Since the SSD is quadratic w.r.t to intensity differences of paired voxels in the registered images, both the penalty for intensity discrepancies and the *rate* at which it grows can become arbitrarily high. This renders registration particularly vulnerable to strong local intensity biases, introduced for instance by topology changes in the imaged objects or by acquisition artefacts. Fig. 3 demonstrates such an occurrence where the shortcomings of SSD result in a qualitatively poor registration. Furthermore residual misalignments between structures of interest in the pair of registered images typically yield higher intensity residuals than those observed at background voxels, as seen in Fig. 4a. Sources of model bias and acquisition noise cannot be captured together in a plausible manner with a single, spatially uniform noise level. In other words, the SSD noise model lacks both the flexibility and robustness required to cope with the various sources of discrepancy in the intensity profiles of the fixed and warped images I and J. To address this limitation we propose to model the noise at each voxel $i = 1 \cdots N$ with a mixture of Gaussian distributions $e_i \sim \sum_{l \le L} \pi_l \mathcal{N}(0,\beta_l)$. The corresponding data matching energy is given in Eq. (5), where we denoted by $Z_l = \sqrt{2\pi/\beta_l}$ the normalizing constant for the Gaussian probability distribution function.

$$\mathcal{D}_{\boldsymbol{\beta},\boldsymbol{\pi},L}(I,J;\Psi) = -\sum_{i=1}^{N}\log\sum_{l=1}^{L}\frac{\pi_l}{Z_l}\exp-\frac{\beta_l}{2}\left(J[\mathrm{c_i}] - \mathrm{I}[\mathrm{C_i}]\right)^2 \tag{5}$$

At each voxel, the residue is implicitely associated to a(n unknown) component of the mixture. This naturally yields a spatially varying model of noise that is better suited to render the complexity of noise patterns in medical images. Unlike previous work (Simpson et al., 2013; Le Folgoc et al., 2014) we do

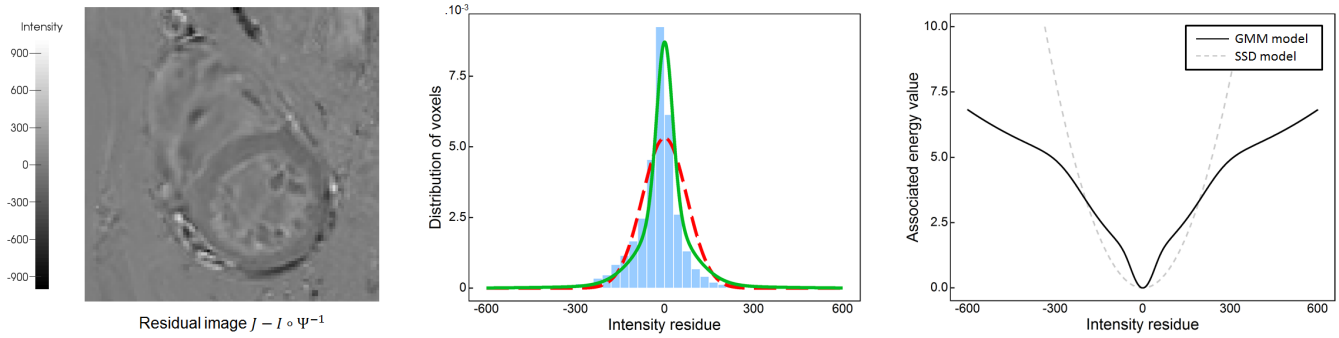


Figure 4: (Left) Example residual image after pairwise image registration, as per the example of Fig. 3. Artefacts and structures of variable appearance between registered images stand out in a distinct manner, much unlike ambient noise. Note that the intensity of the cardiac muscle itself differed between moving and fixed image. (Middle) Histogram of intensity residuals, with SSD and GMM fits overlayed respectively in red and green. (Right) Energy profiles for the SSD (grey dashes) and GMM (black line). The penalty is plotted as a function of the difference of intensities in the registered images. The mixture-of-Gaussians model achieves robustness by incorporating concave inflexions that result in a soft threshold on the penalty incurred for large intensity residuals.

not assume that the noise varies smoothly across the image, as patterns arising from misalignment and imaging artefacts are local in nature. Moreover the parameters L, $\boldsymbol{\pi} = \{\pi_l\}_{l\leq L}$ and $\boldsymbol{\beta} = \{\beta_l\}_{l\leq L}$ of the mixture will be jointly inferred from the data – so that it fits the expected distribution of intensity residuals between registered images. Fig. 4b shows the histogram of intensity residuals obtained after registering the images depicted in the 2D example of Fig. 3, along with the Gaussian mixture fit jointly during the registration process. From the standpoint of energy-based formulations, the procedure effectively learns from the data the most appropriate data matching energy among a prior family of candidates. The respective profiles of the standard SSD loss and the learned Gaussian mixture (GMM) loss are comparatively displayed in Fig 4c for the aforementioned example. The characteristic inflexion of the GMM loss, with a reduced growth rate as the intensity residual becomes higher, is responsible for its robustness towards intensity artefacts compared to the standard SSD quadratic loss. From a computational standpoint, Eq. (5) fortunately admits variational quadratic upperbounds that serve as mathematically sound proxies for the exact loss and make it as convenient to use as the SSD. It is in fact handled as an iteratively reweighted (voxelwise) SSD when necessary. A similar procedure is described fully by *e.g.* Archambeau and Verleysen (2007). For the sake of simplicity and clarity of exposition, most of the derivations in this paper are therefore presented for the SSD loss.

Another limitation of SSD shared by all aforementioned variants is to assume that each voxel provides an independent value of intensity. This assumption does not hold in practice however (Simpson et al., 2012): the residual between the warped image $I \circ \Psi^{-1}$ and its counterpart J exhibits local spatial correlations, either intrinsic to the image acquisition and pre-processing (*e.g.* image pre-smoothing, image upsampling) or introduced as a consequence of registration misalignments. Ignoring local correlations in the noise pattern leads to an artificial increase in the number of independent observations and induces overconfidence in the data term. On the other hand, modeling precisely the noise structure would come at a significant computational cost. Instead, we follow Simpson et al. (2012) in artificially downweighting the SSD term by a factor α that captures redundancies in voxelwise observations, based on a virtual decimation procedure suggested by Groves et al. (2011).

A final fallacy in our generative model of data stems not from the way residuals are modeled, but from the later assumption that the intensity profile can be evaluated with infinite accuracy at any point in the moving image; whereas we actually rely on interpolation of a discrete scalar field. In regions where strong intensity gradients occur between adjacent voxels (*e.g.* in the case of neighbouring hypo- and hyper-intense regions), ignoring interpolation errors leads to an unreasonably high sub-voxel confidence in the optimally paired coordinates $C_i = \Psi^{-1}(\mathrm{c_i})$. Indeed the region in which the intensity value $I[\mathrm{C_i}]$ in the moving image coincides, up to the noise level, with the value $J[\mathrm{c_i}]$ in the fixed image may then shrink down in the gradient direction unreasonably below sub-voxel dimensions. To account for interpolation uncertainty, Wachinger et al. (2014) model interpolation as a noisy observation of the hidden, true intensity value and propose an approximate scheme to marginalize over latent variables. We opt instead for a simple scheme that couples efficiently with a Laplace approximation of the data likelihood, as explained in section 3.1.

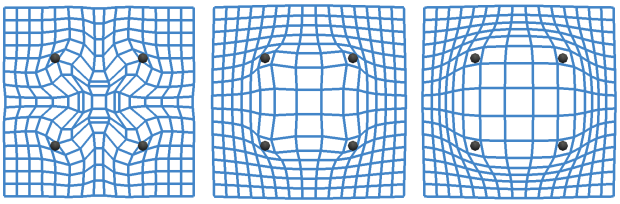
2.2. *Representation of displacements*

In the context of non-rigid registration, an admissible space of transformations should be specified. In this work we restrict ourselves to a small deformation framework, $\Psi^{-1} = \mathrm{Id} + \mathrm{u}$, with a parameterized representation of the displacement field $\mathrm{u}\colon \mathrm{x} \in \mathbb{R}^d \mapsto \mathrm{u(x)} \in \mathbb{R}^d$. We constrain the displacement field to be expressed over a dictionary $\{\phi_k\}_{k=1}^M$ of radial basis functions, specifically Gaussian kernels $\phi_k(\mathrm{x}) = K_{S_k}(\mathrm{x}_k, \mathrm{x})\,\mathbf{I}$ where $\mathbf{I}$ is the $d \times d$ identity matrix and

$$K_S(x, y) = \exp -\frac{1}{2}(x-y)^\intercal S^{-1}(x-y)\,. \tag{6}$$

In other words, the displacement field u is parameterized by a set of weights $\boldsymbol{w}_k \in \mathbb{R}^d$ associated to each basis ϕ_k:

$$\mathrm{u}_{\boldsymbol{w}}(\mathrm{x}) = \sum_{1\leq k\leq M} \phi_k(\mathrm{x})\boldsymbol{w}_k = \boldsymbol{\phi}(\mathrm{x})^\intercal \boldsymbol{w}\,. \tag{7}$$



Penalizing compressibility ($D = \mathrm{div}$) | Elastic membrane energy ($D = \nabla$) | Bending (thin-plate) energy ($D = \nabla^2$)

Figure 5: Impact of the regularization model. Displacements are parameterized by isotropic Gaussian kernels of set width $\sigma = 0.25$. From left to right, the regularizer varies. The data consists of 4 points regularly sampled on the unit circle, forming an axis aligned square, pulled twice as far away from the origin as they initially were. The warped grid obtained by regression is displayed along with the ground truth displacement.

$\boldsymbol{\phi}(\mathrm{x}) = \left(\phi_1(\mathrm{x}) \cdots \phi_M(\mathrm{x})\right)^\intercal$ and $\boldsymbol{w}^\intercal = \left(w_1^\intercal \cdots w_M^\intercal\right)$ are respectively the concatenation, for $k = 1 \cdots M$, of $\phi_k(\mathrm{x})$ and w_k.

The basis centers x_k span a predefined regular grid of points, typically the whole range of voxel centers. The kernel width S_k is also allowed to vary and spans a user-predefined set of values $S_1, S_2, \cdots, S_q$. This allows for a redundant, multiscale representation of displacements. In other words we benefit both from a compact representation *via* larger kernels, and from the ability to capture finer local details *via* smaller kernels.

This dictionary of basis functions can be seen as a finite-dimensional approximation to the space $\mathcal{H}_S$ spanned by Gaussian kernels of a given width $S \leq \min_k\{S_k\}$. Such spaces have attractive properties and are related in the literature to spaces of *currents* (Gori et al., 2013). In particular, we derive in AppendixA a family of *analytical*, computationally efficient probabilistic priors over $\mathcal{H}_S$.

2.3. Transformation Prior

In non-rigid registration, the displacement $\mathrm{u}_{\boldsymbol{w}}$ is insufficiently constrained by the data and some regularizing prior has to be imposed over its parameters. This prior distribution encapsulates our knowledge of the deformation and our modeling assumptions (see for instance Sotiras et al. (2013) for an exhaustive review of deformation priors). We will consider Gaussian priors of the form

$$p(\boldsymbol{w}|\lambda, \{\mathbf{A}_k\}) \propto \exp -\frac{1}{2}\left\{\lambda \boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w} + \sum_{k=1}^{M} \boldsymbol{w}_k^\intercal \mathbf{A}_k \boldsymbol{w}_k\right\} \tag{8}$$

where λ and $\{\mathbf{A}_k\}_{k=1\cdots M}$ are so called hyperparameters. The motivation for such a prior is two-fold.

Regularity control. Gaussian priors in the form of Eq. (9) let us penalize physically implausible deformations. They have been commonly used in the literature starting with Broit (1981), both because of their natural interpretability and soundness in mechanical terms, and their convenience from an algorithmic and computational standpoint.

$$q(\boldsymbol{w}|\lambda) \propto \exp -\frac{1}{2}\lambda \boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w} \tag{9}$$

The structure of the precision matrix $\mathbf{R}$ can be adjusted to penalize the magnitude of the first derivative (Gee and Bajcsy, 1998) or higher order derivatives (Rueckert et al., 1999; Ashburner, 2007; Ashburner and Ridgway, 2013), effectively encoding a wide range of priors.

We specifically consider the subset of quadratic forms that exploit the structure of the space of Gaussian kernels introduced in Section 2.2; namely priors of the type $\boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w} = \|D\mathrm{u}_{\boldsymbol{w}}\|^2_{\mathcal{H}}$, with $D\mathrm{u}$ any differential operator acting on u. Details are reported in AppendixA. By doing so, we obtain among others an analytical, computationally efficient implementation of a membrane energy with $D = \nabla$, a bending (thin plate) energy with $D = \nabla^2$, and a penalty favoring incompressible, divergence free, behaviours by setting $D = \mathrm{div}$. Fig. 5 illustrates the respective impacts of such penalties.

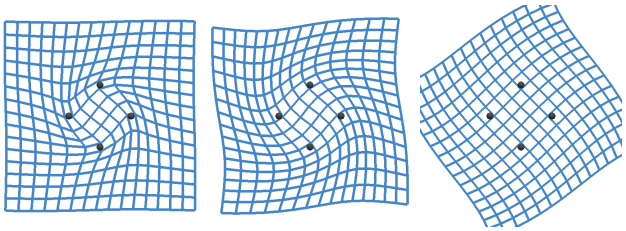
Basis selection. The second factor in our prior, recalled in Eq. (10), induces the basis selection mechanism that we exploit in this article.

$$q(\boldsymbol{w}|\{\mathbf{A}_k\}) \propto \prod_{k=1}^{m} \exp -\frac{1}{2}\boldsymbol{w}_k^\intercal \mathbf{A}_k \boldsymbol{w}_k \tag{10}$$

The additional term $\boldsymbol{w}_k^\intercal \mathbf{A}_k \boldsymbol{w}_k$ for each basis ϕ_k lets us penalize independently the recourse to this basis to capture the displacement, by penalizing high magnitudes of its associated weight $\boldsymbol{w}_k$. The limit case of infinite $\boldsymbol{A}_k$ actually constrains $\boldsymbol{w}_k$ to be null and thus forbids the use of ϕ_k to represent the signal. In section 2.6 we propose a principled way to determine optimal values of the set $\{\mathbf{A}_k\}_{k=1\cdots M}$, from which most of them turn out to be infinite: we thus obtain a sparse representation of the displacement from our initial, over-complete dictionary.

2.4. Sparse Coding & Registration

Despite not conveying a clear biophysical meaning, sparsity-inducing priors retain a two-fold motivation. The first benefit is in terms of algorithmic complexity. Unless resorting to low parametric models, the shear size of the parametrization makes



Basis scale: $\sigma = 0.25$ | Basis scale: $\sigma = 1.0$ | Basis scale: $\sigma = 4.0$

Figure 6: Impact of the basis scale on the inferred transform. From left to right, the displacement field is parameterized by isotropic Gaussian kernels of increasing width. The data consists of 8 points regularly sampled on the unit circle. The underlying motion is a rotation of $\pi/4$ radian. Only four of these eight rotated samples are displayed for readibility. The scale of the bases used to represent the transform strongly affects the area of influence of the data points, as can be seen from the scale at which the regressed transform resembles a global rotation.

direct optimization cumbersome without the recourse to sophisticated solvers. The computation of exact covariance matrices that are typically involved in probabilistic approaches also becomes unfeasible, while diagonal approximations used in their stead ignore significant interactions induced by the data and priors, and encoded by off-diagonal terms. Secondly, basis selection mechanisms adaptively constrain the admissible space of deformations, automatically tuning the degrees of freedom to the smallest sufficient set able to capture the observed displacement. Coupled with a multiscale set of basis functions, this yields a data-driven, automatic spatial refinement of the granularity of the displacement field that greatly complements the otherwise scale-blind L_2 regularization. Adaptive, multiscale regularization was shown to yield state-of-art results *e.g.* in denoising natural scenes (Fanello et al., 2014). It is also critical for medical image registration as the spread of informative structures often lacks spatial homogeneity, yet such salient structures must somehow drive the registration in otherwise textureless areas. Moreover, the amount of noise and artefacts may vary across the image in hardly predictable patterns – and the degree of coarseness at which the displacement is modeled should be adaptively refined in consequence. Fig. 6 illustrates the key impact of the scale at which a displacement is inferred when relying on limited observations, on a simple toy example where the underlying motion is a rotation.

While L_1 priors have met with widespread success and use in all areas of sparse coding, including for registration Shi et al. (2013), their usefulness in a probabilistic formulation is greatly diminished by their tendency to spoil the joint estimation of model parameters and the derivation of uncertainty estimates. L_1 priors are also likely to underperform (w.r.t. the degree of sparsity) in the presence of strongly correlated explanatory variables. For a more extensive review and benchmark of L_1 and bayesian sparse learning methods, we refer the reader to the work of Mohamed et al. (2012). The prior of Eq. (10) was first introduced by Tipping (2001) for regression and classification tasks with the so called Relevance Vector Machine. The authors demonstrated its relevance for sparse coding when used in conjunction with the framework of Automatic Relevance Determination (MacKay, 1992). Bishop and Tipping (2000) offer an alternative sparse Bayesian learning (SBL) view on the Relevance Vector Machine, where they opt for a Variational Bayes treatment. Wipf and Nagarajan (2008) further investigate links between the SBL and ARD frameworks and resulting schemes. Alternatively, Eq. (8) can be interpreted as a generalized spike-and-slab prior (Mitchell and Beauchamp, 1988) despite using a different parametrization, provided that each $\mathbf{A}_k$ is constrained to a binary state – either null or infinite.

2.5. Hyper-priors

In this work, we treat hyperparameters $\{\mathbf{A}, \beta, \lambda\}$ as frequentist parameters rather than latent random variables. For a fully Bayesian formulation, prior distributions would be imposed over model parameters. When conjugate priors are available this yields a very elegant model over which, for instance, Variational Bayes approximate inference may be used to derive closed form iterative updates. We refer the reader to the work of Simpson et al. (2012) for an application to image registration.

However, in absence of strong prior knowledge over the values of λ or β, broad uninformative priors are typically chosen. In the limit this effectively yields identical updates to the ones we derive with point estimates of the parameters while simplifying the exposition. Moreover the uniform prior on $\mathbf{A}_k$ has the appealing benefit of making our inference scheme invariant to rescaling of basis functions, as will be seen from Eq. (22), (23). This is highly desirable as we rely on a multiscale dictionary of basis functions (section 2.2) whose scaling factors may otherwise be hard to relate.

2.6. Model Inference

Our probabilistic model is summarized in a graphical manner in Fig. 2. The registration task now consists in inferring the displacement from the observed data and the prescribed graphical model. We describe the high-level approach in what follows. Section 3 presents the method from an algorithmic standpoint.

Hyperparameter inference. We first estimate the values of the model hyperparameters $\{\{\mathbf{A}_i\}, \lambda, \beta\}$ by maximizing the so called *marginal likelihood* of the data $p(D|\mathbf{A}, \lambda, \beta)$. This framework is known as that of type-II maximum likelihood in the statistical literature. Note that in computing the marginal likelihood, we integrate over the parameters $\boldsymbol{w}$ (these parameters are *marginalized* out):

$$p(D|\mathbf{A}, \lambda, \beta) = \int_{\boldsymbol{w}} p(D|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\mathbf{A}, \lambda) d\boldsymbol{w} \tag{11}$$

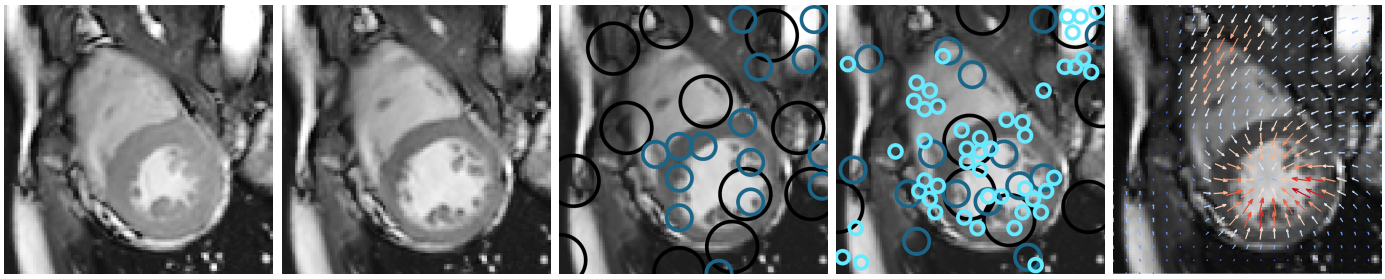
where for convenience of notation we introduce the block diagonal matrix $\mathbf{A} = \mathrm{diag}(\mathbf{A}_i)$.

Posterior distribution of model parameters. Given the optimal parameters $\mathbf{A}^*, \beta^*, \lambda^* = \arg\max_{\mathbf{A},\beta,\lambda} p(D|\mathbf{A}, \lambda, \beta)$, we can derive statistics of interest on the transformation ψ by first computing the *posterior* distribution of its parametric representation $\boldsymbol{w}$ conditioned on the observed data D. Bayes' rule asserts that:

$$p(\boldsymbol{w}|D, \mathbf{A}^*, \beta^*, \lambda^*) = \frac{p(D|\boldsymbol{w}, \beta^*) p(\boldsymbol{w}|\mathbf{A}^*, \lambda^*)}{p(D|\mathbf{A}^*, \lambda^*, \beta^*)} . \tag{12}$$

In particular, the maximum of Eq. (12) minimizes Eq. (1). This bridges the gap with the classical framework in which registration is seen as the task of optimizing a functional. In Eq. (12) however, the hyperparameters assume *optimal* values $\mathbf{A}^*, \beta^*, \lambda^*$ defined w.r.t. the dataset of interest D. Moreover, the posterior distribution does not merely encode a point estimate of $\boldsymbol{w}$. Its higher-order moments relate to the variability in the inferred parameters. We consider a Gaussian approximation $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ of Eq. (12) around its posterior mode to obtain an approximate second moment of the distribution as Σ.

Predictive distribution of displacements. Our model Eq. (7) implies a linear relationship between the displacement value $\mathrm{u}(\mathrm{x}) = \boldsymbol{\phi}(\mathrm{x})^\intercal \boldsymbol{w}$ at any $\mathrm{x} \in \mathbb{R}^d$ and the model parameters $\boldsymbol{w}$.



a. Moving image | b. Fixed image | c. Adaptive parametrization (In the initial steps) | d. Adaptive parametrization (At the end) | e. Output displacement field

Figure 7: Basis selection mechanism displayed on an example 2D registration between slices of cardiac MR images, respectively at end systole (a) and end diastole (b). (c,d) Bases selected in the initial steps of the algorithm vs. at the end, using isotropic Gaussian RBFs at different scales. The circles correspond to the isocontour at one standard deviation from the center of the basis. This is for the sake of readibility – the actual area of influence of a given basis covers two to three times its standard deviation. In the initial steps, the displacement is captured at a coarse level and the extreme regularization forces a somewhat uniform spread of bases. Later on, bases are clustered in regions of high displacement: bases at finer scales help automatically refine the captured displacement in regions where it is called for, most notably close to the myocardium. (e) Inverse displacement field output by the algorithm (scaled by a factor of 2), smoothly varying across the whole image despite using a sparse subset of bases.

Formally, the posterior distribution of u(x) is then given by Eq. (13), where $\boldsymbol{\theta} = \{\mathbf{A}, \beta, \lambda\}$ denotes the set of hyperparameters and δ the Dirac distribution.

$$p(\mathrm{u}(\mathrm{x})|D, \boldsymbol{\theta}^*) = \int_{\boldsymbol{w}} \delta_{\boldsymbol{\phi}(\mathrm{x})^\intercal \boldsymbol{w}}[\mathrm{u}(\mathrm{x})]\, p(\boldsymbol{w}|D, \boldsymbol{\theta}^*) d\boldsymbol{w} \tag{13}$$

If the posterior distribution $p(\boldsymbol{w}|D, \boldsymbol{\theta}^*)$ of $\boldsymbol{w}$ is normally distributed $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, u(x) is in turn Gaussian with mean $\boldsymbol{\phi}(\mathrm{x})^\intercal \boldsymbol{\mu}$ and covariance $\boldsymbol{\phi}(\mathrm{x})^\intercal \Sigma \boldsymbol{\phi}(\mathrm{x})$. More generally, $\mathrm{u}|D, \boldsymbol{\theta}^* : \mathrm{x} \mapsto \mathrm{u}(\mathrm{x})$ is a Gaussian process with mean $\bar{\mathrm{u}}(\,\cdot\,) = \boldsymbol{\phi}(\,\cdot\,)^\intercal \boldsymbol{\mu}$ and covariance

$$\mathrm{Cov}\,(\mathrm{u}(\mathrm{x}), \mathrm{u}(\mathrm{y})) = \boldsymbol{\phi}(\mathrm{x})^\intercal \Sigma \boldsymbol{\phi}(\mathrm{y})\,. \tag{14}$$

Full posterior vs. frequentist posterior. Strictly speaking we would in fact rather estimate the *full* posterior, where hyperparameters have been integrated out as per Eq. (15). When using uniform priors over hyperparameters, the posterior probability $p(\boldsymbol{\theta}|D)$ is proportional to the marginal likelihood $p(D|\boldsymbol{\theta})$.

$$p(\mathrm{u}(x)|D) = \int_{\boldsymbol{\theta}} p(\mathrm{u}(x)|D, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \tag{15}$$

The full posterior generally cannot be computed without resorting to MCMC estimates. However if hyperparameters are well determined by the data, $p(\boldsymbol{\theta}|D) \approx \delta_{\boldsymbol{\theta}^*}$ becomes highly peaked around its mode $\boldsymbol{\theta}^*$. The quality of this assumption is discussed at greater length by Tipping (2001). The full posterior of Eq. (15) is then approximately equal to the frequentist posterior of Eq. (13). This motivates our two-step approach of first looking for the maximum likelihood hyperpameters $\boldsymbol{\theta}^*$ before computing the frequentist posterior.

3. Inference schemes

As described in the previous section (2.6), our inference strategy is based on the maximization of the marginal likelihood as per Eq. (11). Unfortunately, the closed form evaluation of this type of integral is generally intractable, whereas its approximation via MCMC sampling schemes can often be prohibitively costly. Tipping (2001) notes that when the likelihood and the prior are normally distributed, the data conveniently also follows a Gaussian distribution; furthermore the form of Eq. (11) becomes such that tractable maximization schemes can be derived with respect to the hyperparameters. We propose a Gaussian approximation of the likelihood term and extend the inference strategy of (Tipping et al., 2003) to the broader class of priors described in 2.3.

3.1. Approximation of the Model Likelihood

The assumption of Gaussianity of the data given the model parameters $\boldsymbol{w}$ does not stand for registration purposes and several strategies can be considered to approximate the model likelihood. In earlier work (Le Folgoc et al., 2014), a Taylor expansion of the likelihood around one of its local maxima was applied, resulting in a dense block matching step. Here instead Laplace's method is used around the current estimate of the mode $\boldsymbol{\mu}_{\mathrm{MP}}$ of the posterior distribution, found by quasi Newton (BFGS) optimization on Eq. (12). In other words, as in conventional energy-based registration, we numerically solve for the minimizer $\boldsymbol{\mu}_{\mathrm{MP}}$ of Eq. (16) using the current estimate of hyperparameters $\mathbf{A}, \lambda, \beta$:

$$\mathcal{E}(\boldsymbol{w}) = \mathcal{D}_\beta(I, J; \Phi\boldsymbol{w}) + \frac{1}{2}\boldsymbol{w}^\intercal(\mathbf{A} + \lambda\mathbf{R})\boldsymbol{w} \tag{16}$$

Most notably however, Eq. (16) in effect only involves the sparse subset of selected bases ϕ_k for which $\mathbf{A}_k < +\infty$. Proceeding with Laplace's approximation of the data likelihood and dropping the term involving the Hessian of the image[1], we arrive at Eq. (17):

$$\mathcal{D}(I, J; \boldsymbol{w}, \beta) \approx \frac{1}{2}\sum_{i=1}^{N} (\mathrm{t_i} - \boldsymbol{\phi}(\mathrm{c_i})\boldsymbol{w})^\intercal\, \beta\mathbf{H}_i\, (\mathrm{t_i} - \boldsymbol{\phi}(\mathrm{c_i})\boldsymbol{w})\,. \tag{17}$$

[1] This outer product approximation is justified in *e.g.* (Bishop et al., 2006).

Algorithm 1 Sparse Bayesian registration algorithm

Initialize $\mathbf{A}_k = \infty$ for all k ($\mathcal{S} = \emptyset$)
Initialize $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Sigma} = 0$
repeat
  Update β according to Eq. (29).
  Find the posterior mode $\boldsymbol{\mu}_{\mathrm{MP}}$ of Eq. (12) for the current values of $\mathbf{A}, \lambda, \beta$ by quasi-Newton (BFGS) search.
  Compute a quadratic approximation of the likelihood around $\boldsymbol{\mu}_{\mathrm{MP}}$: Eq. (17),(18)
  Recompute $\boldsymbol{\mu}$ ($= \boldsymbol{\mu}_{\mathrm{MP}}$) and $\boldsymbol{\Sigma}$ in full from Eq. (21)
  Recompute $\boldsymbol{q}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{\kappa}_k$ in full from Eq. (D.3), (D.4) and Eq. (D.5), (D.6), for all k
  for p iterations **do**
    $\forall k \in \mathcal{S}$ (resp. $k \notin \mathcal{S}$), compute the gain $\max_{A_k} \Delta l(\mathrm{A}_k)$ in marginal likelihood obtained by updating or deleting k from $\mathcal{S}$ (resp. adding k to $\mathcal{S}$), using Eq. (24) and AppendixB.
    Select the most favorable action i s.t. $\max_{A_i} \Delta l(\mathrm{A}_i) \geq \max_{A_k} \Delta l(\mathrm{A}_k)$ for all k.
    Set $A_i^* = \arg\max_{A_i} l(\mathrm{A}_i)$ and update $\mathcal{S}$.
    Update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ via rank-one matrix identities from Eq. (C.1), (C.4), (C.3).
    Update $\boldsymbol{q}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{\kappa}_k$ for all k using rank-one updates *e.g.* Eq. (D.7), (D.8) and Eq. (D.5), (D.6).
  end for
  Update λ according to Eq. (26).
until no action leads to a significant increase in marginal likelihood.

The virtual observations $\mathrm{t_i} \in \mathbb{R}^{\mathrm{d}}$ and the confidence tensors $\mathbf{H}_i \in \mathcal{M}_{d\times d}$ are given by Eq. (18), where $\mathrm{C_i} = \mathrm{c_i} + \mathrm{u_{MP}(c_i)}$ stands for the current (posterior) estimate of the pairing for $\mathrm{c_i}$.

$$\mathrm{t_i} = \mathrm{u_{MP}(c_i)} - \frac{I(\mathrm{C_i}) - J(\mathrm{c_i})}{\|\nabla I(\mathrm{C_i})\|^2} \nabla I(\mathrm{C_i})\,, \quad \mathbf{H}_i = \nabla I(\mathrm{C_i}) \nabla I(\mathrm{C_i})^\intercal \quad (18)$$

These virtual pairings immediately relate to the optical flow: if we dropped the confidence tensors $\mathbf{H}_i$, the above would yield an approximation of Eq. (1),(12) much in the spirit of the *demons* algorithm (Thirion, 1998; Cachier and Ayache, 2004). The tensors $\mathbf{H}_i$ vary sharply across the image however, *e.g.* as edges or boundaries are crossed. They assign anisotropic, spatially varying confidence in voxelwise pairings and account for how informative and structured the image is at the point of interest.

In particular the confidence $\beta\mathbf{H}_i$ in the virtual voxelwise pairing $t_i + c_i$ can grow arbitrarily high for arbitrarily high intensity gradients. These expressions result from linearizing the intensity profile I around the current pairing C_i, and are blind to interpolation uncertainty in evaluating $I(C_i)$ and $\nabla I(\mathrm{C_i})$. To address this shortcoming, we propose to replace $\beta\mathbf{H}_i$ by

$$\left((\beta\mathbf{H}_i)^{-1} + \mathbf{D}_{\mathrm{int}}\right)^{-1} = \frac{1}{1 + \mathrm{tr}[\beta\mathbf{H}_i\mathbf{D}_{\mathrm{int}}]} \cdot \beta\mathbf{H}_i \quad (19)$$

which implements a soft upper threshold on the precision, as a heuristic for interpolation uncertainty. $\mathbf{D}_{\mathrm{int}}$ acts as a minimum covariance: it is a diagonal matrix set to the square of –say– half the voxel spacing to prevent unreasonable subvoxel confidence.

3.2. Form of the Marginal Likelihood

We now assume that the data $\boldsymbol{t}$ is generated by corrupting the true signal $\boldsymbol{u} = \Phi\boldsymbol{w}$ with Gaussian noise $\boldsymbol{e} \sim \mathcal{N}(0, \beta\mathbf{H})$:

$$\boldsymbol{t} = \Phi\boldsymbol{w} + \boldsymbol{e} \quad (20)$$

Note that in block form, Eq. (17) yields an approximate likelihood model in the form of Eq. (20). Given the hyperparameters of the model, the posterior distribution of $\boldsymbol{w}$ conditioned on the data – obtained by combining the likelihood and prior with Bayes' rule according to Eq. (12) – is Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with

$$\boldsymbol{\mu} = \Sigma\Phi^\intercal\beta\mathbf{H}\boldsymbol{t} \quad \Sigma = (\Phi^\intercal\beta\mathbf{H}\Phi + \lambda\mathbf{R} + \mathbf{A})^{-1} \quad (21)$$

The integrand in Eq. (11) involves the product of two Gaussian factors, and by integrating out the weights $\boldsymbol{w}$ we obtain the *evidence* or *marginal likelihood* for the hyperparameters:

$$p(\boldsymbol{t}|\mathbf{A}, \lambda, \beta) = |\,2\pi\,\mathbf{C}\,|^{-1/2} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{t}^\intercal\mathbf{C}^{-1}\boldsymbol{t}\right\} \quad (22)$$

where by identification $\mathbf{C}^{-1} = \beta\mathbf{H} - (\beta\mathbf{H})\Phi\Sigma\Phi^\intercal(\beta\mathbf{H})$. In other words, the distribution of the data $\boldsymbol{t}$ conditioned on the hyperparameters $\{\mathbf{A}, \lambda, \beta\}$ is Gaussian $\mathcal{N}(0, \mathbf{C})$. Furthermore, it follows from the Woodbury matrix identity that

$$\mathbf{C} = (\beta\mathbf{H})^{-1} + \Phi(\mathbf{A} + \lambda\mathbf{R})^{-1}\Phi^\intercal\,. \quad (23)$$

Interestingly, the process of setting the hyperparameters can then be seen as fitting a covariance model $\mathbf{C}$ to the data $\boldsymbol{t}$ via a maximum likelihood approach. Note also that the two factors in Eq. (22) have antagonistic effects: while the left hand term penalizes covariance matrices $\mathbf{C}$ that waste mass (via $|\mathbf{C}|$), the right hand term gives incentive to spend mass to better explain the data $\boldsymbol{t}$. This compromise leads to sparsity, as revealed by a careful look at the form of $\mathbf{C}$ in Eq. (23). Indeed, regardless of the value of $\mathbf{A}$, part of the data is explained *for free* by the contribution $(\beta\mathbf{H})^{-1}$ of the noise to $\mathbf{C}$; thus only a few degrees of freedom need be active ($\mathbf{A}_k < \infty$) to fully explain the data.

3.3. Hyperparameter inference

Following the strategy discussed in 2.6, we wish to maximize the marginal likelihood (Eq. (22)), or equivalently its logarithm, with respect to the model hyperparameters. We consider schemes that monotonically converge towards a local maximum by *iteratively* maximizing, or merely increasing, the evidence w.r.t. either one of the hyperparameters $\mathbf{A}$, β or λ.

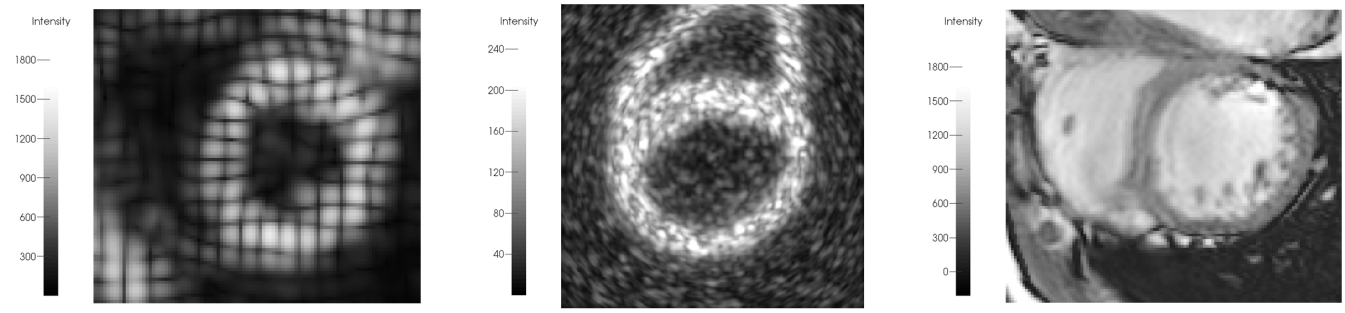


Figure 8: Example slices for the cardiac imaging modalities that we experiment on. (Left) 3D tagged MR image (Middle) 3D echocardiographic image (Right) 3D cine SSFP MR image. In all experiments the same flexible model of registration is applied successfully, regardless of the artefacts and patterns peculiar to each modality.

Maximization w.r.t. A — a hill-climbing scheme. The procedure relies on a sequence of additions and deletions of candidate basis functions starting from an empty set $\mathcal{S} = \emptyset$ of active bases (all $\mathbf{A}_k$'s set to ∞). This notably avoids the $O(M^3)$ cost that would stem from the matrix inversion in Eq. (21), if all M bases were included in the active set $\mathcal{S}$ at the start (all $\mathbf{A}_k < \infty$). At each iteration, we take a single action among the addition of a previously inactive basis ($k \notin \mathcal{S}$), or the update or deletion of an active one ($k \in \mathcal{S}$), in a principled way. Specifically, we implement the action that leads to the largest gain in evidence. This is possible because, when all other hyperparameters $\mathbf{A}_{-k}$, λ, β are fixed, the contribution $l(\mathbf{A}_k)$ of a given basis to (the logarithm of) the evidence for any value of its associated hyperparameter $\mathbf{A}_k$ can be singled out in the form of Eq. (24). Details about the statistics involved $\boldsymbol{\kappa}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ and the maximization of $l(\mathbf{A}_k)$ are left to AppendixB.

$$l(\mathbf{A}_k) = \log|\mathbf{A}_k + \boldsymbol{\kappa}_k| - \log|\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k| + \boldsymbol{q}_k^\intercal \{\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k\}^{-1} \boldsymbol{q}_k \tag{24}$$

Naturally, as bases are added or removed from the active set $\mathcal{S}$ via updates of their associated hyperparameter, the potential contribution $\max_{\mathbf{A}_k} l(\mathbf{A}_k)$ of other bases is subject to change. In practice, the statistics $\boldsymbol{\kappa}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ indeed depend on $\mathbf{A}_{-k}$, λ, β via the moments $\boldsymbol{\mu}$ and Σ of the posterior distribution (Eq. (21)). Therefore, after updating a given $\mathbf{A}_i$, we exploit rank-one matrix identities to update these moments with a complexity of at most $O(|\mathcal{S}|^2)$ (AppendixD). We then update all the $\boldsymbol{\kappa}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ with a complexity of $O(|\mathcal{S}|)$ per basis. Therefore the worst-case complexity of the updates is of $O(|\mathcal{S}|^2 + M|\mathcal{S}|)$, as opposed to $O(M^3)$ for EM updates. The decrease in complexity is rather extreme, as typical respective orders of magnitude of $|\mathcal{S}|$ and M for registration would be 10^2 and 10^6.

To gain a better intuition on how this scheme proceeds, let us look back at the quantities involved in Eq. (24). Firstly, $\boldsymbol{q}_k$ captures the relevance of ϕ_k towards providing a better explanation for the data. It is related to the projection of the residual on the basis ϕ_k. When L_2 regularization is used ($\lambda > 0$), $\boldsymbol{q}_k$ also involves the projection of the residual on nearby active bases ϕ_j ($j \in \mathcal{S}$). Secondly $\boldsymbol{s}_k$ captures the amount of overlap between the basis ϕ_k under consideration and those already in the active set $\mathcal{S}$, therefore accounting for competition between strongly correlated bases. $\boldsymbol{\kappa}_k$ arises from the added regularization $\exp -\frac{1}{2}\lambda \boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w}$, which was absent in the regular RVM ($\lambda = 0$). The sum $\mathbf{A}_k + \boldsymbol{\kappa}_k$ entering Eq. (24) highlights the competition between the shrinkage (sparsity-inducing) mechanism due to $\mathbf{A}_k$ and the L_2-norm energy regularization that is accounted for by the quantity $\boldsymbol{\kappa}_k$.

Maximization w.r.t λ and β — EM updates. We estimate the hyperparameters λ, β using Expectation Maximization updates. Instead of directly maximizing our target criterion, the marginal-likelihood $p(\boldsymbol{t}|\mathbf{A}, \lambda, \beta)$, EM proceeds by maximizing a surrogate quantity, the expected complete log-likelihood:

$$\arg\max_{\lambda,\beta} \mathbb{E}_{p(\boldsymbol{w}|D,\mathbf{A}_*,\beta_*,\lambda_*)} \left[\log p(\boldsymbol{t}, \boldsymbol{w}|\mathbf{A}_*, \lambda, \beta)\right] . \tag{25}$$

The expectation is taken with respect to the current estimate of the posterior distribution $p(\boldsymbol{w}|D, \mathbf{A}_*, \beta_*, \lambda_*)$ of $\boldsymbol{w}$. β_*, λ_* are quantities fixed at their current values, as opposed to variables to optimize. In practice, we iterate between reestimation of the noise level β and of the regularization trade-off λ, fixing the other parameter in turn. The EM scheme is appealing as it guarantees an increase in marginal likelihood, despite maximizing a surrogate. Furthermore, when the distributions involved are Gaussian, the expected complete log-likelihood (25) can be expressed directly in terms of the mean and covariance matrix of the Gaussian distributions. For the regularizing trade-off λ, this leads to maximizing the smooth, convex function:

$$\lambda^* = \arg\max_{\lambda \geq 0} -\frac{\lambda}{2}\mathrm{tr}(\Sigma \mathbf{R}) + \frac{1}{2}\log|\mathbf{A} + \lambda \mathbf{R}| - \frac{\lambda}{2}\boldsymbol{\mu}^\intercal \mathbf{R} \boldsymbol{\mu} . \tag{26}$$

The solution can be found numerically in inexpensive $O(|\mathcal{S}|)$ iterations after a single singular value decomposition in $O(|\mathcal{S}|^3)$. Alternatively, if we restrict $\mathbf{A}_k$ to be either null or infinite as suggested in section 2.4, we obtain the simpler analytical update:

$$|\mathcal{S}| \cdot \lambda^* = \boldsymbol{\mu}^\intercal \mathbf{R} \boldsymbol{\mu} + \mathrm{tr}(\Sigma \mathbf{R}) . \tag{27}$$

Leaving optimization details aside, we can instead gain some insight into the EM update by examining the solution of the

constrained optimization problem Eq. (26). Setting aside the case where the constraint is active ($\lambda^* = 0$), the solution λ^* cancels the gradient of the function to maximize. From a careful analysis of the terms involved in the gradient, the following identity holds at λ^*:

$$\mathbb{E}_{p(\boldsymbol{w}|\mathbf{A}_*,\lambda^*)}\left[\boldsymbol{w}^\mathsf{T}\mathbf{R}\boldsymbol{w}\right] = \mathbb{E}_{p(\boldsymbol{w}|D,\mathbf{A}_*,\beta_*,\lambda_*)}\left[\boldsymbol{w}^\mathsf{T}\mathbf{R}\boldsymbol{w}\right] \tag{28}$$

λ^* is chosen such that a sample drawn from the updated prior has the same energy as a sample drawn from the inferred posterior, on average. In other words, the EM update modifies the strength λ of the prior to best reflect what has been inferred from the data. The noise level β is updated according to Eq. (29). It accounts for a bias between the estimated and true displacements, and for the variance in the estimated displacement.

$$N \cdot \beta^{*-1} = \|\boldsymbol{t} - \Phi\boldsymbol{\mu}\|^2_{\mathbf{H}} + \mathrm{tr}(\Sigma\Phi^\mathsf{T}\mathbf{H}\Phi)\,. \tag{29}$$

In other words, β^{*-1} is set to the expected error (averaged over pixels) if sampling from the current estimate of the posterior.

3.4. Algorithmic overview

Algorithm 1 summarizes our pipeline for registration. The algorithm mostly works by iterative refinements of the subset of active basis functions $\{\phi_j, j \in \mathcal{S}\}$ with fast updates, based on an *approximate* likelihood model around the current mode of the displacement parameters $\boldsymbol{\mu}_{\mathrm{MP}}$. Every now and then, the noise and regularization parameters λ, β are updated, which makes it necessary to recompute every statistic in full. Before doing so, we re-estimate the posterior mode $\boldsymbol{\mu}_{\mathrm{MP}}$ from the *true* likelihood model and the current (reduced) set $\mathcal{S}$ of active bases and recompute the likelihood approximation around this mode. To further accelerate the pipeline, the scheme can be coupled with a multiresolution pyramidal scheme, starting with downsampled (smoothed) versions of the image I and J and progressively moving through the pyramid of images to the images I and J at full resolution.

4. Experiments & Results

We experiment with the proposed sparse Bayesian framework on tasks of cardiac motion tracking. The goal in cardiac motion tracking is to accurately recover the hidden motion of the cardiac muscle over the course of the cardiac cycle from a time series of 3D images. The first experimental setup aims at clarifying and analyzing the empirical behaviour of the proposed algorithm and focuses on a simple example of 2D pairwise registration. All other sections involve full $3D + t$ motion tracking on various imaging modalities, namely cine SSFP sequences, tagged sequences and a synthetic ultrasound dataset. Fig. 8 displays example 2D slices from frames of each modality.

The experimental setup is identical across all modalities, and indeed we aim to demonstrate that the proposed framework is flexible enough to adapt seamslessly to the peculiarities of each dataset. For all 3D experiments, the multiscale parametrization of the displacement field consists of isotropic Gaussian kernels at two scales, of respective variance $S_1 = 20^2$ mm^2 and

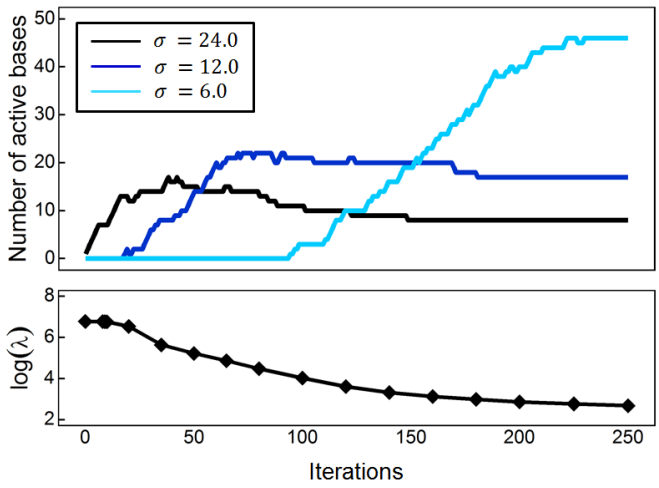


Figure 9: Basis selection mechanism and its coupling with the jointly estimated regularization level, across iterations. (Top) We track the addition, update or deletion of dictionary bases in the active parametrization of the displacement field. Three distinct scales are used in our multiscale representation of displacements. Each of the 3 curves records the number of active dictionary bases of a given scale, throughout iterations. (Bottom) Value of the regularization parameter λ plotted against the number of iterations run since the beginning of the registration. The parameter λ is re-estimated every few iterations only during registration. Each point on the curve corresponds to an actual reestimate of its value. The curve connects those points for the sake of visualization.

$S_2 = 10^2$ mm^2, plus an anisotropic Gaussian kernel of variance 10^2 mm^2 in the short axis plane and 20^2 mm^2 along the long axis. Indeed since our framework imposes no restriction on the parametrization of the displacement field, a natural way to put this advantage to use is to introduce anisotropic bases of potential anatomical relevance. All scales are jointly optimized upon. As explained in section 3.4, the multiscale representation of the displacement field is coupled with a classic multiresolution, pyramidal scheme on the images of interest themselves – they are downsampled by a factor 2 (and smoothed) at three different resolutions. In other words the lowest resolution level is subsampled by a factor of 4 compared to the original image. Finally, we use a bending (thin-plate) energy as a regularizer (Section 2.3) and we constrain the hyperpriors $\mathbf{A}_k$ to a binary state (sections 2.4 and 3.3) to prevent competition between two types of regularization: the regularity induced by $\lambda\mathbf{R}$ on the derivatives of the displacement and that induced by $\mathbf{A}_k$ on the magnitude of specific weights w_k. Note also that our registration scheme does not make use of pre-segmentations of regions of interest, which could be challenging or otherwise impractical to obtain.

4.1. Self-tuning registration algorithm: an analysis

As explained in the previous sections, the parameters introduced in the model of registration are inferred jointly during the course of the algorithm. We now leverage the example 2D registration of Fig. 3 7 to provide some insight into our schemes and analyze the convergence of key parameters throughout iterations.

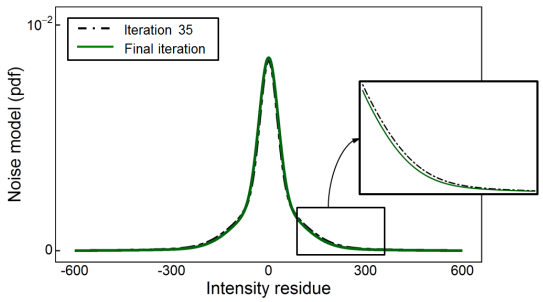


Figure 10: Inferred noise model, *i.e.* inferred expected distribution of intensity residuals between the fixed and warped images. The learned distribution is comparatively shown at an early stage in the registration (black dashed line) and at convergence (green solid line). The curves are nearly indistinguishable, hinting towards fast convergence in this example. The zoom on the distribution tails still evidences a slightly lower noise level at the end of registration, as intuitively expected.

Basis selection & regularization. From an algorithmic standpoint, the algorithm proceeds by iteratively adding dictionary bases into – or deleting them from – the active parametrization of the displacement field. Every few iterations, the noise model and regularization level λ are re-adjusted based on the current estimate of the displacement and its uncertainty. Fig. 9 demonstrates how basis selection mechanisms empirically combine with parameter re-estimation throughout iterations to provide a seamless convergence towards a reasonable local minimum.

The regularization level λ is initialized thanks to a heuristic that provides a "very large" value for the registration at hand. Initially this effectively prevents the addition of finer dictionary bases in the model, whose impact on the signal regularity is too high at this stage. Instead coarse bases are favoured, which capture the global trends in the observed displacement. The regularization level is consequently refined to reflect the actual regularity in this inferred displacement. As λ decreases towards a more sensible value, finer bases get incorporated in the active set to capture finer local details of the visible motion, or to ensure that these finer details of the inferred motion blend smoothly with the rest of the displacement field. In case of significant overlap between a subset of fine bases and a coarse basis, the basis at the coarsest scale may automatically be removed – as it no longer contributes towards a better explanation of the data. Towards the last iterations, algorithmic convergence has roughly been reached. This is evidenced by the plateau value of λ and by the fact that most actions consist in small updates of active bases (specifically, their orientation) rather than in the addition or deletion of dictionary bases.

Fig. 7 further illustrates this mechanism of basis selection, with the location of bases in the active parametrization being depicted at two points in time – in the initial steps of the algorithm and at the end.

Noise model estimation. The noise model is jointly estimated over the course of the algorithm. In all experiments it displayed a fast convergence towards its final inferred distribution. This is examplified by Fig. 10, where the probability density functions corresponding to the Gaussian mixture inferred at an early iteration and at the final iteration are hardly distinguishable. This partly reflects the fact that our registration experiments involve small displacements, with the coarse patterns in the underlying motion being captured early on. The Gaussian mixture is also quick to adapt to changes in the distribution of intensity residuals that arise from our multiresolution pyramidal scheme, when hopping from a smoothed downsampled image to the next level in the pyramid of images, as seen from Fig 11. The jumps from a coarser image resolution to a finer image resolution occur at iteration 10 and iteration 20.

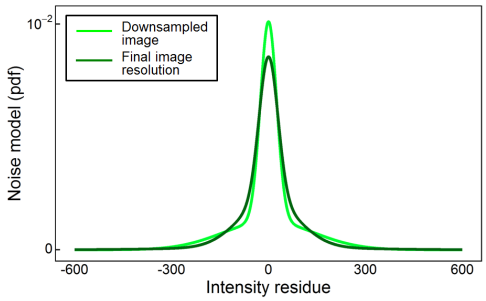


Figure 11: Inferred noise model at the beginning of registration and at the end. The registration scheme relies on a multiresolution representation of registered images, jumping from a coarse resolution to a finer resolution at predetermined iterations. The inferred distribution of intensity residuals is shown for the lowest resolution image (light green) and finest resolution image (dark green). As expected the noise model learned on downsampled, smoothed images has a higher probability of low noise (higher peak around 0) but also of high noise (higher tails) due to the increased misalignment at the beginning of registration.

Robustness w.r.t. initialization. Beyond the seamless convergence of model parameters over the course of the algorithm, one may also wonder whether their estimated final value displays consistency across a range of possible initializations. Fig. 12 provides evidence towards the empirical robustness of the estimated level of regularity λ w.r.t. its initial value. The final value of λ typically varies by significantly less than an order of magnitude when initialized from a range of values covering several orders of magnitude.

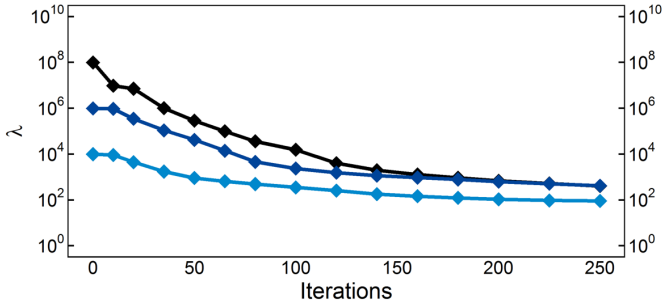


Figure 12: Robustness of the inferred regularity level w.r.t. its initial estimation. The 2D registration is run 3 times, and initialized each time with a differing level of regularity λ (respectively 10^4, 10^6, 10^8). Each curve shows the evolution of λ over the course of the associated run. While the initial value of λ spans 4 orders of magnitude, its final estimate varies by at most a factor of 4 across runs.

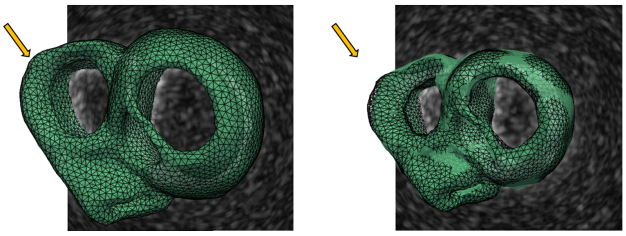


Figure 13: Ground truth mesh (green transparent surface) vs. reference mesh transported *via* registration (overlayed black wireframe). The extrapolated motion out of the field of view (where the arrow points) remains close to the ground truth. The maximum error does not exceed 4mm. Best seen by zooming in.

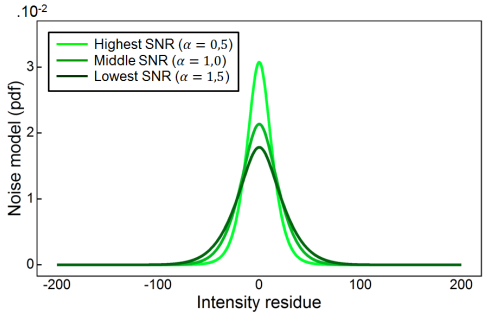


Figure 15: Evolution of the inferred noise model for increasing SNRs (Signal-to-Noise Ratios). As expected, the distribution becomes more sharply peaked around 0 at higher SNRs, which expresses a higher prevalence of small intensity residuals.

4.2. Synthetic 3D Ultrasound Cardiac Dataset

We first demonstrate our approach on synthetic sequences of 3D ultrasound data provided as part of a registration challenge organized for the 2012 MICCAI workshop on Statistical Atlases and Computational Models of the Heart (STACOM). Details on the challenge methodology can be found in De Craene et al. (2013) along with participant results. These synthetic images count approximately 10 million voxels each, at an extremely fine isotropic resolution of 0.33mm. Without further optimization of our code w.r.t. RAM management we had to downsample them by a factor of 2 to process them. We thus worked at a resolution of 0.66mm at the finest level.

The appeal of this benchmark is to offer a dense ground truth in terms of motion and strain inside the cardiac muscle, as displacements in the myocardium are directly prescribed from the output of an electromechanical model of the heart as part of the workflow of image generation. For each sequence of images, the ground truth consists of a sequence of meshes of the left and right ventricles deformed over the cardiac cycle. The data extracted from such ground truth meshes can be compared to that obtained by deforming the mesh at a reference time point (namely, end diastole) throughout the cardiac cycle with the transformation output by the proposed registration approach.

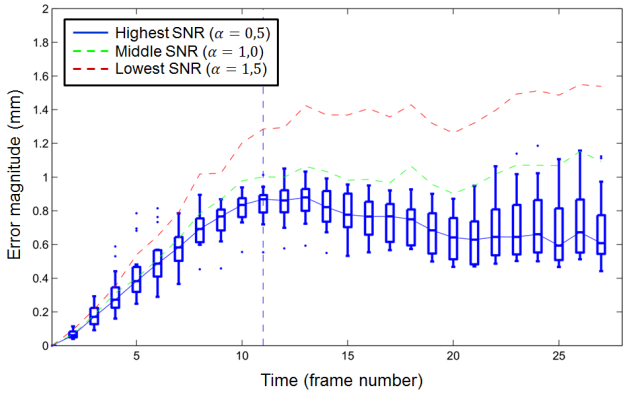


Figure 14: Accuracy benchmark on the 3D US STACOM 2012 normal dataset, reporting the median tracking error over time for varying SNRs (blue, green, red curves). For the reference SNR (in blue), quartiles are overlayed (boxplots) to picture the dispersion of error values.

Prior to any quantitative assessment, we first comment that the visual and qualitative behaviour of the proposed approach was found to be satisfactory, even in terms of extrapolation – as indeed the inferred motion remained consistent in areas of the right ventricle that fell outside of the field of view (Fig. 13). This indicates an effective regularization mechanism, despite being automatically tuned.

We evaluate the accuracy of our approach on a first subset of sequences that image the same motion at various Signal-to-Noise Ratios (SNRs). Because the proposed approach infers a consistent motion both inside and outside of the field of view, we find natural to assess its accuracy from statistics based on the whole mesh. This slightly departs from the methodology of De Craene et al. (2013) where part of the left ventricle only is considered. Fig. 14 reports the median point-to-point error in the inferred displacement for each time frame, where the median statistics is computed from every node in the mesh. At the best SNR, the highest error is observed around end systole with a median of 0.83mm, although the spread of error values becomes wider in the last frames. This falls in the same range as that reported for challenge participants by De Craene et al. (2013) – although slightly higher than the most accurate methodology. Of course part of the error is likely to be attributable to the use of downsampled, smoothed images with a resolution of 0.66mm as opposed to 0.33mm. Besides as the signal to noise ratio degrades, we observe as expected a global trend of increased error magnitude. The increased SNR impacts the noise model learned by the proposed approach, as seen from Fig. 15, which in turn becomes more conservative in its estimates of displacements.

Fig. 16a reports (Green Lagrangian) strain measures at end systole averaged over AHA segments. This provides indirect evidence of the relevance of the automatically tuned regularity level λ and of the displacement parametrization. Ground truth values of strain obtained from the corresponding ground truth mesh are compared to those estimated from the output of registration. Variations in the strain across segments are generally well captured, even more so for its longitudinal and circumferential components. Similarly to most methodologies however, the radial strain – which captures the thickening of the muscle during the contraction – appears to be globally somewhat un-

derestimated in the left ventricle. Such bias in the radial strain might result from a slight bias in the estimated placement of the endo- and/or epicardium. This might indicate a slightly conservative estimate of displacements due to a coarse parametrization or over-regularized transformation. The following table provides statistics on the number of bases of each scale used for the parametrization of the displacement field, for the normal case at highest SNR.

Basis type	Median # (Q1 – Q3)
σ = 20mm	17.5 (14.25 – 19)
anisotropic σ	15 (11.25 – 20.5)
σ = 10mm	34 (31.25 – 38)
Total	**64.5 (60 – 71)**

Table 1: Number of bases at each scale in the active parametrization of the displacement field (pooled over all frames in the sequence). Median, first and third quartiles are reported.

The number of active bases on these sequences is typically smaller than that used in our experiments on cine and tagged data, with a lesser reliance on fine-scale bases. It may evidence increased conservatism in the estimated displacements, as well as indicate greater regularity of the synthetic ground truth motion. The benchmark also provides datasets that aim at reproducing pathological cardiac function, including a case where certain AHA segments become quasi akinetic due to ischemy. Fig. 16b summarizes estimated regional strains for this case, with qualitative retrieval of the ischemic segments (bolded contours), as emphasized by the comparison with the normal case. The accuracy on the ischemic case is similar to that of the normal case at identical SNR, with a median error at end systole of 0.80mm.

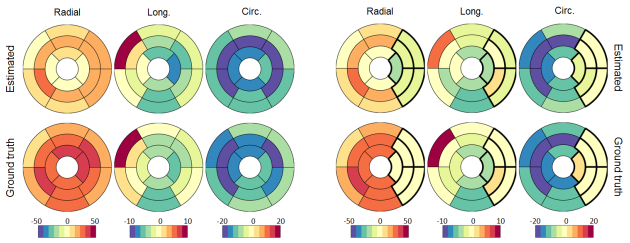


Figure 16: Bull's eye plots of the radial, longitudinal and circumferential strain components at end-systole, averaged over AHA segments: estimated (top) and ground truth (bottom). A healthy case (left) and an ischemic case (right) are reported.

4.3. STACOM 2011 tagged MRI benchmark

In 2011 the MICCAI workshop on Statistical Atlases and Computational Models of the Heart (STACOM) proposed a cardiac motion analysis challenge. The challenge datasets, aimed at evaluating the accuracy of motion tracking algorithms, are openly hosted by the Cardiac Atlas Project. The data includes a set of 15 sequences of 3D tagged MR images. Fig. 8 (Left) shows an example slice for such an image. The grid-like tags overlayed on the region of interest allow to follow the motion of keypoints on the boundary of, or inside the cardiac muscle. Each sequence in the dataset thus comes with a corresponding set of 12 landmarks, the motion of which was manually tracked over time. The landmarks are typically located where tags intersect, and divided in three groups of 4 points in the basal, mid-ventricular and apical areas of the left ventricle. Details of the experimental setting along with challenger results are provided and analyzed by Tobon-Gomez et al. (2013).

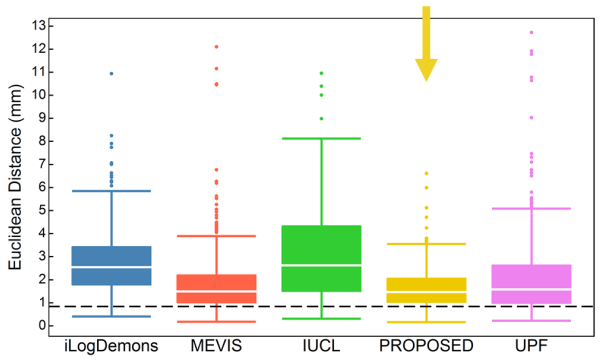


Figure 17: Accuracy benchmark on the 3D tag STACOM 2011 dataset, reporting box-plots of tracking errors on all methodologies. The doted black line represents the average inter-observer variability.

We validate our approach on this dataset of real 3D tagged MR sequences. Manual landmarks serve as ground truth from which the accuracy of our methodology at End-Systole (ES) is assessed. Fig. 17 summarizes challengers' results along with ours. The proposed approach achieves state-of-art results on this benchmark with a median accuracy of 1.46mm. As a point of comparison, the variability in the landmark tracking was estimated as part of the challenge methodology at 0.84mm. We perform two simple statistical tests to quantify the statistical significance of the increase in accuracy of our methodology compared to the challengers: a pairwise Student-t test and a pairwise Kolmogorov-Smirnov test. The tests are run for each pair of samples involving the proposed approach against a challenge participant's. The Student-t test aims at detecting significant differences in the true mean error of our method versus a challenger's, whereas the Kolmogorov-Smirnov test more generally aims at detecting whether the underlying distribution of errors differ. Figures are reported in Table 2 and provide some evidence towards a significant improvement from at least 3 of the 4 methodologies.

Challenger	Student-t p-value	KS p-value
iLogDemons	$\mathbf{< 2.2 \cdot 10^{-16}}$	$\mathbf{< 2.2 \cdot 10^{-16}}$
MEVIS	**0.0099**	0.1385
IUCL	$\mathbf{< 2.2 \cdot 10^{-16}}$	$\mathbf{< 2.2 \cdot 10^{-16}}$
UPF	$\mathbf{2.45 \cdot 10^{-5}}$	**0.00024**

Table 2: Statistical significance of the increase in accuracy on the STACOM 2011 3D motion tracking challenge. We report p-values of pairwise tests for the proposed approach versus each participant's. Bolded values highlight significant improvements at the 5% significance level.

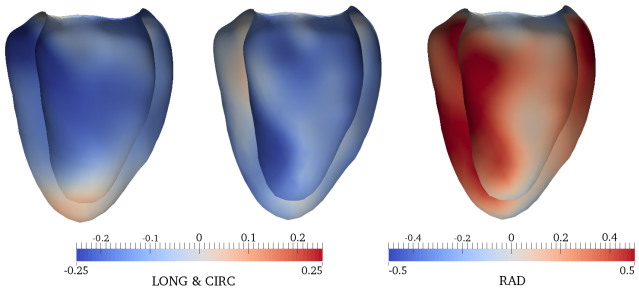


Figure 18: Strain at ES, computed from the 3D tag data of volunteer V9.

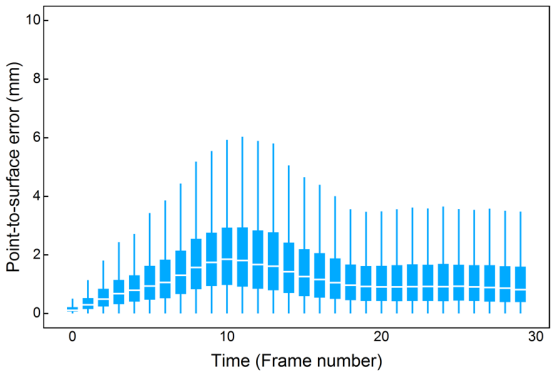


Figure 20: Accuracy benchmark on the cine SSFP STACOM 2011 dataset, reporting median error over time along with quartiles. Surfaces reconstructed from slice-by-slice 3D+t segmentation serve as ground truth. Points on the discrete contours delineated at time 0 are transported over time with the registration output and point-to-surface distances are gathered. For each time step, errors over all 15 sequences and all contour points are pooled. Errors at time 0 are induced by the surface reconstruction.

Rather than over-emphasizing the reach of these statistical tests we would like the reader to appreciate the ability of the proposed formulation to achieve in a quasi automatic manner results qualitatively and quantitatively on par with the state of the art, as demonstrated on this benchmark. In particular it should be emphasized that all parameters involved in the proposed formulation – the noise model and regularity level λ, the active parametrization of the displacement field – were automatically determined during registration. All 3D registration experiments reported in this section involve little user interaction, in effect run from the same model and settings. Interestingly the strain maps and mesh deformations produced by our scheme, as illustrated for instance in Fig. 18, also appear to be qualitatively on par with the best challenge results in that respect, and superior to that of the closest competing methodology accuracy wise (please refer to Tobon-Gomez et al. (2013) for a direct counterpart to Fig. 18). This hints towards the fact that the automatically adjusted weights of the data energy versus the regularization energy, beyond their theoretical grounds, are of practical relevance. Finally we report in Table 3 the number of bases of each scale used for the parametrization of the displacement field.

Basis type	Median # (Q1 – Q3)
σ = 20mm	17 (14.75 – 20)
anisotropic σ	30 (25 – 33.25)
σ = 10mm	100 (90 – 110.25)
Total	**148 (134 – 160)**

Table 3: Number of bases at each scale in the active parametrization of the displacement field (pooled over all sequences and all frames). Median, first and third quartiles are reported.

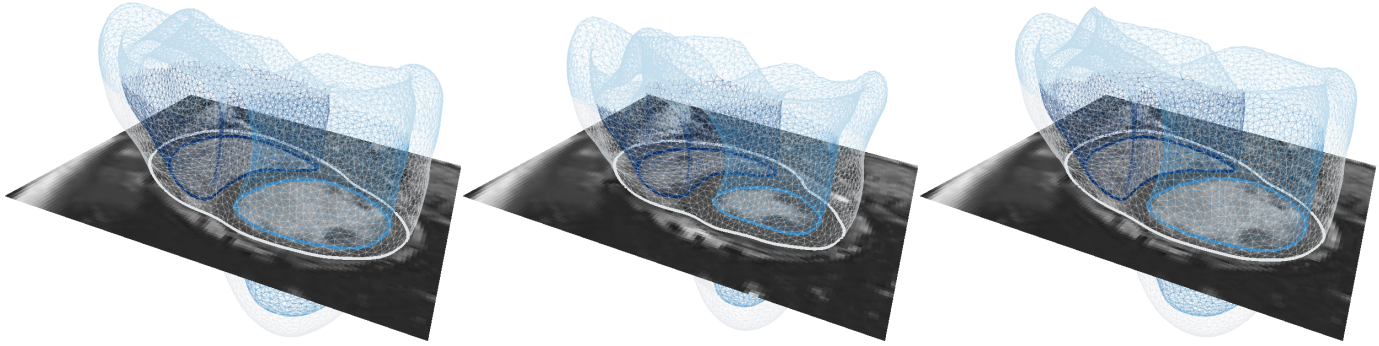
4.4. Cine MRI dataset: qualitative results and uncertainty

In addition to tagged MRI sequences, the STACOM 2011 challenge datasets include 15 sequences of cine SSFP MR cardiac images. For each of the 15 volunteers the left and right ventricles are imaged in 3D, over 30 frames covering the cardiac cycle. Original images had a low inter-slice resolution of 8mm compared to the in-plane resolution of 1.25mm, and we upsampled them (typically by a factor of 5) prior to the registration process to prevent a degradation of numerical accuracy. To obtain a ground truth by direct manual tracking of landmarks over time was deemed difficult for this image modality. Instead the accuracy of the proposed algorithm was evaluated by cross-comparison with direct 3D+t segmentation results. Specifically, the endocardium was delineated over time on 2D slices using the freely available software Segment[2] (Heiberg et al., 2005), yielding a 3D point set of discretized contours. A 3D surface was then reconstructed as the zero level set of a signed distance map computed by radial basis interpolation, after estimating the normal to the surface at every point in the set from a local neighborhood[3]. We then assessed the discrepancy between the reference end diastole segmentation transported over time *via* the output of registration, and the surface estimated by direct segmentation of the endocardium at each time step.

Fig. 20 summarizes the distribution of errors over time, pooled over all 15 sequences and all contour points, displaying the evolution of key quantile-based statistics. The median error reaches a satisfactory maximum of 1.82mm for frame 10, which roughly coincides with the end systole time for all volunteers. As a point of comparison, the volumes under consideration have a spacing of 1.25mm in the short-axis plane (*i.e.* within slices) and 8mm along the long-axis (*i.e.* inter-slice). The wide spread of error values partly reflects the challenge in obtaining a 3D segmentation of the endocardium that remains consistent over time (*e.g.* due to the variable appearance of papillary muscles). Misalignment of short-axis slices in 3D volumes, which may arise from the (slice by slice) image acquisition process, also accounts for some of the largest discrepancies. We observed no evident spatial pattern in the distribution of errors, although the segmentation rarely reached the very tip of the apical region.

The mixture-of-Gaussians noise model proved adequate here again, as distinct components captured variations in the level of noise of an order of magnitude (a factor of 10 between the standard deviations of extreme components). Indeed regions of interest change appearance over time and motion, and tend

[2] http://segment.heiberg.se
[3] http://hdl.handle.net/10380/3149



Frame 0
(End Diastole)

Frame 11
(End Systole)

Frame 29
(End Diastole)

Figure 19: Example registration for the cine SSFP dataset of volunteer V5. We propagate the segmentation from the reference frame to the rest of the time-series with the output of the registration. The resulting mesh is overlayed on a 2D slice and visualized at three representative timesteps. The 3D mesh attests to the regularity of the underlying transform, and to its coherence over the cardiac cycle.

to be assigned higher noise levels than the baseline acquisition noise level. Voxels in basal slices, with visible outflow tracts and apparent topology changes, also tend to fall in the noisiest components. Finally, Fig. 21 attests to the high variability (several orders of magnitude) of the optimal model parameter λ for varying sequences and time steps, which would render its manual estimation via a trial-and-error or cross-validation approach cumbersome. The apparent bimodality of the histogram might reflect the fact that cardiac phases with significant contraction or relaxation, around end systole, alternate with phases of lesser motion around end diastole.

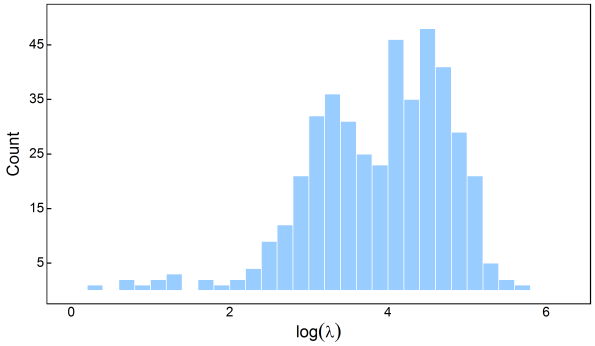


Figure 21: Histogram of inferred values for the regularity hyperparameter λ, pooled over all 15 sequences and 30 frames per sequence.

Finally, because of the strong anisotropy in spacing (lower inter-slice resolution) that is characteristic of the image acquisition process, cine MR data examplifies the necessity of modeling uncertainty in the interpolation of discrete intensity profiles (cf. sections 2.1 and 3.1). When not accounting for it the regularity of the inferred transform was systematically found, upon visual inspection, to be of lesser quality in the direction of lower resolution. This behaviour is to be expected if the scheme is blind to its increased reliance on interpolation to find locations of matching intensity values in the moving image. Image upsampling prior to registration also relies on interpolation and was accounted for in an identical manner. Table 4 reports statistics on the number of bases of each scale used for the parametrization of the displacement field.

Basis type	Median # (Q1 – Q3)
σ = 20mm	18 (10 – 28)
anisotropic σ	29 (21 – 38)
σ = 10mm	44 (29 – 57)
Total	**93 (75 – 111)**

Table 4: Number of bases in the active parametrization of the displacement field (pooled over all sequences and all frames), at each scale. Median, first and third quartiles are reported.

5. Discussion and conclusions

In this paper we proposed a data-driven, spatially adaptative, multiscale parametrization of deformations for registration. It uses larger kernels in regions of high uncertainty due to *e.g.* lack of image gradients or incoherent information in registered images, and uses smaller kernels where a finer motion can be estimated with confidence from local cues in paired images. This is achieved in a Bayesian framework, so that the approach retains natural advantages of probabilistic formulations such as the joint inference of registration hyperparameters. In effect this yields a self-tuning algorithm of general scope with attractive capabilities in terms of achieved accuracy and regularity. The prime contribution of this work is a procedure for fast marginal likelihood maximisation in sparse Bayesian models, that relaxes the assumptions made by Tipping et al. (2003) for the fast Relevance Vector Machine at virtually no algorithmic cost. It broadens the scope of the RVM much in the same direction as the mixed $L1$-$L2$ Elastic Net regularization (Zou and Hastie, 2005) extends the $L1$-norm LASSO regularization (Tibshirani, 1996). This scheme applies to the wide range of classification and regression tasks that share the same abstract graphical representation as Fig. 2 or variants thereof.

While we left the question of uncertainty quantification mostly unadressed, we note that the proposed framework provides us with a readily computable, compact $|\mathcal{S}| \times |\mathcal{S}|$ covariance matrix $\boldsymbol{\Sigma}$ that summarizes uncertainty on the degrees of freedom in the active parametrization. The covariance on transformation parameters can be turned into directional estimates of uncertainty at any point in space by simple linear algebra, or can be sampled from at a marginal $O(|\mathcal{S}|^2)$ cost to efficiently explore the *joint* variability of the full transformation. Sampling the transformation itself, unlike sampling displacements independently at each point in space, preserves correlations in the displacement of close-by points. This can be instrumental in deriving empirical estimates of uncertainty on integral geometrical quantities. For instance, Fig. 1(b) reports estimates of uncertainty in the volume enclosed over time by the endocardium surface, as segmented on the reference frame (at time 0), for a cine MRI sequence (volunteer 5). For the same volunteer, Fig. 1(c) summarizes, in the form of a tensor map, the uncertainty in the inferred displacement field at end-systole, accounting for uncertainty in the output of each frame-to-frame registration between end-diastole and end-systole. Tensors are rasterized at the voxel centers of the end-systole frame. Each tensor encodes (the square root of) the 3×3 covariance matrix of the displacement at a given point and is evidently elongated in directions of higher uncertainty. Due to voxel spacing anisotropy in the cine SSFP dataset, the direction of higher uncertainty is, consistently across space, aligned with the long-axis. The color scheme thus encodes the second principal direction of highest uncertainty. Steep intensity gradients in the underlying image typically translate into directions where tensors are least elongated. Tensor magnitude and principal directions vary smoothly across space, as estimates of uncertainty incorporate information of a local (and in fact, global) nature. The yellow dashed line gives a visual cue as to the position of the left ventricle endocardium boundary. Future work will have to assess extensively the quality of the approximate posterior returned by our fast scheme compared to the exact model posterior (as explored *e.g.* via MCMC techniques), and the empirical agreement between the exact model posterior and intuition.

We experimented with the proposed framework of registration on tasks of motion tracking on dynamic cardiac data. A flexible noise model based on mixtures of Gaussian distributions was introduced and performed suitably on all tested modalities, advantageously replacing models of smoothly varying noise. This is likely due to an increased ability to represent intricate spatial patterns of intensity residuals arising from acquisition noise and artefacts, registration misalignment and variable appearance of organs over time. Despite using generic multiscale RBFs, the inferred parametrizations of 3D displacements were highly sparse, typically involving no more than a hundred degrees of freedom. We note that tagged MR images encouraged the recourse to finer bases; beyond the higher resolution of these volumes (compared to cine SSFP data), it is likely that tags were found to be reliable, informative structures along all directions of motion. While good accuracy was achieved on synthetic echocardiographic time series with a reduced number of bases, the synthetic motion from which the sequences were reconstructed is likely to have enjoyed greater regularity as well.

Despite not relying on temporal regularization to help in tracking the motion of the cardiac muscle (as in *e.g.* De Craene et al. (2012)), the temporal and spatial consistency of deformations was judged satisfactory, as evidenced by Fig. 1(a), 19. Still, incorporating temporal regularization and moving towards a large deformation framework (Beg et al., 2005; Arsigny et al., 2006) with geodesic-by-part trajectories may be fruitful for this application. It could also constitute a step in moving from a data-driven parametrization of displacements (beneficial to the quality of registration) towards an anatomically relevant parametrization. Indeed the proposed framework of basis selection may be suitable for learning a parametric atlas of motion from a small dataset of 3D+t images in the spirit of *e.g.* Allassonnière et al. (2007); Durrleman et al. (2013); Gori et al. (2013).

Acknowledgments. The first author was partly funded by the Microsoft Research – Inria Joint Centre, and part of this work was funded by the European Research Council through the ERC Advanced Grant MedYMA (2011-291080) on Biophysical Modeling and Analysis of Dynamic Medical Images.

AppendixA. Closed form regularizers for Gaussian reproducing kernel Hilbert spaces

Gaussian reproducing kernel Hilbert space. Given a $d \times d$ symmetric positive definite (s.p.d.) matrix S and the Gaussian kernel $K_S(x, y) = \exp -\frac{1}{2}(x-y)^\intercal S^{-1}(x-y)$, we consider the space $\mathcal{H}_S^d$ of integrable d-vector fields $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $\|f\|_S < +\infty$, where

$$\|f\|_S^2 = \frac{1}{(2\pi)^d} \int_\xi \|\widehat{f}(\xi)\|_2^2 \, \widehat{K}_S^{-1}(\xi) d\xi \tag{A.1}$$

involves the Fourier transform $\hat{f} = \mathcal{F}[f]$ of f, defined by $\hat{f}(\xi) = \int_x \exp\{-ix^\intercal \xi\} f(x) dx$. Endowed with the inner product A.2

$$\langle f | g \rangle_S = \frac{1}{(2\pi)^d} \int_\xi \widehat{f}(\xi)^\dagger \widehat{g}(\xi) \, \widehat{K}_S^{-1}(\xi) d\xi \tag{A.2}$$

$(\mathcal{H}_S^d, \langle \, | \, \rangle_S)$ is a reproducing kernel Hilbert space with attractive theoretical and algorithmic properties. Perhaps more intuitively, $\mathcal{H}_S^d$ is the completion of the space spanned by all finite combinations of $K_S(x, \cdot)\boldsymbol{\alpha}$, for $x \in \mathbb{R}^d$ and $\boldsymbol{\alpha} \in \mathbb{R}^d$.

A multiscale space. From A.1 and the properties of Gaussian kernels under Fourier transform, the atom $K_{\tilde{S}}(x, \cdot)\boldsymbol{\alpha}$ lies in $\mathcal{H}_S^m$ if and only if $\tilde{S} > S/2$, in the sense of positive definiteness. In particular, given a sequence $S_1 \leq \cdots \leq S_q$ of $d \times d$ s.p.d. matrices, their associated r.k.h.s. are nested: $\mathcal{H}_{S_1}^d \supseteq \cdots \supseteq \mathcal{H}_{S_q}^d$. This property leads to a principled framework to represent displacements in a multiscale fashion, jointly regularized at all scales.

Closed form regularizers. The successive partial derivatives $\partial_{x^{i_1} \ldots x^{i_p}} f$ of elements $f \in \mathcal{H}_S^d$ exist and all lie in $\mathcal{H}_S^d$ (Zhou,

2008). As such, we may consider the family of regularizers of the form $\mathcal{R}_D(f) = \|Df\|_S^2$, where D is a differential operator. For any f that can be written over a finite number of atoms $K_{S_k}(x_k,\cdot)\boldsymbol{\alpha}_k$, the properties of Gaussian kernels under Fourier transform, multiplication (by Gaussian kernels) and summation yield closed form expressions for such regularizers. We illustrate this on two classic penalty terms for registration: the membrane and bending energies ($D = \nabla^s$, $s = 1, 2$). Recall that for any p-uplet of integers $i_1 \cdots i_p \in \{1, \cdots, d\}$, $\mathcal{F}[\partial_{x^{i_1}\cdots x^{i_p}} f](\xi) = j^p \xi^{i_1} \cdots \xi^{i_p} \mathcal{F}[f](\xi)$. Thus for any $f \in \mathcal{H}_S^d$,

$$\mathcal{F}[\nabla \cdot f](\xi) = j\xi^\intercal \mathcal{F}[f](\xi)\,. \tag{A.3}$$

and for any even integer s,

$$\mathcal{F}[\nabla^s f](\xi) = (j\xi)^s \mathcal{F}[f](\xi) \tag{A.4}$$

The Fourier transform of a Gaussian kernel is given by $\mathcal{F}[K_{S_k}(x_k,\cdot)](\xi) = |2\pi S_k|^{1/2} e^{-j\xi^\intercal x_k} \exp\{-\frac{1}{2}\xi^\intercal S_k \xi\}$, which is again Gaussian. Regrouping the exponential factors when appropriate, we derive the following expression for inner products of the form $\langle \nabla^s(K_{S_k}\boldsymbol{\alpha}_k) \,|\, \nabla^s(K_{S_l}\boldsymbol{\alpha}_l) \rangle_S$:

$$C_{S_k,S_l,d} \cdot \boldsymbol{\alpha}_k^\intercal \left\{ \int_\xi \|\xi\|^{2s} \mathcal{F}[K_{S_{k,l}}(z_{k,l},\cdot)](\xi) d\xi \right\} \boldsymbol{\alpha}_l \tag{A.5}$$

where $C_{S_k,S_l,d} = \frac{1}{(2\pi)^d}\left(\frac{|S_k|\cdot|S_l|}{|S|\cdot|S_{k,l}|}\right)^{1/2}$ is a constant, $S_{k,l} = S_k + S_l - S$ and $z_{k,l} = x_l - x_k$. Recognizing the inverse Fourier transform of $\nabla^{2s}[K_{S_{k,l}}(\vec{0},\cdot)]$ evaluated at $z_{k,l}$, we finally obtain that

$$\left\| \nabla^s \left(\sum\nolimits_k K_{S_k}(x_k,\cdot)\boldsymbol{\alpha}_k \right) \right\|_S^2 = \boldsymbol{\alpha}^\intercal \left[\mathbf{R}_{k,l}^s \right]_{1\le k,l\le M} \boldsymbol{\alpha} \tag{A.6}$$

$$\mathbf{R}_{k,l}^s = \left(\frac{|S_k|\cdot|S_l|}{|S|\cdot|S_{k,l}|} \right)^{1/2} (-\Delta)^s [K_{S_{k,l}}](z_{k,l})\, \mathbf{I} \tag{A.7}$$

for s integer and $\mathbf{I}$ the $d \times d$ identity matrix. Straightforward analytical expressions of $(-\Delta)^s[K_{S_{k,l}}](z_{k,l})$ are obtained for $s = 1, 2$ by computing the derivatives of the Gaussian kernel.

AppendixB. Contribution of a basis to the log marginal likelihood

It follows from Eq. (22) that the log marginal likelihood $\mathcal{L} = \log p(\boldsymbol{t}|\mathbf{A}, \lambda, \sigma^2)$ is given up to additive constant by

$$\mathcal{L} = -\frac{1}{2}\left\{ |\mathbf{C}| + \boldsymbol{t}^\intercal \mathbf{C}^{-1} \boldsymbol{t} \right\} \tag{B.1}$$

with $\mathbf{C} = (\beta\mathbf{H})^{-1} + \Phi\mathbf{L}\Phi^\intercal$, where we define

$$\mathbf{L} \triangleq (\mathbf{A} + \lambda\mathbf{R})^{-1}\,. \tag{B.2}$$

Noting that $\mathbf{C}$ exclusively depends on the basis k via the kth diagonal coefficient of $\mathbf{A}$ and column of Φ, we would like to single out the contribution $l(\mathbf{A}_k)$ of any such basis ϕ_k to the global log marginal likelihood in the form:

$$\mathcal{L} = l(\mathbf{A}_k) + \mathcal{L}_{-k} \tag{B.3}$$

where $\mathcal{L}_{-k}$ does not depend on the basis k. If we denote by $\mathbf{L}_{-k}$ the inverse of the matrix obtained by removing the kth column from $\mathbf{L}^{-1} = \mathbf{A} + \lambda\mathbf{R}$ (or equivalently by setting $\mathbf{A}_k = +\infty$ in $\mathbf{L}$), we see from the Woodbury rank one matrix identity that $\mathbf{L} = \mathbf{L}_{-k} + \mathrm{U}_k \mathrm{L}_{kk} \mathrm{U}_k^\intercal$, with $\mathrm{U}_k^\intercal = \begin{pmatrix}(\lambda\mathbf{L}_{-k}\mathrm{R}_k)^\intercal & \mathrm{I}\end{pmatrix}$ and $\mathrm{L}_{kk} = (\mathbf{A}_k + \boldsymbol{\kappa}_k)^{-1}$, where the $d \times d$ matrix $\boldsymbol{\kappa}_k$ is defined as:

$$\boldsymbol{\kappa}_k \triangleq \lambda\mathbf{R}_{kk} - (\lambda\mathrm{R}_k)^\intercal \mathbf{L}_{-k} (\lambda\mathrm{R}_k) \tag{B.4}$$

By injecting this latter decomposition of $\mathbf{L}$ into the expression of $\mathbf{C}$, we derive a decomposition of $\mathbf{C}$ into the sum of a term that does not depend on the kth basis and of a rank one term:

$$\mathbf{C} = \mathbf{C}_{-k} + (\Phi\mathrm{U}_k)(\mathbf{A}_k + \boldsymbol{\kappa}_k)^{-1}(\Phi\mathrm{U}_k)^\intercal \tag{B.5}$$

Letting $\mathbf{C}_{-k}^{-1} \triangleq (\mathbf{C}_{-k})^{-1}$, a second application of rank one update identities for the determinant and the inverse gives the two following expressions B.6 and B.7 for the two terms in the right-hand side of the log marginal likelihood expression B.1:

$$|\mathbf{C}| = |\mathbf{C}_{-k}| \cdot |\mathbf{A}_k + \boldsymbol{\kappa}_k|^{-1} \cdot |\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k| \tag{B.6}$$

$$\boldsymbol{t}^\intercal \mathbf{C}^{-1} \boldsymbol{t} = \boldsymbol{t}^\intercal \mathbf{C}_{-k}^{-1} \boldsymbol{t} - \boldsymbol{q}_k^\intercal \left(\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k\right)^{-1} \boldsymbol{q}_k \tag{B.7}$$

We introduced the statistics $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ respectively defined as:

$$\boldsymbol{s}_k \triangleq (\Phi\mathrm{U}_k)^\intercal \mathbf{C}_{-k}^{-1} (\Phi\mathrm{U}_k) \tag{B.8}$$

$$\boldsymbol{q}_k \triangleq (\Phi\mathrm{U}_k)^\intercal \mathbf{C}_{-k}^{-1} \boldsymbol{t} \tag{B.9}$$

We thus retrieve the expression Eq. (24) for $l(\mathbf{A}_k)$, which was used without proof in Section 3.3. It is of practical significance to the algorithmic complexity of our schemes that the quantities involved ($\mathbf{L}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) do not actually depend on bases that are not in the active set $\mathcal{S}$ (*i.e.* all bases s.t. $\mathbf{A}_m = +\infty$). Similarly $\boldsymbol{s}_k$, $\boldsymbol{\kappa}_k$ and $\boldsymbol{q}_k$ only involve the set of active bases $\mathcal{S}$ augmented with the kth basis, due of the form of U_k.

The maximization of Eq. (24) under constraint that $\mathbf{A}_k$ is a symmetric positive semidefinite $d \times d$ matrix involves the gradient of the (unconstrained) function $l(\mathbf{A}_k)$:

$$\nabla l(\mathbf{A}_k) = -\boldsymbol{\sigma}_k \left\{ \boldsymbol{q}_k \boldsymbol{q}_k^\intercal - \boldsymbol{s}_k - \boldsymbol{s}_k (\mathbf{A}_k + \boldsymbol{\kappa}_k)^{-1} \boldsymbol{s}_k \right\} \boldsymbol{\sigma}_k \tag{B.10}$$

where $\boldsymbol{\sigma}_k$ is shorthand for $(\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k)^{-1}$, and $\nabla l(\mathbf{A}_k)$ is a $d \times d$ matrix. Since $\boldsymbol{\sigma}_k$ is symmetric positive definite and $\boldsymbol{q}_k \boldsymbol{q}_k^\intercal$ is of rank one, $\nabla l(\mathbf{A}_k)$ has at most one negative eigenvalue. More precisely, if $\boldsymbol{q}_k \boldsymbol{q}_k^\intercal - \boldsymbol{s}_k$ is negative then $\nabla l(\mathbf{A}_k)$ is positive definite for all $\mathbf{A}_k$ and the improper maximizer of $l(\mathbf{A}_k)$ lies at infinity $\mathbf{A}_k \to +\infty$. Otherwise there is exactly one negative eigenvalue and we look for maximizers of the form $\mathbf{A}_k^{-1} = \alpha_k^{-1} \boldsymbol{\eta}_k \boldsymbol{\eta}_k^\intercal$. This is consistent with the intuitive comment that $\mathbf{A}_k^{-1} \in \mathcal{M}_{d,d}$ cannot be fully determined from a single "observation" $\boldsymbol{q}_k \in \mathbb{R}^d$ and should be degenerate. Rewriting Eq. (24) as a function of α, η leads to maximizing B.11 under constraint that α is positive (dropping the index k for convenience). Note also that B.11 is invariant under reparametrization $\boldsymbol{\eta} \to \nu\boldsymbol{\eta}, \alpha \to \alpha/\nu^2$.

$$l(\alpha, \boldsymbol{\eta}) = -\log\left\{ 1 + \frac{\boldsymbol{\eta}^\intercal s \boldsymbol{\eta}}{\alpha + \boldsymbol{\eta}^\intercal \boldsymbol{\kappa} \boldsymbol{\eta}} \right\} + \frac{(\boldsymbol{q}^\intercal \boldsymbol{\eta})^2}{\alpha + \boldsymbol{\eta}^\intercal (\boldsymbol{\kappa} + \boldsymbol{s}) \boldsymbol{\eta}} \tag{B.11}$$

At a maximizer $\alpha^*, \boldsymbol{\eta}^* = \arg\max_{\alpha,\boldsymbol{\eta}} l(\alpha, \boldsymbol{\eta})$ the constraint is either active ($\alpha^* = 0$) or inactive ($\alpha^* > 0$). If inactive, the solution actually maximizes the unconstrained function B.11 and is given by $\alpha^* = \bar{\alpha}(\bar{\boldsymbol{\eta}})$, $\boldsymbol{\eta}^* = \bar{\boldsymbol{\eta}}$ where

$$\bar{\alpha}(\boldsymbol{\eta}) = \frac{(\boldsymbol{\eta}^\intercal \boldsymbol{s}\boldsymbol{\eta})^2}{(\boldsymbol{q}^\intercal\boldsymbol{\eta})^2 - \boldsymbol{\eta}^\intercal\boldsymbol{s}\boldsymbol{\eta}} - \boldsymbol{\eta}^\intercal\boldsymbol{\kappa}\boldsymbol{\eta}\,, \tag{B.12}$$

$$\bar{\boldsymbol{\eta}} = \boldsymbol{s}^{-1}\boldsymbol{q}\,. \tag{B.13}$$

In this case $l(\alpha^*, \boldsymbol{\eta}^*)$ is simply equal to $\bar{l}(\boldsymbol{\eta}^*)$, where $\bar{l}(\boldsymbol{\eta})$ is defined by B.14 with $\xi(\boldsymbol{\eta}) \triangleq (\boldsymbol{q}^\intercal\boldsymbol{\eta})^2/\boldsymbol{\eta}^\intercal\boldsymbol{s}\boldsymbol{\eta}$.

$$\bar{l}(\boldsymbol{\eta}) \triangleq -\log\xi(\boldsymbol{\eta}) + \xi(\boldsymbol{\eta}) - 1 \tag{B.14}$$

In addition $\bar{l}(\bar{\boldsymbol{\eta}})$ can be seen to always provide an upper bound to the maximum contribution of a basis to the evidence, $\max_{\alpha,\boldsymbol{\eta}} l(\alpha, \boldsymbol{\eta})$. In the case where the constraint is active, $\alpha^* = 0$, we numerically optimize over the unit sphere in $\mathbb{R}^d$ to find $\boldsymbol{\eta}^*$. This case occurs when the l_2-norm regularization is by itself sufficient along the direction $\boldsymbol{\eta}^*$, and no additional shrinkage is deemed necessary. To save on unnecessary computations, we first check that the upper bound $\bar{l}(\bar{\boldsymbol{\eta}})$ to the maximum contribution of the basis k to the evidence is superior to the current best contribution among bases already handled, as this is a necessary condition for $\mathbf{A}_k$ to be updated as this iteration.

AppendixC. Update of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{L}$

Updates of the moments of the posterior distribution $\boldsymbol{\mu}$, $\boldsymbol{\Sigma} = (\Phi^\intercal(\beta\mathbf{H})\Phi + \mathbf{A} + \lambda\mathbf{R})^{-1}$ and of $\mathbf{L} = (\mathbf{A} + \lambda\mathbf{R})^{-1}$ upon deletion from the model, update or addition to the model of a basis i are done similarly to Tipping et al. (2003) and follow from Woodbury identities. Denoting updated quantities with a tilde, we get in the case of deletion:

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \Sigma_i\Sigma_{ii}^{-1}\Sigma_i^\intercal\,, \tag{C.1}$$

$$\tilde{\mathbf{L}} = \mathbf{L} - \mathrm{L}_i\mathrm{L}_{ii}^{-1}\mathrm{L}_i^\intercal \tag{C.2}$$

and

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \Sigma_i(\Sigma_{ii}^{-1}\boldsymbol{\mu}_i)\,. \tag{C.3}$$

These rank one updates carefully avoid matrix-matrix products and have a $O(|\mathcal{S}|^2)$ complexity. In the case of the addition of a basis, we first compute the new column of $\tilde{\boldsymbol{\Sigma}}$ (resp. $\tilde{\mathbf{L}}$) before updating its full body as:

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \tilde{\Sigma}_i\tilde{\Sigma}_{ii}^{-1}\tilde{\Sigma}_i^\intercal\,, \tag{C.4}$$

$$\tilde{\mathbf{L}} = \mathbf{L} + \tilde{\mathrm{L}}_i\tilde{\mathrm{L}}_{ii}^{-1}\tilde{\mathrm{L}}_i^\intercal\,, \tag{C.5}$$

where the column $\tilde{\Sigma}_i$ (resp. $\tilde{\mathbf{L}}_i$) is given in $O(|\mathcal{S}|^2)$ by

$$\tilde{\Sigma}_i = \begin{pmatrix}\boldsymbol{\Sigma}\Pi_i\tilde{\Sigma}_{ii}\\ \tilde{\Sigma}_{ii}\end{pmatrix}, \quad \tilde{\mathrm{L}}_i = \begin{pmatrix}\mathbf{L}(\lambda\mathrm{R}_i)\tilde{L}_{ii}\\ \tilde{L}_{ii}\end{pmatrix} \tag{C.6}$$

and

$$\tilde{\Sigma}_{ii} = (\boldsymbol{s}_i + \boldsymbol{\kappa}_i + \mathbf{A}_i)^{-1}\,, \quad \tilde{L}_{ii} = (\boldsymbol{\kappa}_i + \mathbf{A}_i)^{-1}\,. \tag{C.7}$$

Π_i is the column vector of $d \times d$ matrices defined by $\Pi_i = \Phi^\intercal(\beta\mathbf{H})\phi_i + \lambda\mathbf{R}_i$. The case of the update of a basis i is treated as a deletion followed by an addition, updating $\boldsymbol{s}_i$ and $\boldsymbol{\kappa}_i$ in-between these actions as they are needed in Eq. (C.7).

AppendixD. Update of $\boldsymbol{s}_i$, $\boldsymbol{\kappa}_i$ and $\boldsymbol{q}_i$

From the resolvent identity, we note that $\Sigma\Phi^\intercal(\beta\mathbf{H})\Phi\mathbf{L} = \mathbf{L} - \Sigma$. Using such relationships after developing the factors in B.8 and B.9, we derive alternative expressions for $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$:

$$\boldsymbol{s}_m = \phi_m^\intercal(\beta\mathbf{H})\phi_m - \Pi_m^\intercal\boldsymbol{\Sigma}_{-m}\Pi_m + (\lambda\mathrm{R}_m)^\intercal\mathbf{L}_{-m}(\lambda\mathrm{R}_m) \tag{D.1}$$

$$\boldsymbol{q}_m = \phi_m^\intercal(\beta\mathbf{H})\boldsymbol{t} - \Pi_m^\intercal\boldsymbol{\Sigma}_{-m}\Phi^\intercal(\beta\mathbf{H})\boldsymbol{t} \tag{D.2}$$

where Π_m is a column vector of 3×3 matrices defined by $\Pi_m = \Phi^\intercal(\beta\mathbf{H})\phi_m + \lambda\mathbf{R}_m$. Π_m can be interpreted as the inner product of basis m with all the active bases w.r.t an appropriate metric, in the sense that its jth coefficient is given by: $\Pi_{jm} = \phi_j^\intercal(\beta\mathbf{H})\phi_m + \lambda\,(D\phi_j|D\phi_m)_{\mathcal{H}}$. In the specific case where $\lambda = 0$, we retrieve the quantities and expressions derived by Tipping et al. (2003) for the RVM. We found useful to introduce surrogate quantities $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ respectively defined according to D.3 and D.4:

$$\boldsymbol{t}_m \triangleq \phi_m^\intercal(\beta\mathbf{H})\phi_m - \Pi_m^\intercal\boldsymbol{\Sigma}\Pi_m + (\lambda\mathrm{R}_m)^\intercal\mathbf{L}(\lambda\mathrm{R}_m) \tag{D.3}$$

$$\boldsymbol{r}_m \triangleq \phi_m^\intercal(\beta\mathbf{H})\boldsymbol{t} - \Pi_m^\intercal\boldsymbol{\Sigma}\Phi^\intercal(\beta\mathbf{H})\boldsymbol{t} \tag{D.4}$$

These quantities merely differ from $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$ in that the index $-m$ was dropped from $\boldsymbol{\Sigma}_{-m}$ and $\mathbf{L}_{-m}$. Our underlying motivation is to update simpler quantities $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ that still retain a straightforward link to the statistics $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$ of interest for the computation of $l(\mathbf{A}_m)$. Indeed, for a basis l that does not lie in the model, $\boldsymbol{\Sigma}_{-l} = \boldsymbol{\Sigma}$ and $\mathbf{L}_{-l} = \mathbf{L}$. Therefore, the quantities under consideration coincide: $\boldsymbol{s}_l = \boldsymbol{t}_l$ and $\boldsymbol{q}_l = \boldsymbol{r}_l$. For a basis j that lies in the model and noting that $\boldsymbol{\Sigma}_{-j} = \boldsymbol{\Sigma} - \Sigma_j\Sigma_{jj}^{-1}\Sigma_j^\intercal$, we obtain the statistics of interest efficiently as:

$$\begin{aligned}\boldsymbol{s}_j = \boldsymbol{t}_j &+ \left[\Pi_j^\intercal\Sigma_j\right]\Sigma_{jj}^{-1}\left[\Pi_j^\intercal\Sigma_j\right]^\intercal\\ &- \left[(\lambda\mathrm{R}_j)^\intercal\mathrm{L}_j\right]L_{jj}^{-1}\left[(\lambda\mathrm{R}_j)^\intercal\mathrm{L}_j\right]^\intercal\end{aligned} \tag{D.5}$$

$$\boldsymbol{q}_j = \boldsymbol{r}_j + \left[\Pi_j^\intercal\Sigma_j\right]\Sigma_{jj}^{-1}\mu_j \tag{D.6}$$

Thus, we always maintain the quantities $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ (for every basis) and recompute $\boldsymbol{s}_m$ and $\boldsymbol{q}_m$ either at no cost for inactive bases or, for bases in the active set $\mathcal{S}$, in $O(|\mathcal{S}|\cdot d)$. Updates of $\boldsymbol{t}_m$ and $\boldsymbol{r}_m$ upon deletion from the model, update or addition to the model of a basis i are done similarly to Tipping et al. (2003), in $O(|\mathcal{S}|\cdot d)$ per basis. For instance, in the addition case, it follows from Woodbury identities that

$$\begin{aligned}\tilde{\boldsymbol{t}}_m = \boldsymbol{t}_m &- \left[\Pi_m^\intercal\tilde{\Sigma}_i\right]\tilde{\Sigma}_{ii}^{-1}\left[\Pi_m^\intercal\tilde{\Sigma}_i\right]^\intercal\\ &+ \left[(\lambda\mathrm{R}_m)^\intercal\tilde{\mathrm{L}}_i\right]\tilde{L}_{ii}^{-1}\left[(\lambda\mathrm{R}_m)^\intercal\tilde{\mathrm{L}}_i\right]^\intercal\end{aligned} \tag{D.7}$$

and

$$\tilde{\boldsymbol{r}}_m = \boldsymbol{r}_m - \left[\Pi_m^\intercal\tilde{\Sigma}_i\right]\boldsymbol{r}_i \tag{D.8}$$

where $\tilde{\boldsymbol{r}}_m$ and $\tilde{\boldsymbol{t}}_m$ denote updated quantities, as opposed to quantities prior to the update $\boldsymbol{r}_m$ and $\boldsymbol{t}_m$. The quantities indexed by i are computed (once for all bases) following AppendixC.

Allassonnière, S., Amit, Y., Trouvé, A., 2007. Towards a coherent statistical framework for dense deformable template estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69, 3–29.

Archambeau, C., Verleysen, M., 2007. Robust bayesian clustering. Neural Networks 20, 129–138.

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A log-euclidean framework for statistics on diffeomorphisms, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006. Springer, pp. 924–931.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95–113.

Ashburner, J., Ridgway, G.R., 2013. Symmetric diffeomorphic modelling of longitudinal structural MRI. Frontiers in Neuroscience 6.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision 61, 139–157.

Bishop, C.M., Tipping, M.E., 2000. Variational relevance vector machines, in: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.. pp. 46–53.

Bishop, C.M., et al., 2006. Pattern recognition and machine learning. volume 1. springer New York.

Broit, C., 1981. Optimal registration of deformed images .

Cachier, P., Ayache, N., 2004. Isotropic energies, filters and splines for vector field regularization. Journal of Mathematical Imaging and Vision 20, 251–265.

Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the pasha algorithm. Computer vision and image understanding 89, 272–298.

Chandrashekara, R., Mohiaddin, R.H., Rueckert, D., 2004. Analysis of 3-d myocardial motion in tagged mr images using nonrigid image registration. IEEE Transactions on Medical Imaging 23, 1245–1250.

De Craene, M., Marchesseau, S., Heyde, B., Gao, H., Alessandrini, M., Bernard, O., Piella, G., Porras, A., Saloux, E., Tautz, L., et al., 2013. 3d strain assessment in ultrasound (STRAUS): A synthetic comparison of five tracking methodologies. IEEE Transactions on Medical Imaging .

De Craene, M., Piella, G., Camara, O., Duchateau, N., Silva, E., Doltra, A., Dhooge, J., Brugada, J., Sitges, M., Frangi, A.F., 2012. Temporal diffeomorphic free-form deformation: Application to motion and strain estimation from 3d echocardiography. Medical Image Analysis 16, 427–450.

Durrleman, S., Allassonnière, S., Joshi, S., 2013. Sparse adaptive parameterization of variability in image ensembles. International Journal of Computer Vision 101, 161–183.

Fanello, S.R., Keskin, C., Kohli, P., Izadi, S., Shotton, J., Criminisi, A., Pattacini, U., Paek, T., 2014. Filter forests for learning data-dependent convolutional kernels, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE. pp. 1709–1716.

Gee, J.C., Bajcsy, R.K., 1998. Elastic matching: Continuum mechanical and probabilistic analysis. Brain warping 2.

Gori, P., Colliot, O., Worbe, Y., Marrakchi-Kacem, L., Lecomte, S., Poupon, C., Hartmann, A., Ayache, N., Durrleman, S., 2013. Bayesian atlas estimation for the variability analysis of shape complexes, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 267–274.

Groves, A.R., Beckmann, C.F., Smith, S.M., Woolrich, M.W., 2011. Linked independent component analysis for multimodal data fusion. Neuroimage 54, 2198–2217.

Hachama, M., Desolneux, A., Richard, F.J., 2012. Bayesian technique for image classifying registration. Image Processing, IEEE Transactions on 21, 4080–4091.

Heiberg, E., Wigstrom, L., Carlsson, M., Bolger, A., Karlsson, M., 2005. Time resolved three-dimensional automated segmentation of the left ventricle, in: Computers in Cardiology, 2005, IEEE. pp. 599–602.

Janoos, F., Risholm, P., Wells III, W., 2012. Bayesian characterization of uncertainty in multi-modal image registration. Biomedical Image Registration , 50–59.

Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N., 2014. Sparse bayesian registration, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 235–242.

MacKay, D.J., 1992. Bayesian interpolation. Neural computation 4, 415–447.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023–1032.

Mohamed, S., Heller, K.A., Ghahramani, Z., 2012. Bayesian and l1 approaches for sparse unsupervised learning. Proceedings of the 29th International Conference in Machine Learning .

Richard, F.J., Samson, A.M., Cuénod, C.A., 2009. A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences. Stat Comput 19.

Risholm, P., Janoos, F., Norton, I., Golby, A.J., Wells III, W.M., 2013. Bayesian characterization of uncertainty in intra-subject non-rigid registration. Medical image analysis 17, 538–555.

Rohde, G.K., Aldroubi, A., Dawant, B.M., 2003. The adaptive bases algorithm for intensity-based nonrigid image registration. Medical Imaging, IEEE Transactions on 22, 1470–1479.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast mr images. IEEE Transactions on Medical Imaging 18, 712–721.

Sabuncu, M.R., Van Leemput, K., 2011. The relevance voxel machine (rvoxm): a bayesian method for image-based prediction, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer, pp. 99–106.

Shi, W., Jantsch, M., Aljabar, P., Pizarro, L., Bai, W., Wang, H., ORegan, D., Zhuang, X., Rueckert, D., 2013. Temporal sparse free-form deformations. Medical Image Analysis 17, 779–789.

Simpson, I.J., Schnabel, J.A., Groves, A.R., Andersson, J.L., Woolrich, M.W., 2012. Probabilistic inference of regularisation in non-rigid registration. NeuroImage 59, 2438–2451.

Simpson, I.J., Woolrich, M.W., Cardoso, M.J., Cash, D.M., Modat, M., Schnabel, J.A., Ourselin, S., 2013. A bayesian approach for spatially adaptive regularisation in non-rigid registration. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013 , 10–18.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: A survey. IEEE Transactions on Medical Imaging 32, 1153–1190.

Stefanescu, R., Pennec, X., Ayache, N., 2004. Grid powered nonlinear image registration with locally adaptive regularization. Medical image analysis 8, 325–342.

Thirion, J.P., 1998. Image matching as a diffusion process: an analogy with maxwell's demons. Medical image analysis 2, 243–260.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267–288.

Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. The Journal of Machine Learning Research 1, 211–244.

Tipping, M.E., Faul, A.C., et al., 2003. Fast marginal likelihood maximisation for sparse bayesian models, in: Workshop on artificial intelligence and statistics, Jan.

Tipping, M.E., Lawrence, N.D., 2005. Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis. Neurocomputing 69, 123–141.

Tobon-Gomez, C., De Craene, M., McLeod, K., Tautz, L., Shi, W., Hennemuth, A., Prakosa, A., Wang, H., Carr-White, G., Kapetanakis, S., et al., 2013. Benchmarking framework for myocardial tracking and deformation algorithms: An open access database. Medical Image Analysis .

Wachinger, C., Golland, P., Reuter, M., Wells, W., 2014. Gaussian process interpolation for uncertainty estimation in image registration, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 267–274.

Wells III, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multimodal volume registration by maximization of mutual information. Medical Image Analysis 1, 35–51.

Wipf, D.P., Nagarajan, S.S., 2008. A new view of automatic relevance determination, in: Advances in neural information processing systems, pp. 1625–1632.

Zhou, D.X., 2008. Derivative reproducing properties for kernel methods in learning theory. Journal of computational and Applied Mathematics 220, 456–463.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.