Sparse Bayesian registration of medical images for self-tuning of parameters and spatially adaptive parametrization of displacements

Loïc Le Folgoc, Hervé Delingette, Antonio Criminisi, Nicholas Ayache

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

We extend Bayesian models of non-rigid image registration to allow not only for the automatic determination of registration parameters (such as the trade-off between image similarity and regularization functionals), but also for a data-driven, multiscale, spatially adaptive parametrization of deformations. Adaptive parametrizations have been used with success 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. Under the proposed model, a sparsity-inducing prior on transformation parameters complements the classical smoothness-inducing prior, and favors parametrizations that use few degrees of freedom. As a result, 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. The space of possible parametrizations consists of arbitrary combinations of basis functions chosen among any preset, widely overcomplete (and typically multiscale) dictionary. Inference is tackled in an efficient Variational Bayes framework. In addition we propose a flexible mixture-of-Gaussian model of data that proves to be more faithful for a variety of image modalities than the sum-of-squared differences. The performance of the proposed approach is demonstrated on time series of (cine and tagged) magnetic resonance and echocardiographic cardiac images. The proposed algorithm matches the state-of-the-art on benchmark datasets evaluating accuracy of motion and strain, and is highly automated.

Keywords: Non-rigid registration, Bayesian modelling, Sparse structured prior, Variational Bayes, ARD, Cardiac Imaging

1. Introduction

Non-rigid image registration is the ill-posed task of inferring a deformation Ψ from a pair of observed (albeit 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 $\mathcal{R}$:

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

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 trade-off 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. Unfortunately, this assumption generally does not hold. Probabilistic interpretations of registration recently emerged as a way to automate the process (Richard et al., 2009; Simpson et al., 2012; Risholm et al., 2013). 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 trade-off parameter itself can be treated as a hidden random variable, equipped with a broad prior distribution, and jointly inferred with Ψ or integrated out. In practice, analytical inference is precluded and various strategies are devised for approximate inference. Risholm et al. (2013) characterize the distributions of interest from MCMC samples. This is a most principled and accurate approach provided that enough samples can be drawn within the available computational bud-

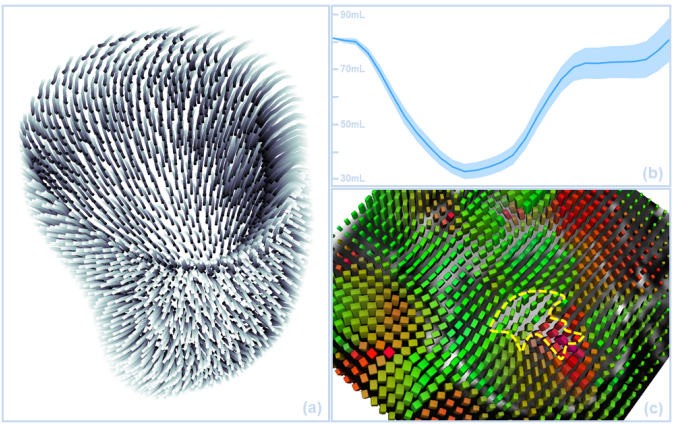


Figure 1: (a) Trajectories of points on the endocardium, following the registration of a time series of cardiac MR images by the proposed approach. (b) LV volume over time and 99.7% confidence interval. (c) Tensor visualization of directional uncertainty at end-systole, rasterized at voxel centers of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

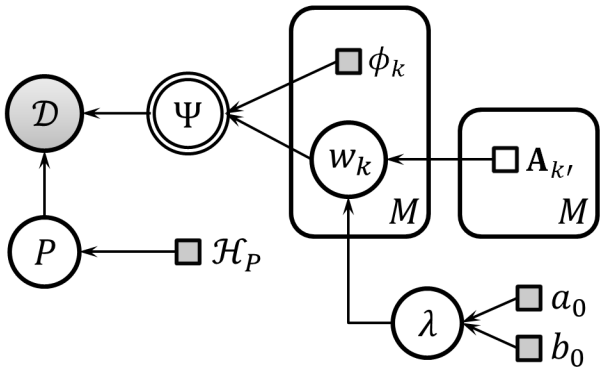


Figure 2: Graphical model of registration. The generative model of data D involves a transformation Ψ of space, and noise governed by a set of underlying parameters P. Hyperpriors (with hyperparameters $\mathcal{H}_P$) are in turn imposed over the noise parameters. The transformation is parametrized as a linear combination of predefined basis functions $\{\phi_k,\, k = 1 \cdots M\}$ with associated weights $\boldsymbol{w}_k$. Priors on the transformation smoothness and on the relevance of individual bases introduce additional parameters λ and $\mathbf{A}_{k'}$. Random variables are circled, hyperparameters are squared. Arrows capture conditional dependencies. Shaded nodes are observed variables or fixed hyperparameters. The transformation Ψ is fully determined by its parent nodes (the ϕ_k and $\boldsymbol{w}_k$), hence the doubly circled node. The content of plates is replicated (M times).

get. 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 Bayes framework (Simpson et al., 2012, 2015). 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. In this article, we propose to extend the Bayesian framework of registration to automatically select the optimal location and scale of bases parametrizing the transformation.

Spatial refinement of the parametrization was previously handled heuristically (Rohde et al., 2003), or led to alternative formulations of registration via spatially anisotropic filtering Stefanescu et al. (2004). Dynamic refinements of the displacement space have also been proposed by Glocker et al. (2008); Parisot et al. (2014) for MRF-based discrete registration. The displacement quantization is refined using local, min-marginal based estimates of uncertainty. Dynamic quantization (see also Tang and Hamarneh (2013) and Heinrich et al. (2016)) does not affect the registration energy however: its purpose is simply to accelerate convergence towards the optimum. In our work, the registration cost function forces the complexity of the mapping to adapt to the underlying dataset: finer bases are introduced only in the presence of coherent image information and motion, while coarser bases ensure better extrapolation of the motion to textureless, uninformative regions. In that spirit of model selection, Stewart et al. (2003) use an information criterion to choose regionally among a limited pool of deformation models (*e.g.* similarity, affine, quadratic), but do not address combinatorial issues arising in fully non-rigid registration from the number of possible parametrizations. Shi et al. (2013) couple sparse optimization with a multiscale free-form representation of deformations, demonstrating gains in registration accuracy. Here and to our knowledge, for the first time, basis selection in registration is approached on principled grounds within a probabilistic framework.

We propose a Bayesian model of registration that allows to automatically infer from the data the optimal parametrization of displacements, along with all model parameters. The inference scheme is efficient and tractable for real scale non-rigid registration tasks. The model and inference strategy are based on the Relevance Vector Machine (Tipping, 2001; Tipping et al., 2003), a generic approach to sparse regression and classification. To make it suitable for registration, where smooth solutions are looked for, we extend it to richer Gaussian priors with arbitrarily structured covariance, at no cost in algorithmic complexity. We also generalize the approach to multivalued regression (regression of vector fields), so as to preserve the natural invariance of the problem to changes of coordinate frames.

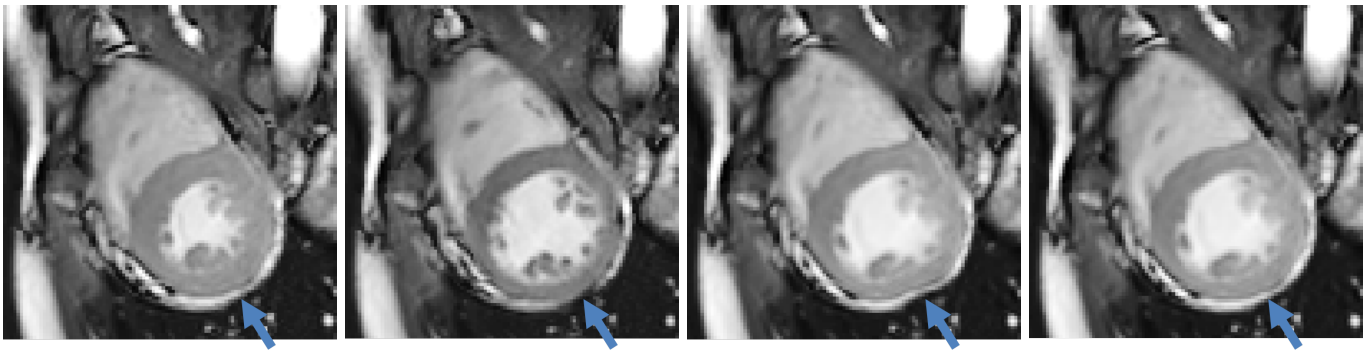
This article expands on earlier work of the authors (Le Folgoc et al., 2014) in several ways. Inference is fully presented within a variational Bayes framework. We propose a different approximation of the likelihood term, effectively removing a computational bottleneck: the voxelwise, local optimization of the image similarity *via* dense block-matching. It is replaced by a step where the registration energy is optimized w.r.t. the reduced parametrization. Finally we introduce a flexible noise model that is more robust to acquisition noise and artefacts, adapting over a range of image modalities.

Section 2 describes the statistical model of pairwise registration. The inference strategy is exposed in section 3. Section 4 reports experiments 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

Image registration assumes images to relate *via* some transformation of space such as, in a medical context, when imaging the motion of organs throughout a sequence of time frames. Registration then aims at recovering the unknown transformation of space from the observed data, which is formally an inference problem. We now specify a generative model of the data given the transformation Ψ, along with a sparse structured prior over the admissible set of transformations. Fig. 2 provides a graphical depiction of the model.

The sparse structured prior model was previously proposed by Sabuncu and Van Leemput (2012) for image-based tasks of classification and regression. Inference followed the guidelines of Tipping (2001), and was later accelerated for specific priors exploiting sparsely connected graphs (Ganz et al., 2013). In Section 3 we develop alternative inference schemes that are applicable with no restriction: the structured part of the prior, irrelevant to the algorithmic complexity, may be arbitrarily defined. The gain in algorithmic complexity reflects the way in which the later work of Tipping et al. (2003) accelerates the original Relevance Vector Machine (Tipping, 2001). This ef-



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. 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. The inferred motion is qualitatively closer to our expectations.

fectively renders the approach applicable to non-rigid registration.

2.1. Data Likelihood

A good transformation Ψ should adequately map the datasets, up to some misalignment and residual error attributable to the data formation process. The knowledge of this process is captured in a likelihood model, which assigns a probability $p(D|\Psi;P)$ for the data D to be observed under some transformation Ψ (often conditioned on a set of hyperparameters P). The likelihood typically assumes the form of a Boltzmann distribution:

$$p(D|\Psi;P) \propto \exp -\mathcal{D}(D,\Psi;P)\,, \tag{2}$$

which explicitely bridges the gap with the classic optimization framework of Eq. (1). For pairwise registration of images, 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 stands for the inverse variance (precision) of the noise across the image. The SSD model can be described in a more familiar manner by the energy of Eq. (4), where $\{v_i\}_{i=1}^N$ is the list of voxel centers in the fixed image and $V_i = \Psi^{-1}(v_i)$ are the paired coordinates in the moving image.

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

Since the SSD is quadratic w.r.t to intensity differences of paired voxels, both the penalty for intensity discrepancies and the *rate* at which it grows can become arbitrarily high. As seen in Fig. 3, this renders registration vulnerable to strong local intensity biases, introduced for instance by acquisition artefacts or by topology changes in the imaged objects. In addition residual misalignments between structures of interest tend to yield higher intensity residuals than those observed at background voxels (see for instance 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 is neither robust nor flexible enough.

To address this limitation we propose to model the noise $e_i \sim \sum_{1\leq l\leq L}\pi_l\mathcal{N}(0,\beta_l)$ at each voxel $i = 1\cdots N$ with a mixture of L Gaussian distributions. Implicitely at each voxel, the residue $e_i = J[v_i] - I[V_i]$ is independently assigned to one of the L components, with π_l the probability of being assigned to the lth component $\mathcal{N}(0,\beta_l)$. Introducing a set of binary assignment variables $\{z_{i1}\cdots z_{iL}\}$ for the ith voxel, such that $z_{il} = 1$ if assigned to the lth component and $z_{il} = 0$ otherwise, the above can be summarized as:

$$p(J|I,\Psi,\boldsymbol{\beta},\boldsymbol{z}) \propto \prod_{i=1}^{N}\exp -\frac{1}{2}\left(\textstyle\sum_{l=1}^{L} z_{il}\beta_l\right)\left(J[v_i] - I[V_i]\right)^2 \tag{5}$$

$$p(\boldsymbol{z}|\boldsymbol{\pi}) \propto \prod_{i=1}^{N}\prod_{l=1}^{L}\pi_l^{z_{il}} \tag{6}$$

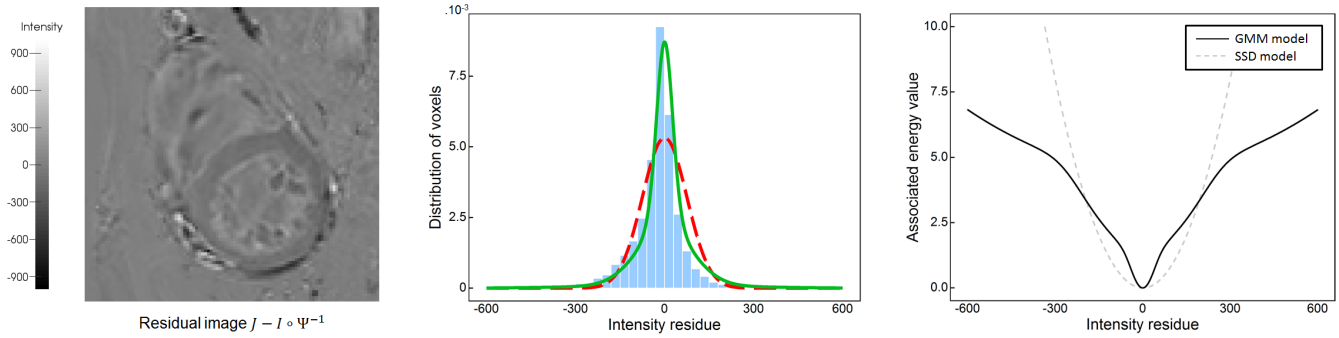


Figure 4: (Left) Example residual image following registration. Artefacts and structures that changed appearance from one image to the other stand out much unlike ambient noise. Note that the intensity of the cardiac muscle itself differed in the pair of images. (Middle) Histogram of intensity residuals, with SSD and GMM fits overlayed respectively in red and green. (Right) Energy profiles for the SSD (grey dashes) and GMM (black line). The voxelwise penalty is plotted as a function of the intensity residual. GMM achieves robustness thanks to concave inflexions that result in a soft threshold on the penalty incurred for large intensity residuals.

with $\boldsymbol{\beta} = \{\beta_1 \cdots \beta_l\}$, $\boldsymbol{\pi} = \{\pi_1 \cdots \pi_l\}$ and $\boldsymbol{z} = \{z_{il}\}_{i=1\cdots N, l=1\cdots L}$. This yields a spatially varying model of noise that is better suited to render the complexity of noise patterns in medical images. Unlike in previous work (Simpson et al., 2013; Le Folgoc et al., 2014) the noise here is not assumed to vary smoothly across the image, as patterns arising from misalignment and imaging artefacts are local in nature. Integrating over assignment variables, we explicitely retrieve the mixture-of-Gaussian structure:

$$p(J|I,\Psi,\boldsymbol{\beta},\boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{l=1}^{L} \frac{\pi_l}{Z_l} \exp -\frac{\beta_l}{2} \left(J[v_i] - I[V_i]\right)^2 \quad (7)$$

where $Z_l = \sqrt{2\pi/\beta_l}$ stands for the normalizing constant for the Gaussian probability distribution function. The corresponding data matching energy is given in Eq. (8):

$$\mathcal{D}_{\boldsymbol{\beta},\boldsymbol{\pi},L}(J;I,\Psi) = -\sum_{i=1}^{N} \log \sum_{l=1}^{L} \frac{\pi_l}{Z_l} \exp -\frac{\beta_l}{2} \left(J[v_i] - I[V_i]\right)^2 \quad (8)$$

Fig. 4b shows the histogram of intensity residuals for the example registration of Fig. 3, along with the learned Gaussian mixture (jointly fit during registration). The profiles of the standard SSD loss and the Gaussian mixture (GMM) loss are displayed in Fig 4c. 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. Mixtures of Gaussian (or Student-t) distributions have long been used as building blocks for robust autoregressive models (Roberts and Penny, 2002; Tipping and Lawrence, 2005), including in medical imaging (Penny et al., 2007) but remain uncommon for registration (Leventon and Grimson, 1998).

A limitation of SSD shared by all aforementioned variants is to assume that voxelwise intensity residuals are independent. This assumption does not hold (Simpson et al., 2012). In practice, 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 over-confidence in the data term. On the other hand, modeling precisely the noise structure would come at a significant computational cost. Here, we follow Simpson et al. (2012) in artificially downweighting the data term by a factor α that captures redundancies in voxelwise observations, based on a virtual decimation procedure suggested by Groves et al. (2011).

2.2. Representation of displacements

We proceed in a small deformation framework, $\Psi^{-1} = \mathrm{Id} + \mathrm{u}$, with a parameterized representation of the displacement field $\mathrm{u}\colon \mathrm{x} \in \mathbb{R}^d \mapsto \mathrm{u}(\mathrm{x}) \in \mathbb{R}^d$. The displacement field is expressed over a dictionary $\{\phi_k\}_{k=1}^M$ of Gaussian radial basis functions, $\phi_k(\mathrm{x}) = K_{S_k}(\mathrm{x}_k, \mathrm{x})\,\mathbf{I}$ where $\mathbf{I}$ is the $d \times d$ identity matrix and

$$K_S(x,y) = \exp -\frac{1}{2}(x-y)^\intercal S^{-1}(x-y)\,. \quad (9)$$

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

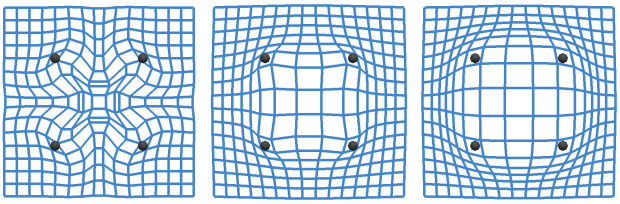
$$\mathrm{u}_{\boldsymbol{w}}(\mathrm{x}) = \sum_{1 \le k \le M} \phi_k(\mathrm{x})\boldsymbol{w}_k = \boldsymbol{\phi}(\mathrm{x})\boldsymbol{w}\,. \quad (10)$$

$\boldsymbol{\phi}(\mathrm{x}) = \left(\phi_1(\mathrm{x}) \cdots \phi_M(\mathrm{x})\right)$ and $\boldsymbol{w}^\intercal = \left(\boldsymbol{w}_1^\intercal \cdots \boldsymbol{w}_M^\intercal\right)$ are respectively the concatenation, for $k = 1 \cdots M$, of $\phi_k(\mathrm{x})$ and $\boldsymbol{w}_k$.

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 yields a redundant, multiscale representation of displacements. Larger kernels make the representation more compact, whereas smaller kernels allow to capture finer local details. The genericity of the approach w.r.t. the choice of dictionary is discussed in AppendixH.

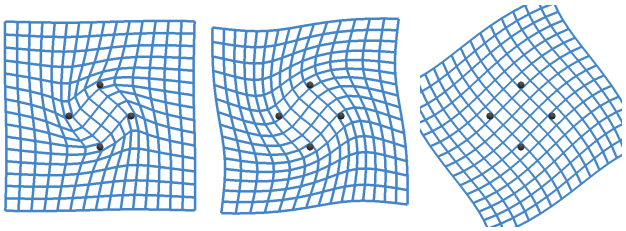
2.3. Transformation Prior

In non-rigid registration, the displacement $\mathrm{u}_{\boldsymbol{w}}$ is insufficiently constrained by the data and some regularizing prior has to be



Penalizing compressibility ($D = \text{div}$) | 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.



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 affects the area of influence of the data points and the scale at which the regressed transform resembles a global rotation.

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

$$p(\boldsymbol{w}|\lambda, \{\mathbf{A}_k\}) \propto \exp -\frac{1}{2}\left\{\lambda \boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w} + \sum_{k=1}^{M} \boldsymbol{w}_k^\intercal \mathbf{A}_k \boldsymbol{w}_k\right\} \tag{11}$$

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

Regularity control. Gaussian priors in the form of Eq. (12) 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.

$$q(\boldsymbol{w}|\lambda) \propto \exp -\frac{1}{2}\lambda \boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w} \tag{12}$$

The structure of the precision matrix $\mathbf{R}$ can be adjusted to penalize the magnitude $\|D\mathrm{u}\|^2$ of the first derivative of the displacement field (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 recall in AppendixH how to compute $\mathbf{R}$ efficiently using Fourier analysis for general families of basis functions instead of relying on costly numerical integration. With the parametrization of displacements given in section 2.2, classical energies are in fact implemented in closed-form. In this work, we specifically rely on a bending (thin-plate) energy ($D = \nabla^2$).

Basis selection. The second factor in our prior, recalled in Eq. (13), induces the desired basis selection mechanism.

$$q(\boldsymbol{w}|\{\mathbf{A}_k\}) \propto \prod_{k=1}^{m} \exp -\frac{1}{2}\boldsymbol{w}_k^\intercal \mathbf{A}_k \boldsymbol{w}_k \tag{13}$$

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 $\boldsymbol{w}_k$. Each $\mathbf{A}_k$ is an arbitrary $d \times d$ symmetric positive matrix, so that $\boldsymbol{w}_k$ can be penalized in a different manner along different orientations. The improper limit case of infinite $\mathbf{A}_k$ actually constrains $\boldsymbol{w}_k$ to be null and thus forbids the use of ϕ_k to represent the signal. In section 3 we determine optimal values of the set $\{\mathbf{A}_k\}_{k=1\cdots M}$ in a principled manner, from which most of them turn out to be infinite: we thus obtain a sparse representation of the displacement from the initial, over-complete dictionary.

Introducing $\mathbf{A} \triangleq \text{diag}(\mathbf{A}_1 \cdots \mathbf{A}_M)$ as the block diagonal matrix whose kth $d \times d$ diagonal block is $\mathbf{A}_k$, the full prior takes the more compact form $p(\boldsymbol{w}|\lambda, \mathbf{A}) \propto \exp -\frac{1}{2}\boldsymbol{w}^\intercal(\mathbf{A} + \lambda\mathbf{R})\boldsymbol{w}$.

2.4. Hyperpriors

The value of model parameters $\boldsymbol{\beta}$, $\boldsymbol{\pi}$, λ and $\{\mathbf{A}_k\}_{k=1\cdots M}$ is unknown. We regard them as additional model variables and endow them with prior distributions. When possible, the choice of conjugate priors facilitates inference. The noise levels $\boldsymbol{\beta} = \{\beta_1 \cdots \beta_L\}$ for each component of the Gaussian mixture are assigned independent Gamma priors $\Gamma(\beta_l|c_0, d_0) \propto \beta_l^{c_0-1} e^{-d_0\beta_l}$. The noise mixture proportions $\boldsymbol{\pi} = \{\pi_1 \cdots \pi_L\}$ are equipped with a Dirichlet prior $\text{Dir}(\boldsymbol{\pi}|\eta_0) \propto \prod_l \pi_l^{\eta_0-1}$. λ is endowed with a Gamma prior $\Gamma(\lambda|a_0, b_0) \propto \lambda^{a_0-1} e^{-b_0\lambda}$. In absence of strong prior knowledge, broad uninformative priors can be chosen ($a_0 = b_0 = c_0 = d_0 = \eta_0 \to 0$).

An improper uniform prior is taken over basis penalties $\mathbf{A}_k$, with the added benefit of making inference invariant to rescaling of basis functions. Moreover given the inference strategy of section 3, AppendixB and AppendixE, optimality conditions state that $\mathbf{A}_k^{-1} = \alpha_k^{-1}\boldsymbol{n}_k\boldsymbol{n}_k^\intercal$ is at most rank-one, with $\alpha_k \in \mathbb{R}_+ \cup \{+\infty\}$. We found advantageous to further restrain α_k to be either 0 or $+\infty$. In other words, along any given direction, $\mathbf{A}_k$ either constrains $\boldsymbol{w}_k$ to be null or does not constrain it whatsoever. This prevents direct competition between the two regularization mechanisms of Eq. (12) and (13). Moreover, the Gamma prior over λ then becomes conjugate to $p(\boldsymbol{w}|\mathbf{A}, \lambda)$.

2.5. Related work: Sparse Coding & Registration

Sparsity-inducing priors have a two-fold motivation. The first benefit is in terms of algorithmic complexity. Unless resorting to low parametric models, the size of the parametrization makes 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 discard significant interactions induced by the data and priors. Secondly, basis selection mechanisms adaptively constrain the space of deformations, automatically tuning the degrees of freedom to the smallest set sufficient 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 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), but also in medical image registration (Shi et al., 2013). Fig. 6 gives a naive insight into the key impact of scale when limited data is available.

L_1 priors have been widely used in all areas of sparse coding, including for registration (Shi et al., 2013). Other sparsity-inducing norms such as k-support norms and variants (Argyriou et al., 2012; Belilovsky et al., 2015a), that improve over the performance of the L_1 norm w.r.t. the degree of sparsity in presence of strongly correlated explanatory variables, have recently been proposed. They were shown to be attractive on tasks of functional MR imaging (Jenatton et al., 2012; Belilovsky et al., 2015b). Here, we turn instead towards sparse Bayesian learning, with the prospect of joint estimation of model parameters and that of uncertainty quantification. For an extensive review of sparse methods, we refer the reader to the work of Bach et al. (2012), and to that of Mohamed et al. (2012) for a benchmark of L_1 and bayesian sparse learning methods. The prior of Eq. (13) 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. (11) 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.

3. Model Inference

Bayesian inference summarizes both prior knowledge $p(\boldsymbol{\theta})$ and data-driven information $p(D|\boldsymbol{\theta})$ on model parameters and hyperparameters $\theta = \{\boldsymbol{w}, \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{z}, \lambda, \mathbf{A}\}$ within a posterior distribution

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(D)} \,. \tag{14}$$

The goal of Bayesian inference is thus to characterize the joint posterior $p(\boldsymbol{\theta}|D)$ or to characterize marginals of interest, such as the marginal posterior distribution $p(\boldsymbol{w}|D)$ of transformation parameters for the purpose of registration. Exact inference is precluded and we proceed in the framework of variational Bayes (VB) inference (Bishop et al., 2006).

3.1. Variational Bayes inference

VB inference approximates the true posterior $p(\boldsymbol{\theta}|D)$ among a restricted family of variational posterior distributions $q(\boldsymbol{\theta})$ that benefits analytical and computational derivations. The objective under VB inference is to minimize the Kullbach-Leibler divergence $\mathrm{KL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|D))$ or equivalently to maximize a lower bound $\mathcal{L}(q)$ of the log-evidence. This equivalence follows from the following identity, where on the left-hand side the evidence $p(D)$ for the model is constant w.r.t. q:

$$\log p(D) = \mathrm{KL}\left(q(\boldsymbol{\theta}) \,\middle\|\, p(\boldsymbol{\theta}|D)\right) \underbrace{- \int_{\boldsymbol{\theta}} q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, D)} d\boldsymbol{\theta}}_{\triangleq \mathcal{L}(q)} \tag{15}$$

When the true posterior lies within the variational family, the (non negative) Kullbach-Leibler divergence is minimized (zero) when $q^*(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|D)$. In practice however, the choice of variational family q reflects a trade-off between the accuracy of the approximation $q^*(\boldsymbol{\theta})$ and its actual tractability.

The mean-field approximation assumes q to factorize over subsets of model variables. In our case, we consider variational distributions for which the transformation parameters and individual penalties, the regularization level, the noise levels, the mixture proportions and the voxel assignments factorize:

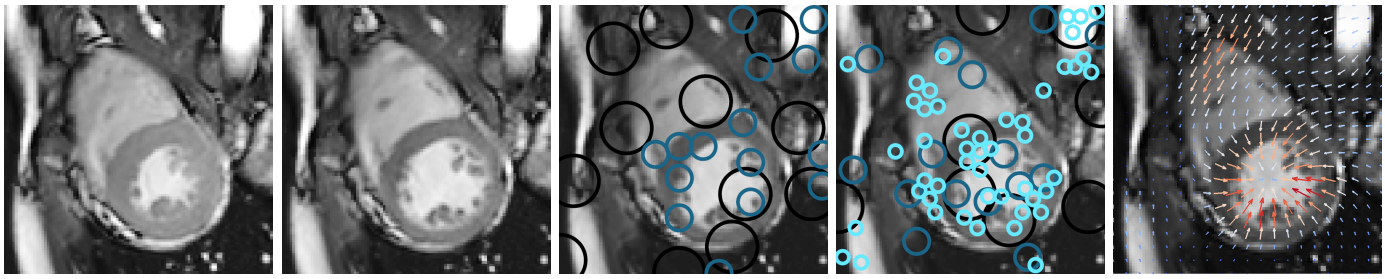
$$q(\boldsymbol{\theta}) = q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w}, \mathbf{A}) q_\lambda(\lambda) q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) q_{\boldsymbol{\pi}}(\boldsymbol{\pi}) q_z(\boldsymbol{z}) \,. \tag{16}$$

Let q_I be any one of the individual factors and q_{-I} the product of remaining factors, *e.g.* q_λ and $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w}, \mathbf{A}) q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) q_{\boldsymbol{\pi}}(\boldsymbol{\pi}) q_z(\boldsymbol{z})$ respectively. Let $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_{-I}$ denote corresponding subsets of variables within $\boldsymbol{\theta}$. Exploiting the factorization, each factor $q_I(\boldsymbol{\theta}_I) \approx p(\boldsymbol{\theta}_I|D)$[1] can be seen to give an approximation of a given marginal of interest. The optimum q^* among variational posteriors compatible with this factorization is known from calculus of variations to satisfy, for each individual factor q_I:

$$\log q_I^* = \langle \log p(\boldsymbol{\theta}, D) \rangle_{q_{-I}^*} + \text{const} \,. \tag{17}$$

Here, $\langle \cdot \rangle_{q_{-I}}$ denotes expectation w.r.t. $q_{-I}(\boldsymbol{\theta}_{-I})$. Eq. (17) naturally suggests inference schemes that update each factor q_I in turn until convergence, guaranteeing decrease of the objective $\mathcal{L}(q)$ at each iteration. Moreover when the Bayesian model uses conjugate exponential distributions, mean-field VB updates are considerably simplified. Each factor $q_I^*(\boldsymbol{\theta}_I)$ lies in the same exponential family as the corresponding prior $p(\boldsymbol{\theta}_I|\boldsymbol{\theta}_{-I})$ so that VB inference resolves into much more practical updates of the exponential distribution *parameters*. For instance, the

[1] $q_I(\boldsymbol{\theta}_I) = q_I(\boldsymbol{\theta}_I) \int_{\boldsymbol{\theta}_{-I}} q_{-I}(\boldsymbol{\theta}_{-I}) \, d\boldsymbol{\theta}_{-I} = \int_{\boldsymbol{\theta}_{-I}} q(\boldsymbol{\theta}) \, d\boldsymbol{\theta}_{-I} \approx \int_{\boldsymbol{\theta}_{-I}} p(\boldsymbol{\theta}|D) \, d\boldsymbol{\theta}_{-I}$



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 (cf. sec 4.3), respectively at ES (a) and ED (b). (d,e) Bases selected in the initial steps of the algorithm vs. at the end. The locations and scales of the Gaussian RBFs are indicated by circles (isocontour at 1 std). (c) Inverse displacement field output by the algorithm (scale factor: 2), smoothly varying across the whole image.

optimal variational posterior for λ is again Gamma distributed, $q^*_\lambda(\lambda) = \Gamma(\lambda|a,b)$, with closed form expressions for a, b. In the proposed model, most conditional probabilities do belong to conjugate exponential families. One exception is the likelihood $p(D|\boldsymbol{w},\boldsymbol{\beta},z)$. To enable analytical derivations of $q_{\boldsymbol{\beta}}$, q_z and $q_{\boldsymbol{w},\mathbf{A}}$, a Gaussian approximation of the likelihood is used (section 3.4). In addition the family of variational posteriors $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A})$ is further restricted (section 3.2).

3.2. *Constraining $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A})$ for fast sparse Bayes inference*

From the factorization of Eq. (16) and optimality conditions of Eq. (17), $q^*_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A})$ unfortunately does not simplify into a convenient distribution. Without loss of generality yet, $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A}) = q(\boldsymbol{w}|\mathbf{A})q_{\mathbf{A}}(\mathbf{A})$, where $q_{\mathbf{A}}(\mathbf{A}) \triangleq \int_{\boldsymbol{w}} q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A})\,d\boldsymbol{w}$ is a variational approximation to $p(\mathbf{A}|D)$. We propose to constrain the variational posterior $q_{\mathbf{A}}(\mathbf{A})$ to take the form of a Dirac distribution $q_{\mathbf{A}}(\mathbf{A}) \triangleq \delta_{\hat{\mathbf{A}}}(\mathbf{A})$ with all its mass assigned at the value $\mathbf{A}_k = \hat{\mathbf{A}}_k$ for $k = 1 \dots M$, so that $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A}) = q(\boldsymbol{w}|\hat{\mathbf{A}})\delta_{\hat{\mathbf{A}}}(\mathbf{A})$. Under this assumption, the optimum

$$q^*_{\boldsymbol{w},\mathbf{A}} = \underset{\hat{\mathbf{A}},\, q(\boldsymbol{w}|\hat{\mathbf{A}})}{\arg\max}\, \mathcal{L}(q) \triangleq q^*(\boldsymbol{w}|\mathbf{A}^*)\delta_{\mathbf{A}^*}(\mathbf{A}) \tag{18}$$

can be derived by calculus of variations in two steps. Given any $\hat{\mathbf{A}}$, $q^*(\boldsymbol{w}|\hat{\mathbf{A}})$ satisfies optimality conditions similar to Eq. (17):

$$\log q^*(\boldsymbol{w}|\hat{\mathbf{A}}) = \langle \log p(\boldsymbol{\theta}_{-\boldsymbol{w},-\hat{\mathbf{A}}}, \boldsymbol{w}, \hat{\mathbf{A}}, D)\rangle_{q_{-\boldsymbol{w},-\hat{\mathbf{A}}}} + \text{const}\,. \tag{19}$$

Reinjecting this expression into Eq. (18) turns it into a maximization w.r.t. $\hat{\mathbf{A}}$ only, so that the maximizer $\mathbf{A}^*$ of $\mathcal{L}(q)$ can be shown to maximize the following quantity:

$$\mathbf{A}^* = \underset{\hat{\mathbf{A}}}{\arg\max} \int_{\boldsymbol{w}} \exp\langle \log p(\boldsymbol{\theta}_{-\hat{\mathbf{A}}}, \hat{\mathbf{A}}, D)\rangle_{q_{-\hat{\mathbf{A}},-\boldsymbol{w}}}\, d\boldsymbol{w} \tag{20}$$

$$= \underset{\hat{\mathbf{A}}}{\arg\max}\; p(\hat{\mathbf{A}}\,|\,D, \langle\lambda\rangle_{q_\lambda}, \langle\boldsymbol{\beta}\rangle_{q_{\boldsymbol{\beta}}}, \langle z\rangle_{q_z})\,, \tag{21}$$

$$= \underset{\hat{\mathbf{A}}}{\arg\max}\; p(D\,|\,\hat{\mathbf{A}}, \langle\lambda\rangle_{q_\lambda}, \langle\boldsymbol{\beta}\rangle_{q_{\boldsymbol{\beta}}}, \langle z\rangle_{q_z})\,. \tag{22}$$

Eq. (21) uses the fact that $p(\boldsymbol{w}|\mathbf{A},\lambda)$ and $p(D|\boldsymbol{w},\boldsymbol{\beta},z)$ belong to exponential families. Eq. (22) follows from Bayes' rule with the improper prior $p(\mathbf{A}) \propto 1$, and shows that maximizing $\mathcal{L}(q)$ w.r.t. $\hat{\mathbf{A}}$ is the same as maximizing the conditional evidence $p(D\,|\,\hat{\mathbf{A}}, \langle\lambda\rangle_{q_\lambda}, \langle\boldsymbol{\beta}\rangle_{q_{\boldsymbol{\beta}}}, \langle z\rangle_{q_z})$. In fact to increase the value of the objective $\mathcal{L}(q)$ w.r.t. $q_{\boldsymbol{w},\mathbf{A}}$ we merely need to increase (not necessarily maximize) the conditional evidence w.r.t. $\hat{\mathbf{A}}$, then update $q(\boldsymbol{w}|\hat{\mathbf{A}})$ according to the optimality condition of Eq. (19).

Based on this remark, we derive an *active set* method that greedily improves on the objective functional $\mathcal{L}(q)$. The active set refers to the subset $\mathcal{S}$ of basis functions ϕ_k for which $\hat{\mathbf{A}}_k$ is finite along at least a direction, as opposed to the inactive set of basis functions for which $\hat{\mathbf{A}}_k$ is infinite and constrains $\boldsymbol{w}_k = 0$. The scheme starts with an arbitrary active set (typically $\mathcal{S} = \emptyset$) and proceeds by updating one $\hat{\mathbf{A}}_k$ at a time, maximizing the quantity of Eq. (22)) w.r.t. this basis function only. This results in adding a new basis to the active set if $\hat{\mathbf{A}}_k$ is made finite along at least a direction, or removing a previously active basis if $\hat{\mathbf{A}}_k$ becomes infinite. The hyperparameter $\hat{\mathbf{A}}_k$ that is selected for an update is the one, among all indices $k = 1 \dots M$, that provides the highest gain w.r.t. the objective. AppendixB shows that, in the case of a Gaussian likelihood, all necessary updates can be performed efficiently using rank-1 linear algebra identities.

3.3. *Related work*

In a simplified setting ($\lambda = 0$), Bishop and Tipping (2000) use the factorization $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A}) = q_{\boldsymbol{w}}(\boldsymbol{w})q_{\mathbf{A}}(\mathbf{A})$ to derive closed-form updates for all factors. Unfortunately, the smoothness-inducing prior destroys model conjugacies on which the authors rely. In addition, the resulting updates have a complexity $O(M^3)$ that does not scale favorably w.r.t the number of basis functions. Finally basis functions that are numerically pruned from the model cannot be reintroduced at a later stage. As an alternative, the evidence-maximization criterion of Eq. (21) was also proposed by Tipping et al. (2003) on the grounds of type-II maximum likelihood inference. Our active set method generalizes their fast marginal likelihood maximization procedure in presence of a smoothness-inducing prior.

3.4. *Gaussian approximation of the likelihood*

Although voxel intensities in the warped and fixed images are related by assumption *via* Gaussian noise (or a mixture thereof), the transformation $\Psi_{\boldsymbol{w}}^{-1}$ acts non-linearly on intensity profiles

and the resulting likelihood w.r.t. $\boldsymbol{w}$ does not belong to a standard family. To retrieve the required conjugacies during updates of $q_{\boldsymbol{\beta}}$, q_z and $q_{\boldsymbol{w},\mathbf{A}}$, a Gaussian approximation of the data likelihood is used. It is derived from an efficient second-order Taylor expansion of the log-likelihood (AppendixA). The Taylor expansion is local: it depends on the point $\boldsymbol{w}$ around which it is computed. For updates of $q_{\boldsymbol{\beta}}$ and q_z, the approximation is used around the known mode of $q(\boldsymbol{w}|\mathbf{A})$. For updates of $q_{\boldsymbol{w},\mathbf{A}}$ however, the approximation is taken at the mode of the true posterior $p(\boldsymbol{w}|D, \mathbf{A}, \langle\boldsymbol{\beta}\rangle_{q_{\boldsymbol{\beta}}}, \langle\boldsymbol{\pi}\rangle_{q_{\boldsymbol{\pi}}}, \langle\lambda\rangle_{q_\lambda})$, which must first be computed. This is done by quasi Newton optimization (L-BFGS) w.r.t. the subset of active variables (AppendixB).

3.5. Algorithm overview

The scheme proceeds according to Algorithm 1. We start with no active bases, $\mathcal{S} = \emptyset$. We cycle between updates of the noise mixture parameters, of the transformation parametrization and parameters, and of the regularization parameter. Prior to updating $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w}, \mathbf{A})$, we update the approximation of the data likelihood. The global objective $\mathcal{L}(q)$ provides an always increasing lower-bound to the evidence $p(D)$ and can be used to monitor convergence. Alternatively, the scheme can simply stop after a certain number of updates to the set of active bases has been performed.

Algorithm 1 Sparse Bayesian registration algorithm

Initialize $\hat{\mathbf{A}}_k = \infty$ for all k ($\mathcal{S} = \emptyset$) and q_λ
repeat
  for T iterations **do**
    Update q_z to $\arg\max_{q_z} \mathcal{L}(q)$ following AppendixD.
    Update $q_{\boldsymbol{\beta}}$ to $\arg\max_{q_{\boldsymbol{\beta}}} \mathcal{L}(q)$ following AppendixD.
    Update $q_{\boldsymbol{\pi}}$ to $\arg\max_{q_{\boldsymbol{\pi}}} \mathcal{L}(q)$ following AppendixD.
  end for
  Update the likelihood approximation (AppendixA).
  Update $\mathbf{A}$ (active set method) to greedily increase $\mathcal{L}(q)$ then set $q(\boldsymbol{w}|\mathbf{A})$ to $\arg\max_{q(\boldsymbol{w}|\mathbf{A})} \mathcal{L}(q)$ (AppendixB, AppendixE, AppendixF, AppendixG)
  Update q_λ to $\arg\max_{q_\lambda} \mathcal{L}(q)$ according to AppendixC.
until no significant increase in $\mathcal{L}(q)$ or maximum number of iterations reached.

3.6. Algorithmic complexity

Updates of the mixture parameters take $O(L \cdot N)$ per pass on the image. The cost is dominated by the computation, for each of the N voxels, of soft-assignments to the L mixture components. Several $O(L \cdot N)$ passes are typically performed. The regularization level λ is updated in $O(|\mathcal{S}|^2)$, where $|\mathcal{S}|$ is the number of active bases. Updates to the parametrization $\mathbf{A}$ occur one basis at a time (a single $\mathbf{A}_k$ is changed): each update takes $O(|\mathcal{S}|^2 + M|\mathcal{S}| + N \log N)$ to maintain necessary statistics, exploiting rank-1 linear algebra identities. As a byproduct, an update of $q(\boldsymbol{w}|\mathbf{A})$ is obtained. An overhead of $O(|\mathcal{S}|^3 + M|\mathcal{S}|^2 + |\mathcal{S}| \cdot N \log N)$ adds up to this, since statistics must be recomputed once in full after the mixture parameters and the likelihood approximation are updated. The likelihood approximation itself involves minimizing a registration energy w.r.t. the subset of active basis parameters ($|\mathcal{S}| \ll M$) to find the posterior mode, then a $O(N)$ cost to compute the Gaussian approximation around this mode.

As a point of comparison, a single gradient descent step when optimizing the classical registration energy of Eq. (1) w.r.t. the full set of variables costs $O(M^2 + N \log N)$, where the left-hand term stems from the gradient of the regularization energy and the right-hand term from the gradient of the data-energy. Exact Hessian computation in absence of sparsity costs $O(M^2 + MN \log N)$ and Hessian inversion is $O(M^3)$.

4. Experiments & Results

4.1. Material

We experiment with the proposed framework on tasks of cardiac motion tracking. The goal is to recover the motion of the cardiac muscle over the course of the cardiac cycle from a time series of 3D images. The first experiment gives insight into the empirical behaviour of the proposed algorithm on a simple example of 2D pairwise registration. Other experiments involve full $3D + t$ motion tracking on various imaging modalities.

The first dataset consists of synthetic sequences of 3D ultrasound data provided as part of the 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). These synthetic images count approximately 10 million voxels each, at a very fine isotropic resolution of 0.33mm. To avoid further optimization of our code in terms of RAM management, we downsampled them by a factor of 2. We thus worked at a resolution of 0.66mm at the finest level. The second and third datasets are hosted by the Cardiac Atlas Project. They were made available following the cardiac motion analysis challenge (Tobon-Gomez et al., 2013) organized for the 2011 MICCAI workshop on Statistical Atlases and Computational Models of the Heart (STACOM). The data includes a set of 15 sequences of real 3D tagged MR images at roughly 1mm × 1mm × 1mm resolution (1 million voxels), and a set of 15 sequences of real cine MR images at about 1.25mm × 1.25mm × 8mm resolution. The tagged sequences contain 20 to 30 frames each, the cine MR sequences 30 frames each. Fig. 8 displays example 2D slices from frames of each modality.

4.2. Details of the experimental setting

The experimental setup is identical across all modalities. The multiscale parametrization of the displacement field consists of isotropic Gaussian kernels of respective variance $S_1 = 20^2$ mm^2 and $S_2 = 10^2$ 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. The proposed framework imposes no restriction on the parametrization of the displacement field and we expect the anisotropy to be of potential relevance given the ventricle anatomy. All hyperparameters are set to uninformative values ($a_0 = b_0 = c_0 = d_0 = \eta_0 \to 0$). The registration is accelerated with a multiresolution pyramidal scheme, starting with

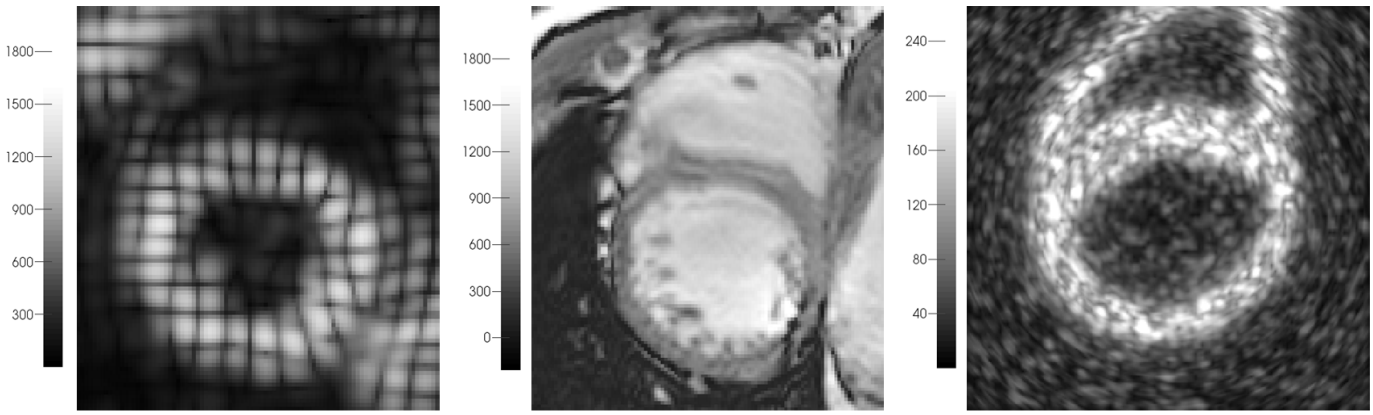


Figure 8: Example slices for the cardiac imaging modalities that we experiment on, with artefacts and patterns peculiar to each modality. (Left) 3D tagged MR image. (Middle) 3D echocardiographic image. (Right) 3D cine MR image.

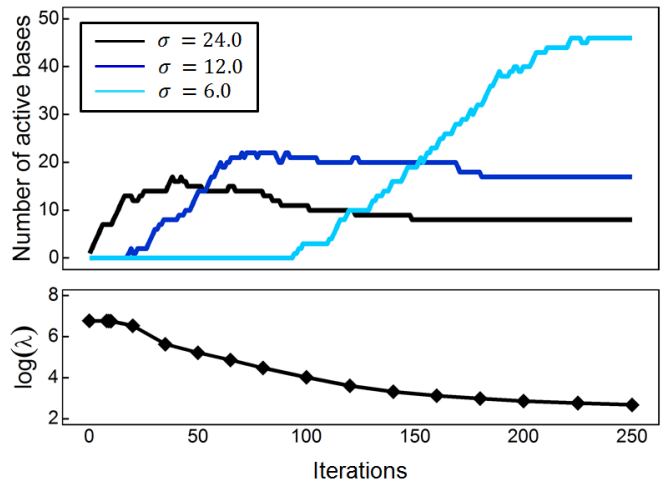


Figure 9: Basis selection mechanism and its coupling with the jointly estimated regularization level, across iterations. (Top) Addition, update or deletion of dictionary bases in the active parametrization of the displacement field across iterations. Three distinct scales are used in the representation of displacements (1 curve per scale). (Bottom) Regularization parameter λ, updated every few iterations, plotted against the number of iterations run since the beginning of the registration.

downsampled (smoothed) versions of images I and J and progressively moving through the pyramid of images to the images I and J at full resolution. Three resolution levels are introduced, downsampling by a factor of 2 at each level. Note that we do not make use of pre-segmentations of regions of interest. Computations were run on an Intel Xeon processor X5660 (@2.80GHz, 6 cores, 12 threads) and took 15 – 30min per image pair for cine MRI, 30 – 45min for tagged MRI and ~ 90min for 3D US.

4.3. Self-tuning registration algorithm: an analysis

We use the example 2D registration of Fig. 3 and Fig. 7 to give some insight into the registration algorithm.

Basis selection & regularization. Fig. 9 demonstrates how basis selection mechanisms empirically combine with parameter re-estimation throughout iterations. A heuristic provides a large initial value for the regularization level λ. This initially discourages the addition of finer dictionary bases, whose impact on the signal regularity is too high at this stage. Coarse bases are added instead to capture the global trends in the observed displacement. The regularization level is consequently refined to reflect the regularity of the inferred displacement. As λ decreases towards a more sensible value, finer bases are 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 be deemed no longer to contribute towards a better explanation of the data and removed. Towards the last iterations, most actions consist in updating the orientation of active bases rather than in additions or deletions from the active set, and λ reaches a plateau as well. Fig. 7 further illustrates this mechanism of basis selection. The location of active bases is shown 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

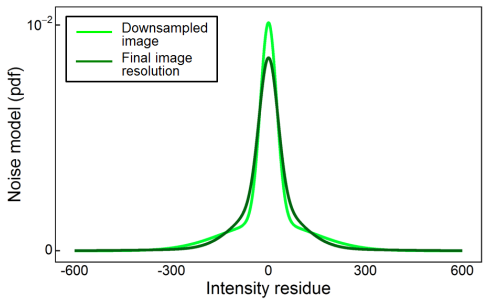


Figure 10: Inferred noise model at the beginning of registration and at the end. The noise model learned on downsampled, smoothed images has a higher probability of low noise (higher peak around 0) but also of high noise (higher tails) due to the increased misalignment at the beginning of registration.

over the course of the algorithm. In all experiments it displayed a fast convergence. The Gaussian mixture also adapts quickly to changes in the distribution of intensity residuals that arise from the multiresolution pyramidal scheme, when hopping from a smoothed downsampled image to the next level in the pyramid of images, as seen from Fig 10.

Robustness w.r.t. initialization. Fig. 11 provides evidence towards the empirical robustness of the estimated level of regularity λ w.r.t. its initial value. The initial value of λ spans 4 orders of magnitude, whereas its final estimate varies by at most a factor of 4 across runs. Empirically, we observe that the regularity level λ decreases monotonically from its starting value towards a reasonable local optimum. As a limitation, it follows that the scheme will typically not recover the expected regularity level if initialized from too low a value of λ.

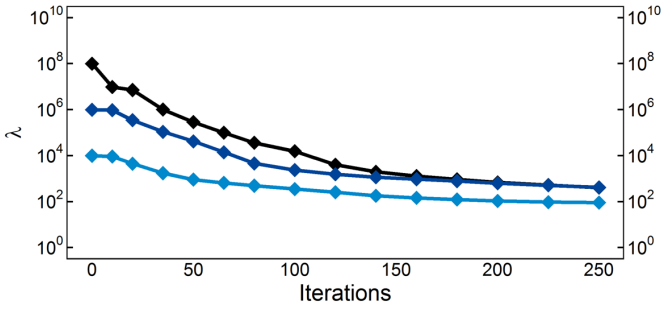
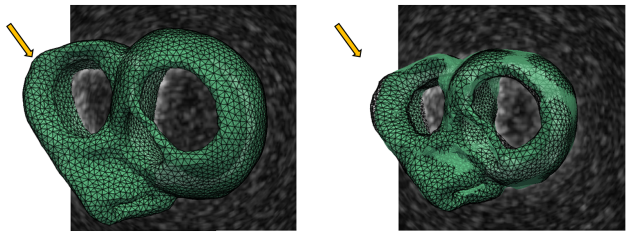


Figure 11: Robustness of the inferred regularity level w.r.t. its initial estimation. The 2D registration is run 3 times, and initialized each time with a differing level of regularity λ (respectively 10^4, 10^6, 10^8). Each curve shows the evolution of λ over the course of the associated run.

4.4. Synthetic 3D Ultrasound Cardiac Dataset

The appeal of this benchmark is to offer a dense ground truth in terms of motion and strain inside the cardiac muscle. The workflow of image synthesis uses the output of a cardiac electromechanical model to prescribe displacements in the myocardium. For each sequence of images, the ground truth consists of a sequence of meshes of the left and right ventricles



end-diastole frame end-systole frame

Figure 12: 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.

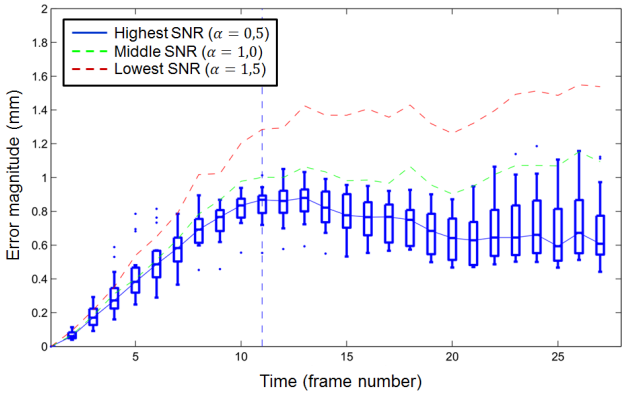


Figure 13: Accuracy benchmark on the 3D US STACOM 2012 normal dataset, reporting the median tracking error over time for varying SNRs (blue, green, red curves). For the reference SNR (in blue), quartiles are overlayed (boxplots) to picture the dispersion of error values.

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. The visual and qualitative behaviour of the proposed approach was found to be satisfactory, even in terms of extrapolation: the inferred motion remained consistent in areas of the right ventricle that fall outside of the field of view (Fig. 12). This hints at an effective regularization mechanism, despite being automatically tuned.

We evaluate the accuracy of the proposed 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. 13 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. As seen from Fig. 14, the increased SNR impacts the noise model (with a higher prevalence of small intensity residuals at high SNR) learned by the proposed approach, which in turn becomes more conservative in its estimates of displacements.

Fig. 15a 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

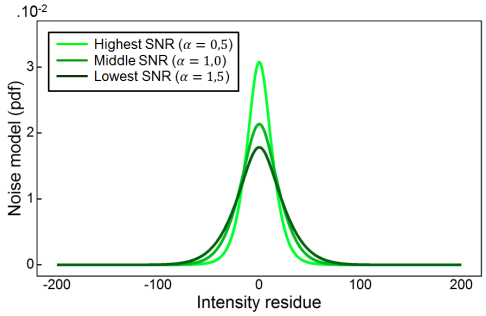


Figure 14: Evolution of the inferred noise model for increasing Signal-to-Noise Ratios.

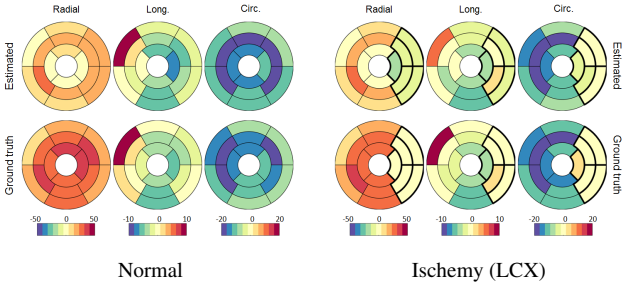


Figure 15: Bull's eye plots of the radial, longitudinal and circumferential strain components at end-systole, averaged over AHA segments: estimated (top) and ground truth (bottom). A healthy case (left) and an ischemic case (right) are reported.

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 underestimated in the left ventricle. This might indicate a slight under-estimation of the endo- and/or epicardium displacement, 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 and all sequences). 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. 15b 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.

4.5. STACOM 2011 tagged MRI benchmark

On this image modality the grid-like tags 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 divided in three groups of 4 points in the basal, mid-ventricular and apical areas of the left ventricle. They serve as ground truth from which registration accuracy is assessed. Details of the experimental setting along with challenger results are provided by Tobon-Gomez et al. (2013).

Fig. 16 summarizes challengers' results along with ours at End-Systole (ES). 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.

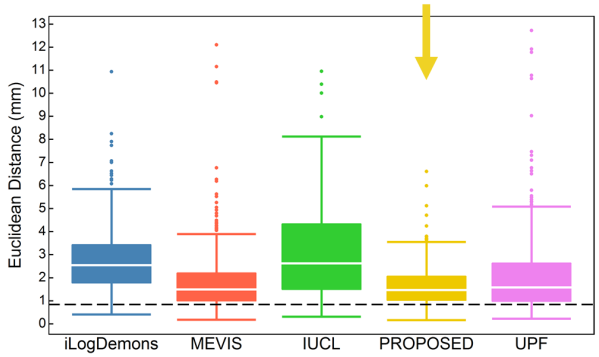


Figure 16: Accuracy benchmark on the 3D tag STACOM 2011 dataset, reporting box-plots of tracking errors on all methodologies. The doted black line represents the average inter-observer variability.

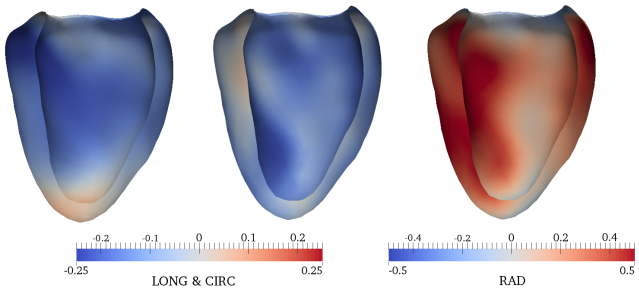


Figure 17: Strain at ES, computed from the 3D tag data of volunteer V9.

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.

The proposed formulation appears to achieve in a quasi automatic manner results qualitatively and quantitatively on par with the state of the art. In particular we insist that most 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. The strain maps and mesh deformations produced by the proposed scheme, as illustrated for instance in Fig. 17, 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. 17). This again hints at the practical viability of the automatically adjusted trade-off between data and regularization energies. 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 (pooled over all cases). Median, first and third quartiles are reported.

4.6. Cine MRI dataset: qualitative results and uncertainty

Original images had a low inter-slice resolution of 8mm compared to the in-plane resolution of 1.25mm. 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

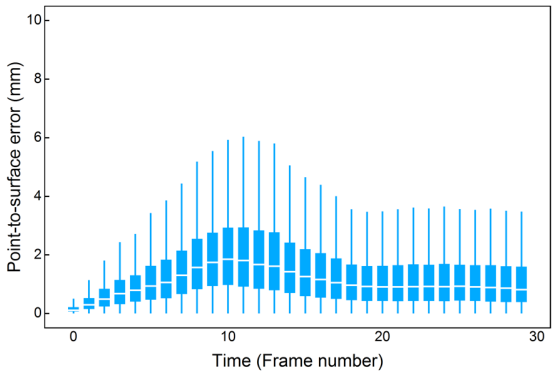


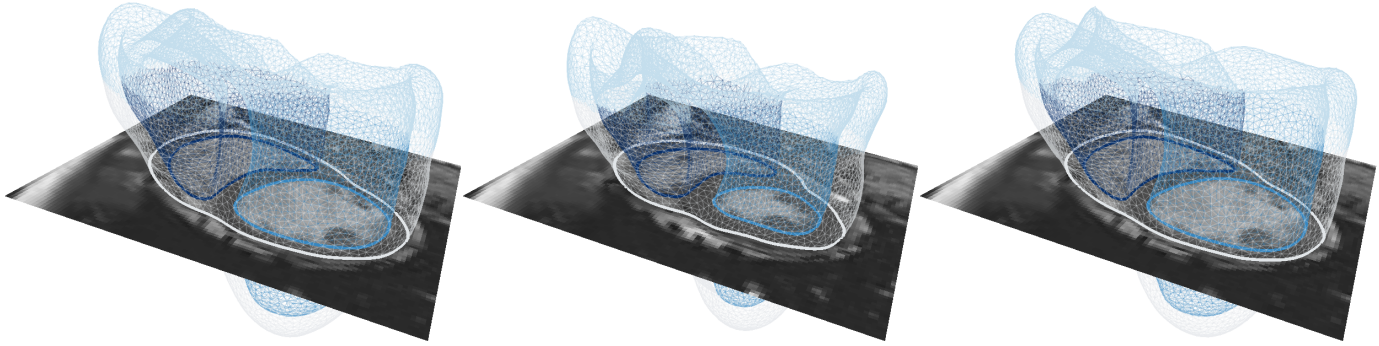
Figure 18: Accuracy benchmark on the cine MR dataset, reporting median error over time along with quartiles. Surfaces reconstructed from slice-by-slice 3D+t segmentation serve as ground truth. Points on the discrete contours at time 0 are transported over time using the registration output. For each time step, point-to-surface distances over all 15 sequences and all contour points are pooled. Errors at time 0 are induced by the surface reconstruction.

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. 18 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 captured variations in the level of noise of an order of magnitude between distinct components (a factor of 10 between the standard deviations of extreme components). The visual aspect of the cardiac muscle changes drastically over time, so that these regions tend to be assigned higher noise levels than the baseline acquisition noise

[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.

level. Voxels in basal slices, with visible outflow tracts and apparent topology changes, also tend to fall in the noisiest components. Fig. 20 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.

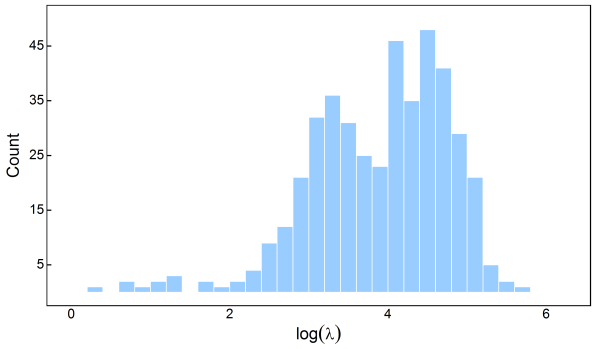


Figure 20: Histogram of inferred values for the regularity hyperparameter λ, pooled over all 15 sequences and 30 frames per sequence.

Finally, 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 (pooled over all cases), at each scale. Median, first and third quartiles are reported.

5. Discussion

5.1. Adequacy of modelling assumptions

Despite using generic RBFs, the inferred parametrizations of 3D displacements were highly sparse, typically involving no more than a hundred degrees of freedom. This shows that the proposed sparsity-inducing mechanism is potent, and benefits both algorithmic complexity and memory usage. Finer bases were used more often in experiments with tagged MRI; in addition to the higher resolution of these volumes compared to cine MRI data, tags may have been regarded as 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.

We did not explicitely make use of temporal regularization (as in *e.g.* De Craene et al. (2012)), but the temporal (and spatial) consistency of deformations remained satisfactory (Fig. 1(a) and Fig. 19). Incorporating temporal regularization and moving towards a large deformation framework (Beg et al., 2005; Arsigny et al., 2006) with geodesic-by-part trajectories may still be advantageous, although technicalities should be addressed to maintain a reasonable computational complexity.

The proposed smoothness-inducing prior is widely used, and realizes a pragmatic trade-off between quality and complexity: it is stationary and involves a single hyperparameter λ. Spatially varying levels of regularization have been used in the recent literature (Simpson et al., 2015; Gerig et al., 2014) to account for imaging artefacts or heterogeneous image content (*e.g.* tissue, blood), at the cost of additional technicalities and approximations in the variational inference. Here, spatially varying regularization occurs indirectly via a spatially varying noise level, and directly thanks to the adaptive parametrization of displacements. The latter results in a non-stationary prior on transformations (with coarser bases naturally enforcing higher regularization).

The sparsity-inducing prior over individual basis functions is parametrized by a positive symmetric precision matrix $\mathbf{A}_k$. This allows to activate a given basis along a single direction while constraining $\boldsymbol{w}_k$ to be null in the orthogonal plane. This type of (at most) rank one $\mathbf{A}_k^{-1} = \alpha_k^{-1}\boldsymbol{n}_k\boldsymbol{n}_k^\intercal$ is in fact optimal under variational Bayes inference. To simplify derivations, an alternative *all or nothing* parametrization $\mathbf{A}_k = \alpha_k\mathbf{I}$ could have been chosen, where a basis function is artificially constrained to be either fully in use or fully pruned along all directions, at the cost of artificially tripling the number of active degrees of freedom.

For mono-modal registration the assumption that intensities between source and target images coincide up to spatially varying noise mostly holds. In this context, the mixture-of-Gaussian model of residuals is flexible yet simple enough to be efficiently and robustly fit jointly during the registration of images of interest, as opposed to beforehand on training data (Leventon and Grimson, 1998; Zhou et al., 2006; Lee et al., 2009; Tang et al., 2012). For higher interpretability of the mixture, the variational Bayes approach allows to select the optimal number of components (Penny et al., 2007; Archambeau and Verleysen, 2007), although this was not pursued here. For multi-modal registration, the mapping between source and target image intensities can also be regressed (Guimond et al., 2001; Janoos et al., 2012).

The assumed independence of voxelwise intensity residuals is not realistic. To avoid placing too much confidence on data, the virtual decimation procedure downweights the data term (by up to two factors of magnitude in 3D experiments) in a mostly ad-hoc but empirically viable manner. Designing spatially varying, correlation-aware models of image discrepancies would address the matter more elegantly, but is outside of the scope of this paper.

Modelling uncertainty in the interpolation of discrete intensity profiles (cf. AppendixA) proved to be appropriate. This is examplified by cine MR data, due to the lower long-axis resolution. If not accounting for it the regularity of the inferred transform in the long axis direction was systematically found, upon visual inspection, to be of lesser quality. This behaviour is expected if the scheme is unaware of its increased reliance on interpolation to match intensity values between images. Image upsampling prior to registration also involves interpolation and was accounted for in an identical manner.

5.2. Implementation & computational load

Running times for the proposed approach are on the same order of magnitude as the state-of-the-art in cardiac motion tracking (De Craene et al., 2013; Tobon-Gomez et al., 2013). Although our implementation is CPU based (with partial multithreading), most computations involve statistics at the level of individual voxels or basis functions and are highly parallelizable. The active set method used to update the active parametrization is technical but computationally inexpensive. Significant improvements in running time may be obtained instead by basic optimization of the energy minimization problem solved to compute the posterior mode for the Laplace approximation[4], by partial parallelization of the VB updates for the noise mixture (AppendixD) and by optimizing convolution filters that we rely on (AppendixH). Limitations of the proposed approach include the technicality of its implementation and its memory consumption.

Upon inspection, the expressions derived for $\langle\lambda\rangle_{q_\lambda}$, $\langle\boldsymbol{\beta}\rangle_{q_\beta}$ and other hyperparameter expectations relate to intuitive quantities, such as averages of voxelwise square intensity residuals, or the energy in the estimated displacement field. Simplified versions of these updates can be used independently of the specifics of the registration scheme as ad-hoc recipes for automated parameter tuning.

5.3. Atlas of motion

In the present work, the reduced parametrization simply benefits the algorithmic complexity of the schemes and the quality of interpolation. If an anatomically relevant parametrization were desired however, we note that the active set method naturally extends to a multi-subject setting: it may be used to yield an optimal, joint parametrization of displacements. This could constitute a basis to learn 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).

5.4. Uncertainty quantification

The proposed method returns a Gaussian variational approximation $q(\boldsymbol{w}|\mathbf{A}) \triangleq \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the posterior distribution $p(\boldsymbol{w}|I, J)$ of transformation parameters. The $|\mathcal{S}| \times |\mathcal{S}|$ covariance matrix $\boldsymbol{\Sigma}$ is of small size and readily computed over the course of the algorithm. 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 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 allows to derive empirical uncertainty estimates on integral geometrical quantities.

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 pointwise displacement and is elongated in directions of higher uncertainty. Due to voxel anisotropy, 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.

[4]The minimization is w.r.t. the active, reduced parametrization and should be significantly faster and easier than its classic counterpart.

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. The agreement between exact and approximate posteriors should be explored in future work.

6. Conclusions

In this paper we proposed an approach to data-driven, spatially adaptative, multiscale parametrizations 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. It is self-tuning, with hyperparameters being jointly inferred during registration. Inference is tractable on real-scale data thanks to an efficient Variational Bayes method.

The core methodological contribution 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. The prior is allowed to encode correlations between explanatory variables so as to favor smooth solutions. The proposed structured sparse Bayesian model itself, and variants thereof, are relevant to a variety of generalized regression problems, including image-based classification and regression tasks.

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. Second order Taylor expansion of the model likelihood

Taking a Gaussian approximation of the likelihood facilitates the variational Bayes inference, as the variational posterior $q(\boldsymbol{w}|\mathbf{A}) \approx \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is then approximately Gaussian (AppendixB). We seek such an approximation of the likelihood – or equivalently we look for a quadratic approximation of the log-likelihood. In this appendix we derive one from the second-order Taylor expansion of $\mathcal{D}_{\beta,z}(J;I,\boldsymbol{w}) = -\log p(D|\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{z})$. Around the point $\boldsymbol{w}^*$, it takes the form:

$$\begin{aligned}\mathcal{D}_{\beta,z}(J;I,\boldsymbol{w}) \approx \mathcal{D}_{\beta,z}(J;I,\boldsymbol{w}^*) + \nabla[\mathcal{D}_{\beta,z}]_{\boldsymbol{w}^*}^\intercal(\boldsymbol{w}-\boldsymbol{w}^*) \\ + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^*)^\intercal \mathrm{H}[\mathcal{D}_{\beta,z}]_{\boldsymbol{w}^*}(\boldsymbol{w}-\boldsymbol{w}^*)\end{aligned} \tag{A.1}$$

where $\nabla[\mathcal{D}_{\beta,z}]_{\boldsymbol{w}^*}$ stands for the gradient at $\boldsymbol{w}^*$ and $\mathrm{H}[\mathcal{D}_{\beta,z}]_{\boldsymbol{w}^*}$ for the Hessian at $\boldsymbol{w}^*$. For the gradient we obtain:

$$\nabla[\mathcal{D}_{\beta,z}]_{\boldsymbol{w}^*} = \sum_{i=1}^{N} \hat{\beta}_i \left(I[V_i] - J[v_i]\right) \boldsymbol{\phi}(v_i)^T \nabla I(V_i)\,, \tag{A.2}$$

where $\hat{\beta}_i = \sum_{1\le l\le L} z_{il}\beta_l$. $\{v_i\}_{i=1}^N$ is the list of voxel centers in the fixed image and $V_i \triangleq \Psi_{\boldsymbol{w}^*}^{-1}(v_i) = v_i + \boldsymbol{\phi}(v_i)\boldsymbol{w}^*$ are the paired coordinates in the moving image. For the Hessian:

$$\begin{aligned}\mathrm{H}[\mathcal{D}_{\beta,z}]_{\boldsymbol{w}^*} = \sum_{i=1}^{N} \hat{\beta}_i\, \boldsymbol{\phi}(v_i)^\intercal \nabla I(V_i) \nabla I(V_i)^\intercal \boldsymbol{\phi}(v_i) \\ + \sum_{i=1}^{N} \hat{\beta}_i \left(I[V_i] - J[v_i]\right) \boldsymbol{\phi}(v_i)^\intercal \mathrm{H}[I](V_i) \boldsymbol{\phi}(v_i)\,.\end{aligned} \tag{A.3}$$

After dropping the term involving the Hessian of the image[5], we arrive at the following approximation of $\mathcal{D}_{\beta,z}(J;I,\boldsymbol{w})$:

$$\mathcal{D}_{\beta,z}(J;I,\boldsymbol{w}) \approx \frac{1}{2}\sum_{i=1}^{N} (\mathrm{t_i} - \boldsymbol{\phi}(v_i)\boldsymbol{w})^\intercal \hat{\beta}_i \mathbf{H}_i (\mathrm{t_i} - \boldsymbol{\phi}(v_i)\boldsymbol{w})\,. \tag{A.4}$$

$\mathrm{t_i} \in \mathbb{R}^d$ and $\mathbf{H}_i \in \mathcal{M}_{d\times d}$ only depend on the point $\boldsymbol{w}^*$ around which the approximation is taken, and are respectively given by Eq. (A.5) and Eq. (A.6), noting $\mathrm{u}^*(v_i) \triangleq \boldsymbol{\phi}(v_i)\boldsymbol{w}^*$.

$$\mathrm{t_i} = \mathrm{u}^*(v_i) - \frac{I(V_i) - J(v_i)}{\|\nabla I(V_i)\|^2}\nabla I(V_i)\,, \tag{A.5}$$

$$\mathbf{H}_i = \nabla I(V_i)\nabla I(V_i)^\intercal\,. \tag{A.6}$$

Eq. (A.4) demonstrates that up to a second order local approximation, registration can be recast as a regression task with a set of (virtual) observations $\mathrm{t_i}$ and associated (heteroscedastic, anisotropic) confidence $\mathbf{H}_i$. This is known as a generalized linear model. In block form, the data likelihood can be expressed as $p(D|\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{z}) \approx p(\boldsymbol{t}|\boldsymbol{w},\hat{\boldsymbol{\beta}})$ with:

$$p(\boldsymbol{t}|\boldsymbol{w},\hat{\boldsymbol{\beta}}) \propto \exp -\frac{1}{2}(\boldsymbol{t}-\Phi\boldsymbol{w})^\intercal \hat{\boldsymbol{\beta}}\mathbf{H}(\boldsymbol{t}-\Phi\boldsymbol{w})\,. \tag{A.7}$$

$\boldsymbol{t}$ denotes the concatenation of all $\mathrm{t_i}$. $\hat{\boldsymbol{\beta}}\mathbf{H}$ is a block diagonal matrix with the ith diagonal block equal to $(\hat{\boldsymbol{\beta}}\mathbf{H})_i \triangleq \hat{\beta}_i\mathbf{H}_i$.

In our particular instance, the virtual pairings relate to the optical flow: if we dropped the confidence tensors $\mathbf{H}_i$, Eq. (A.4) would yield an approximation of Eq. (B.6) 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. The local approximation of Eq. (A.4) transforms an image-based criterion into a landmark-based one, and the proximity to formulations in the related literature (Rohr et al., 2003) is indeed striking in this form.

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

$$\left((\hat{\beta}_i\mathbf{H}_i)^{-1} + \mathbf{D}_{\mathrm{int}}\right)^{-1} = \frac{1}{1+\mathrm{tr}[\hat{\beta}_i\mathbf{H}_i\mathbf{D}_{\mathrm{int}}]}\cdot\hat{\beta}_i\mathbf{H}_i \tag{A.8}$$

[5] This outer product approximation conveniently guarantees positivity of the Hessian and is discussed in *e.g.* (Bishop et al., 2006).

which implements a soft upper threshold on the precision, as a heuristic for interpolation uncertainty. $\mathbf{D}_{\text{int}}$ 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.

AppendixB. VB inference – optimization of $q_{\boldsymbol{w},\mathbf{A}}$

We recall that the variational posterior $q_{\boldsymbol{w},\mathbf{A}}$ is constrained to lie in the family $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A}) = q(\boldsymbol{w}|\hat{\mathbf{A}})\delta_{\hat{\mathbf{A}}}(\mathbf{A})$. AppendixB.1 restates exact optimality conditions and exhibits where the likelihood is involved. Under a Gaussian approximation of the likelihood, (approximate) optimality conditions yield tractable updates of $q(\boldsymbol{w}|\hat{\mathbf{A}})$ and suggest an efficient active set scheme to update $\hat{\mathbf{A}}$ (AppendixB.2). Technicalities for the Gaussian approximation are clarified in AppendixB.3.

AppendixB.1. Exact optimality conditions

From optimality conditions given in section 3.2 by Eq. (19) and keeping explicit only the terms that depend on $\boldsymbol{w}$, we obtain:

$$q^*(\boldsymbol{w}|\hat{\mathbf{A}}) = \frac{p\left(D|\boldsymbol{w},\langle\boldsymbol{\beta}\rangle_{q_\beta},\langle z\rangle_{q_z}\right)\cdot p\left(\boldsymbol{w}|\hat{\mathbf{A}},\langle\lambda\rangle_{q_\lambda}\right)}{p\left(D|\hat{\mathbf{A}},\langle\lambda\rangle_{q_\lambda},\langle\boldsymbol{\beta}\rangle_{q_\beta},\langle z\rangle_{q_z}\right)}, \tag{B.1}$$

where we exploited the linearity of the logarithm of distributions of interest w.r.t. $\boldsymbol{\beta}$, z and λ to take the expectations of those variables w.r.t. the associated variational factors inside those distributions. The numerator of Eq. (B.1) is simply the product of the likelihood times the prior on transformation parameters, with λ, $\boldsymbol{\beta}$ and z fixed at their expected values with respect to their respective variational posteriors. Again from Eq. (22), the optimal $\mathbf{A}^*$ maximizes the denominator of Eq. (B.1):

$$\begin{aligned} &p\left(D\,|\hat{\mathbf{A}},\langle\lambda\rangle_{q_\lambda},\langle\boldsymbol{\beta}\rangle_{q_\beta},\langle z\rangle_{q_z}\right) = \\ &\quad\int_{\boldsymbol{w}} p\left(D|\boldsymbol{w},\langle\boldsymbol{\beta}\rangle_{q_\beta},\langle z\rangle_{q_z}\right)\cdot p\left(\boldsymbol{w}|\hat{\mathbf{A}},\langle\lambda\rangle_{q_\lambda}\right)d\boldsymbol{w} \end{aligned} \tag{B.2}$$

and $q^*_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A}) = q^*(\boldsymbol{w}|\mathbf{A}^*)\delta_{\mathbf{A}^*}(\mathbf{A})$. Maximizing Eq. (B.2) w.r.t. $\hat{\mathbf{A}}$ and evaluating Eq. (B.1) however is hard since the likelihood $p(D|\boldsymbol{w},\langle\boldsymbol{\beta}\rangle_{q_\beta},\langle z\rangle_{q_z})$ is non trivial as a function of $\boldsymbol{w}$. For convenience we drop the expectations $\langle\ \rangle$ from notations in what follows (and the hat over $\mathbf{A}$), as it does not affect derivations.

AppendixB.2. Approximate optimality conditions

The tractability of the inference relies on approximating the data likelihood following AppendixA as a Gaussian distribution $p(D|\boldsymbol{w},\boldsymbol{\beta},z) \approx \mathcal{N}(\boldsymbol{t}|\Phi\boldsymbol{w},\hat{\boldsymbol{\beta}}\mathbf{H})$, with virtual data $\boldsymbol{t}$ interpretable as the concatenation of noisy voxelwise displacements. The prior $p(\boldsymbol{w}|\mathbf{A},\lambda) = \mathcal{N}(0,\mathbf{A}+\lambda\mathbf{R})$ is also Gaussian, yielding in turn a Gaussian variational posterior $q^*(\boldsymbol{w}|\mathbf{A}) \approx \mathcal{N}(\boldsymbol{\mu},\Sigma)$ with:

$$\boldsymbol{\mu} = \Sigma\Phi^\intercal\hat{\boldsymbol{\beta}}\mathbf{H}\boldsymbol{t} \quad \Sigma = (\Phi^\intercal\hat{\boldsymbol{\beta}}\mathbf{H}\Phi + \lambda\mathbf{R} + \mathbf{A})^{-1}. \tag{B.3}$$

Moreover the marginal likelihood $p(D|\mathbf{A},\lambda,\boldsymbol{\beta},z) \approx p(\boldsymbol{t}|\hat{\boldsymbol{\beta}},\mathbf{A},\lambda)$ is Gaussian as well:

$$p(\boldsymbol{t}|\hat{\boldsymbol{\beta}},\mathbf{A},\lambda) = |\,2\pi\,\mathbf{C}\,|^{-1/2}\cdot\exp\left\{-\frac{1}{2}\boldsymbol{t}^\intercal\mathbf{C}^{-1}\boldsymbol{t}\right\} \tag{B.4}$$

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

$$\mathbf{C} = (\hat{\boldsymbol{\beta}}\mathbf{H})^{-1} + \Phi(\mathbf{A}+\lambda\mathbf{R})^{-1}\Phi^\intercal. \tag{B.5}$$

The objective is to increase Eq. (B.4) w.r.t. $\mathbf{A}$. The two factors in Eq. (B.4) 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 mechanically leads to sparsity. Indeed looking at the form of $\mathbf{C}$ in Eq. (B.5), we see that part of the data is explained *for free* by the contribution $(\hat{\boldsymbol{\beta}}\mathbf{H})^{-1}$ of the noise to $\mathbf{C}$ regardless of the right hand term; thus only a few degrees of freedom need be active ($\mathbf{A}_k < \infty$) to fully explain the data.

Algorithm 2 Optimization of $\mathbf{A}$

1: **if** the likelihood approximation was updated **then**
2: Recompute $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ in full from Eq. (B.3).
3: Recompute statistics $\boldsymbol{q}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{\kappa}_k$ in full for all k.
4: **end if**
5: **for** p iterations **do**
6: $\forall k \in \mathcal{S}$ (resp. $k \notin \mathcal{S}$), compute the gain $\max_{\mathbf{A}_k}\Delta l(\mathbf{A}_k)$ in log-evidence obtained by updating or deleting k from $\mathcal{S}$ (resp. adding k to $\mathcal{S}$) from AppendixE.
7: Select the most favorable action l such that: $\max_{\mathbf{A}_l}\Delta l(\mathbf{A}_l) \geq \max_{\mathbf{A}_k}\Delta l(\mathbf{A}_k)$
8: Set $\mathbf{A}_l^* = \arg\max_{\mathbf{A}_l} l(\mathbf{A}_l)$ and update $\mathcal{S}$.
9: Update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ via rank-one identities (AppendixF).
10: Update $\boldsymbol{q}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{\kappa}_k$ for all k using rank-one updates (AppendixG).
11: **end for**

In AppendixE we single out the contribution $\Delta l(\mathbf{A}_k)$ of each basis ϕ_k to the log marginal likelihood given the state of other bases. AppendixF and AppendixG show that the statistics $\boldsymbol{q}_k$, $\boldsymbol{s}_k$ and $\boldsymbol{\kappa}_k$ required to express this contribution can be updated efficiently using rank-one linear algebra identities. This is the basis for the proposed *active set* method, which optimizes $\mathbf{A}$ by updating one $\mathbf{A}_k$ at a time according to Algorithm 2.

AppendixB.3. Around which point is the likelihood approximated?

We have not yet specified around which point the Gaussian approximation is taken. The mode $\boldsymbol{\mu}_{\text{MP}}$ of the (true) posterior

$$p(\boldsymbol{w}|D,\mathbf{A},\langle\boldsymbol{\beta}\rangle_{q_\beta},\langle\boldsymbol{\pi}\rangle_{q_\pi},\langle\lambda\rangle_{q_\lambda})$$

is chosen. This departs from Simpson et al. (2012), who use a Taylor expansion around the mode of the (Gaussian) variational posterior $q_{\boldsymbol{w}}$, and from Le Folgoc et al. (2014) who use a different strategy to derive a quadratic approximation. Here, as in energy-based registration we numerically solve for the minimizer $\boldsymbol{\mu}_{\text{MP}}$ of Eq. (B.6), using the current estimate of hyperparameters $\mathbf{A},\langle\lambda\rangle,\langle\boldsymbol{\beta}\rangle,\langle\boldsymbol{\pi}\rangle$:

$$\mathcal{E}(\boldsymbol{w}) = \mathcal{D}_{\langle\boldsymbol{\beta}\rangle,\langle\boldsymbol{\pi}\rangle}(J;I,\Phi\boldsymbol{w}) + \frac{1}{2}\boldsymbol{w}^\intercal(\mathbf{A}+\langle\lambda\rangle\mathbf{R})\boldsymbol{w} \tag{B.6}$$

with the notable difference, in terms of computational complexity, that Eq. (B.6) only involves the sparse subset $\mathcal{S}$ of active bases. Other weights are effectively constrained to zero due to an infinite penalty $\mathbf{A}_k$. Note that we marginalize over soft assignments $\boldsymbol{z}$ of voxels to components of the noise mixture. This comes at little computational cost and can reasonably be expected to accelerate convergence. The mode is found by quasi Newton (BFGS) optimization, using closed form expressions for the energy and its gradient. Expressions for the energy follow Eq. (B.6) and Eq. (8). The gradient of the data term is given by:

$$\nabla_{\boldsymbol{w}}[\mathcal{D}_{\langle\boldsymbol{\beta}\rangle,\langle\boldsymbol{\pi}\rangle}] = \sum_{i=1}^{N} \tilde{\beta}_i \left(I[V_i] - J[v_i]\right) \boldsymbol{\phi}(v_i)^T \nabla I(V_i)\,. \tag{B.7}$$

where $\tilde{\beta}_i = \sum_{1\leq l\leq L} \rho_{il}\langle\beta_l\rangle$ can be seen as an effective noise level for the ith voxel, and $\rho_{il} \propto \frac{\langle\pi_l\rangle}{Z_l} \exp -\frac{\langle\beta_l\rangle}{2} \left(J[v_i] - I[V_i]\right)^2$ (such that $\sum_l \rho_{il} = 1$) can be seen as a soft assignment of the ith voxel to the lth component.

AppendixC. VB inference – optimization of q_λ

Taking terms that do not depend on λ in the constant, the optimality condition of Eq. (17) rewrites as:

$$\begin{aligned}
\log q_\lambda(\lambda) &= \langle \log p(\boldsymbol{w}|\mathbf{A},\lambda)p(\lambda) \rangle_{q_{\boldsymbol{w},\mathbf{A}} q_{\boldsymbol{\pi}} q_{\boldsymbol{z}} q_{\boldsymbol{\beta}}} + \text{const} && \text{(C.1)}\\
&= \langle \log p(\boldsymbol{w}|\mathbf{A},\lambda)p(\lambda) \rangle_{q_{\boldsymbol{w},\mathbf{A}}} + \text{const} && \text{(C.2)}\\
&= \langle \log p(\boldsymbol{w}|\hat{\mathbf{A}},\lambda)\rangle_{q(\boldsymbol{w}|\hat{\mathbf{A}})} + \log p(\lambda) + \text{const.} && \text{(C.3)}
\end{aligned}$$

Recall that $\langle\cdot\rangle_{q(\theta)} = \int_\theta \cdot\, q(\theta) d\theta$ denotes expectation w.r.t. $q(\theta)$. The second equation uses the fact that the integrand does not depend on either $\boldsymbol{\pi}$, $\boldsymbol{z}$ nor $\boldsymbol{\beta}$. The third uses the fact that $q_{\boldsymbol{w},\mathbf{A}}(\boldsymbol{w},\mathbf{A}) = q(\boldsymbol{w}|\mathbf{A})\delta_{\hat{\mathbf{A}}}(\mathbf{A})$. The prior on λ is Gamma distributed, $p(\lambda) = \Gamma(\lambda|a_0,b_0)$, so that:

$$\log p(\lambda) = -b_0\lambda + (a_0 - 1)\log\lambda + \text{const}(\lambda)\,. \tag{C.4}$$

$p(\boldsymbol{w}|\hat{\mathbf{A}},\lambda) = \mathcal{N}(0|\hat{\mathbf{A}} + \lambda\mathbf{R})$ is a degenerate Gaussian. For all inactive bases $k \notin \mathcal{S}$, $\hat{\mathbf{A}}_k^{-1} = 0$, and the Gaussian degenerates to a Dirac at 0 along the d corresponding dimensions. For all active bases $k \in \mathcal{S}$, the Gaussian degenerates along $d-1$ directions since $\hat{\mathbf{A}}_k^{-1} = \alpha_k^{-1}\boldsymbol{n}_k\boldsymbol{n}_k^\intercal$ is rank one (section 2.4, AppendixE). By assumption $\alpha_k = 0$ so that, noting $\boldsymbol{w}_{\mathcal{S}} = (\boldsymbol{w}_k^\intercal \boldsymbol{n}_k)_{k\in\mathcal{S}}$ and $\boldsymbol{w}_{-\mathcal{S}} = \boldsymbol{w}\backslash\boldsymbol{w}_{\mathcal{S}}$:

$$p(\boldsymbol{w}|\hat{\mathbf{A}},\lambda) = \delta_0(\boldsymbol{w}_{-\mathcal{S}}) \cdot |\frac{\lambda}{2\pi}\mathbf{R}_{\mathcal{S}}|^{1/2} \exp -\frac{\lambda}{2}\boldsymbol{w}_{\mathcal{S}}^\intercal \mathbf{R}_{\mathcal{S}}^{-1}\boldsymbol{w}_{\mathcal{S}}\,, \tag{C.5}$$

where $\mathbf{R}_{\mathcal{S}}$ is an $|\mathcal{S}|\times|\mathcal{S}|$ matrix whose k,lth coefficient is $\boldsymbol{n}_k^\intercal\mathbf{R}\boldsymbol{n}_l$. Finally, since $q(\boldsymbol{w}|\hat{\mathbf{A}}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$:

$$\begin{aligned}
\log q_\lambda(\lambda) = \frac{|\mathcal{S}|}{2}\log\lambda - \frac{1}{2}\Big(\boldsymbol{\mu}^\intercal\mathbf{R}_{\mathcal{S}}\boldsymbol{\mu} + \text{tr}(\boldsymbol{\Sigma}\mathbf{R}_{\mathcal{S}})\Big)\lambda \\
- b_0\lambda + (a_0 - 1)\log\lambda + \text{const}(\lambda)\,.
\end{aligned} \tag{C.6}$$

In other words, $q_\lambda(\lambda) = \Gamma(\lambda|a,b)$ is Gamma distributed with hyperparameters given by Eq. (C.7) and (C.8), and $\langle\lambda\rangle_{q_\lambda} = a/b$.

$$a = a_0 + |\mathcal{S}|/2 \tag{C.7}$$

$$b = b_0 + \Big(\boldsymbol{\mu}^\intercal\mathbf{R}_{\mathcal{S}}\boldsymbol{\mu} + \text{tr}(\boldsymbol{\Sigma}\mathbf{R}_{\mathcal{S}})\Big)/2\,. \tag{C.8}$$

AppendixD. VB inference – optimization of q_π, q_β, q_z

VB updates are stated without proof. Derivations follow the same strategy as previously, starting from the optimality condition of Eq. (17). They follow the same outline as those of Archambeau and Verleysen (2007). The variational posterior for assignments $\boldsymbol{z}$ of voxels to a component of the Gaussian mixture is a product of categorical distributions $q_{\boldsymbol{z}} = \prod_{1\leq i\leq N} C(z_i|\rho_{i1}\cdots\rho_{iL})$ with parameters ρ_{il} interpretable as soft-assignments, summing to 1 ($\forall i$, $\sum_l \rho_{il} = 1$) and given by Eq. (D.1). The variational posterior $q_{\boldsymbol{\pi}}$ for the noise mixture proportions $\boldsymbol{\pi} = \{\pi_1\cdots\pi_L\}$ follows a Dirichlet distribution $\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\eta})$, with $\boldsymbol{\eta} = (\eta_1\cdots\eta_L)$ given by Eq. (D.2). The variational posterior for noise levels $q_{\boldsymbol{\beta}} = \prod_{1\leq l\leq L}\Gamma(\beta_l|c_l,d_l)$ is a product of Gamma distributions over each individual mixture component, with parameter updates given by Eq. (D.3) and (D.4).

$$\rho_{il} \propto \tilde{\pi}_l\tilde{\beta}_l^{1/2} \exp -\tfrac{1}{2}\langle\beta_l\rangle_{q_\beta}\langle e_i^2\rangle \tag{D.1}$$

$$\eta_l = N\bar{\pi}_l + \eta_0 \tag{D.2}$$

$$c_l = \tfrac{1}{2}N\bar{\pi}_l + c_0 \tag{D.3}$$

$$d_l = \tfrac{1}{2}\textstyle\sum_{1\leq i\leq N}\rho_{il}\,\langle e_i^2\rangle + d_0 \tag{D.4}$$

The quantities required for these updates are as follows:

$$\log\tilde{\pi}_l \triangleq \langle\log\pi_l\rangle_{q_\pi} = \psi(\eta_l) - \psi(\textstyle\sum_{1\leq l\leq L}\eta_l) \tag{D.5}$$

$$\log\tilde{\beta}_l \triangleq \langle\log\beta_l\rangle_{q_\beta} = \psi(c_l) - \log d_l \tag{D.6}$$

$$\bar{\pi}_l \triangleq \tfrac{1}{N}\textstyle\sum_{1\leq i\leq N}\rho_{il} \tag{D.7}$$

$$\begin{aligned}
\langle e_i^2\rangle &\triangleq \langle (I[\Psi_{\boldsymbol{w}}^{-1}(v_i)] - J[v_i])^2\rangle_{q(\boldsymbol{w}|\hat{\mathbf{A}})}\\
&\approx \left(I[\Psi_{\boldsymbol{\mu}}^{-1}(v_i)] - J[v_i]\right)^2 + \text{tr}[\boldsymbol{\phi}(v_i)^\intercal\boldsymbol{\Sigma}\boldsymbol{\phi}(v_i)\,\mathbf{H}_i]
\end{aligned} \tag{D.8}$$

$$\langle\beta_l\rangle_{q_\beta} = c_l/d_l \tag{D.9}$$

where $\psi(\cdot)$ stands for the digamma function and $\text{tr}[\cdot]$ for the trace. Eq. (D.8) follows from the quadratic approximation of AppendixA, noting in addition that $q(\boldsymbol{w}|\hat{\mathbf{A}}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ (cf. AppendixB and Eq. (B.3) for the updates of $\boldsymbol{\mu},\boldsymbol{\Sigma}$).

Each factor $q_{\boldsymbol{z}}$, $q_{\boldsymbol{\pi}}$, $q_{\boldsymbol{\beta}}$ is updated in turn for T iterations; each iteration essentially computes then aggregates voxelwise statistics in a single pass over the image of squared intensity residuals $\langle e_i^2\rangle$.

AppendixE. Contribution of a basis to the log marginal likelihood

From Eq. (B.4) we see that the log marginal likelihood $\mathcal{L} = \log p(\boldsymbol{t}|\hat{\boldsymbol{\beta}},\mathbf{A},\lambda)$ is given up to additive constant by

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

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

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

Noting that $\mathbf{C}$ exclusively depends on the basis k via the kth ($d\times d$ block-) diagonal coefficient of $\mathbf{A}$ and the kth column (of $d\times d$ elements) of Φ, we would like to single out the contribution

$l(\mathbf{A}_k)$ of any such basis ϕ_k to the log marginal likelihood in the form:

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

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[6] from $\mathbf{L}^{-1} = \mathbf{A} + \lambda\mathbf{R}$ (or equivalently by setting $\mathbf{A}_k = +\infty$[7] in $\mathbf{L}$), we see from the Woodbury rank one matrix identity[6] 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:

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

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:

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

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 E.6 and E.7 for the two terms in the right-hand side of the log marginal likelihood expression E.1:

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

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

We introduced the statistics $\boldsymbol{s}_k \in \mathcal{M}_{d,d}$ and $\boldsymbol{q}_k \in \mathbb{R}^d$ respectively defined as:

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

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

Eq. (G.1) and Eq. (G.2) in AppendixG provide alternative, more easily interpretable expressions for these quantities. This yields the following expression for $l(\mathbf{A}_k)$:

$$\begin{aligned} l(\mathbf{A}_k) = \log|\mathbf{A}_k + \boldsymbol{\kappa}_k| \; - \; \log|\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k| \\ + \; \boldsymbol{q}_k^\intercal \left\{\mathbf{A}_k + \boldsymbol{\kappa}_k + \boldsymbol{s}_k\right\}^{-1} \boldsymbol{q}_k \,. \end{aligned} \tag{E.10}$$

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. $\mathbf{A}_k = +\infty$). 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. (E.10) under constraint that $\mathbf{A}_k$ is a symmetric positive semidefinite $d \times d$ matrix involves the gradient of the (unconstrained) function $l(\mathbf{A}_k)$:

$$\nabla l(\mathbf{A}_k) = -\boldsymbol{\sigma}_k \left\{\boldsymbol{q}_k \boldsymbol{q}_k^\intercal - \boldsymbol{s}_k - \boldsymbol{s}_k (\mathbf{A}_k + \boldsymbol{\kappa}_k)^{-1} \boldsymbol{s}_k\right\} \boldsymbol{\sigma}_k \tag{E.11}$$

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 $\mathbf{A}_k^{-1} = \alpha_k^{-1} \boldsymbol{n}_k \boldsymbol{n}_k^\intercal$. 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. (E.10) as a function of $\alpha, \boldsymbol{n}$ leads to maximizing E.12 under constraint that α is positive (dropping the index k for convenience). Note also that E.12 is invariant under reparametrization $\boldsymbol{n} \to \nu \boldsymbol{n}, \alpha \to \alpha/\nu^2$.

$$l(\alpha, \boldsymbol{n}) = -\log\left\{1 + \frac{\boldsymbol{n}^\intercal \boldsymbol{s} \boldsymbol{n}}{\alpha + \boldsymbol{n}^\intercal \boldsymbol{\kappa} \boldsymbol{n}}\right\} + \frac{(\boldsymbol{q}^\intercal \boldsymbol{n})^2}{\alpha + \boldsymbol{n}^\intercal (\boldsymbol{\kappa} + \boldsymbol{s}) \boldsymbol{n}} \tag{E.12}$$

At a maximizer $\alpha^*, \boldsymbol{n}^* = \arg\max_{\alpha, \boldsymbol{n}} l(\alpha, \boldsymbol{n})$ the constraint is either active ($\alpha^* = 0$) or inactive ($\alpha^* > 0$). If inactive, the solution actually maximizes the unconstrained function E.12 and is given by $\alpha^* = \bar{\alpha}(\bar{\boldsymbol{n}})$, $\boldsymbol{n}^* = \bar{\boldsymbol{n}}$ where

$$\bar{\alpha}(\boldsymbol{n}) = \frac{(\boldsymbol{n}^\intercal \boldsymbol{s} \boldsymbol{n})^2}{(\boldsymbol{q}^\intercal \boldsymbol{n})^2 - \boldsymbol{n}^\intercal \boldsymbol{s} \boldsymbol{n}} - \boldsymbol{n}^\intercal \boldsymbol{\kappa} \boldsymbol{n} \,, \tag{E.13}$$

$$\bar{\boldsymbol{n}} = \boldsymbol{s}^{-1} \boldsymbol{q} \,. \tag{E.14}$$

In this case $l(\alpha^*, \boldsymbol{n}^*)$ is simply equal to $\bar{l}(\boldsymbol{n}^*)$, where $\bar{l}(\boldsymbol{n})$ is defined by E.15 with $\xi(\boldsymbol{n}) \triangleq (\boldsymbol{q}^\intercal \boldsymbol{n})^2 / \boldsymbol{n}^\intercal \boldsymbol{s} \boldsymbol{n}$.

$$\bar{l}(\boldsymbol{n}) \triangleq -\log \xi(\boldsymbol{n}) + \xi(\boldsymbol{n}) - 1 \tag{E.15}$$

In addition $\bar{l}(\bar{\boldsymbol{n}})$ can be shown to always provide an upper bound to the maximum contribution of a basis to the evidence, $\max_{\alpha, \boldsymbol{n}} l(\alpha, \boldsymbol{n})$. If $\bar{\alpha}(\bar{\boldsymbol{n}}) < 0$, the upper bound is not reachable and the constraint is active, $\alpha^* = 0$. In that case, we numerically optimize over the unit sphere in $\mathbb{R}^d$ to find $\boldsymbol{n}^*$. The case occurs when the l_2-norm regularization is by itself sufficient along the direction $\boldsymbol{n}^*$, and no additional shrinkage is deemed necessary. It is also the default case if we purposedly restrict α to be either null or infinite. To save on unnecessary computations, we first check that the upper bound $\bar{l}(\bar{\boldsymbol{n}})$ to the maximum contribution of the basis k to the evidence is superior to the current best contribution among bases already handled.

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

Updates of the moments of the posterior distribution $\boldsymbol{\mu}, \boldsymbol{\Sigma} = (\Phi^\intercal (\hat{\boldsymbol{\beta}} \mathbf{H}) \Phi + \mathbf{A} + \lambda \mathbf{R})^{-1}$ and of $\mathbf{L} = (\mathbf{A} + \lambda \mathbf{R})^{-1}$ upon deletion from the model, update or addition to the model of a basis l 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[8]:

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \Sigma_l \Sigma_{ll}^{-1} \Sigma_l^\intercal \,, \tag{F.1}$$

[6] This is an abuse of speech for the sake of convenience, when we in fact manipulate blocks of $d \times d$ elements, lines and columns of such $d \times d$ elements.

[7] We abuse notations here again for convenience. We mean to say that $\mathbf{A}_k^{-1} = 0$ or that $\forall \boldsymbol{n} \in \mathbb{R}^d$, $\boldsymbol{n}^\intercal \mathbf{A}_k \boldsymbol{n} \to \infty$.

[8] Whenever necessary we identify square matrices (resp. vectors) of differing dimensionalities if one can be obtained from the other by padding with zeros, *e.g.* in (F.1) (F.2) (F.3), the rank-one correction leaves the lth column and line (resp. coefficient) of the right-hand side of the equation equal to zero.

$$\tilde{\mathbf{L}} = \mathbf{L} - \mathrm{L}_l \mathrm{L}_{ll}^{-1} \mathrm{L}_l^\intercal \tag{F.2}$$

and

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \Sigma_l (\Sigma_{ll}^{-1} \boldsymbol{\mu}_l)\,. \tag{F.3}$$

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:

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \tilde{\Sigma}_l \tilde{\Sigma}_{ll}^{-1} \tilde{\Sigma}_l^\intercal\,, \tag{F.4}$$

$$\tilde{\mathbf{L}} = \mathbf{L} + \tilde{\mathrm{L}}_l \tilde{\mathrm{L}}_{ll}^{-1} \tilde{\mathrm{L}}_l^\intercal\,, \tag{F.5}$$

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

$$\tilde{\Sigma}_l = \begin{pmatrix} \boldsymbol{\Sigma} \Pi_l \tilde{\Sigma}_{ll} \\ \tilde{\Sigma}_{ll} \end{pmatrix}, \quad \tilde{\mathrm{L}}_l = \begin{pmatrix} \mathbf{L}(\lambda \mathrm{R}_l) \tilde{L}_{ll} \\ \tilde{L}_{ll} \end{pmatrix} \tag{F.6}$$

and

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

Π_l is the column vector of $d \times d$ matrices defined by $\Pi_l = \Phi^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\phi_l + \lambda \mathbf{R}_l$. The counterpart of Eq. (F.3) for addition is given by Eq. (F.8):

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} + \tilde{\Sigma}_l \boldsymbol{q}_l\,. \tag{F.8}$$

In particular, $\tilde{\boldsymbol{\mu}}_l = \tilde{\Sigma}_{ll} \boldsymbol{q}_l$. The case of the update of a basis l is treated as a deletion followed by an addition, updating $\boldsymbol{s}_l$ and $\boldsymbol{\kappa}_l$ in-between these actions as they are needed in Eq. (F.7).

AppendixG. Update of $\boldsymbol{s}_k$, $\boldsymbol{\kappa}_k$ and $\boldsymbol{q}_k$

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

$$\boldsymbol{s}_k = \phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\phi_k - \Pi_k^\intercal \boldsymbol{\Sigma}_{-k} \Pi_k + (\lambda \mathrm{R}_k)^\intercal \mathbf{L}_{-k} (\lambda \mathrm{R}_k) \tag{G.1}$$

$$\boldsymbol{q}_k = \phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\boldsymbol{t} - \Pi_k^\intercal \boldsymbol{\Sigma}_{-k} \Phi^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\boldsymbol{t} \tag{G.2}$$

where Π_k is a column vector of $d \times d$ matrices defined by $\Pi_k = \Phi^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\phi_k + \lambda \mathbf{R}_k$. Π_k can be interpreted as the inner product of basis k with all the active bases w.r.t an appropriate metric, in the sense that its jth $d \times d$ coefficient is given by: $\Pi_{jk} = \phi_j^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\phi_k + \lambda \langle D\phi_j | D\phi_k \rangle$. $\boldsymbol{q}_k$ is the projection on the basis ϕ_k of the optimal residual $\boldsymbol{t} - \Phi\boldsymbol{\mu}_{-k}$ for the active set $\mathcal{S}\backslash\{k\}$, with a correction to account for regularization[9]: it is an unnormalized indicator of the relevance of basis ϕ_k to better explain the data. $\boldsymbol{s}_k + \boldsymbol{\kappa}_k = \phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\phi_k + \lambda \mathbf{R}_{kk} - \Pi_k^\intercal \boldsymbol{\Sigma}_{-k} \Pi_k$ measures how much confidence basis ϕ_k aggregates, taking into account its overlap with bases already in the active set (the confidence already captured by overlapping active bases is withdrawn). 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 $\boldsymbol{t}_k$ and $\boldsymbol{r}_k$ respectively defined according to G.3 and G.4:

$$\boldsymbol{t}_k \triangleq \phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\phi_k - \Pi_k^\intercal \boldsymbol{\Sigma} \Pi_k + (\lambda \mathrm{R}_k)^\intercal \mathbf{L} (\lambda \mathrm{R}_k) \tag{G.3}$$

$$\boldsymbol{r}_k \triangleq \phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\boldsymbol{t} - \Pi_k^\intercal \boldsymbol{\Sigma} \Phi^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\boldsymbol{t} \tag{G.4}$$

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

$$\begin{aligned} \boldsymbol{s}_k = \boldsymbol{t}_k &+ \left[\Pi_k^\intercal \Sigma_k\right] \Sigma_{kk}^{-1} \left[\Pi_k^\intercal \Sigma_k\right]^\intercal \\ &- \left[(\lambda \mathrm{R}_k)^\intercal \mathrm{L}_k\right] L_{kk}^{-1} \left[(\lambda \mathrm{R}_k)^\intercal \mathrm{L}_k\right]^\intercal \end{aligned} \tag{G.5}$$

$$\boldsymbol{q}_k = \boldsymbol{r}_k + \left[\Pi_k^\intercal \Sigma_k\right] \Sigma_{kk}^{-1} \mu_k \tag{G.6}$$

Thus, we always maintain the quantities $\boldsymbol{t}_k$ and $\boldsymbol{r}_k$ (for every basis) and recompute $\boldsymbol{s}_k$ and $\boldsymbol{q}_k$ either at no cost for inactive bases or, for bases in the active set $\mathcal{S}$, in $O(|\mathcal{S}| \cdot d)$. Updates of $\boldsymbol{t}_k$ and $\boldsymbol{r}_k$ upon deletion from the model, update or addition to the model of a basis l 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

$$\begin{aligned} \tilde{\boldsymbol{t}}_k = \boldsymbol{t}_k &- \left[\Pi_k^\intercal \tilde{\Sigma}_l\right] \tilde{\Sigma}_{ll}^{-1} \left[\Pi_k^\intercal \tilde{\Sigma}_l\right]^\intercal \\ &+ \left[(\lambda \mathrm{R}_k)^\intercal \tilde{\mathrm{L}}_l\right] \tilde{L}_{ll}^{-1} \left[(\lambda \mathrm{R}_k)^\intercal \tilde{\mathrm{L}}_l\right]^\intercal \end{aligned} \tag{G.7}$$

and

$$\tilde{\boldsymbol{r}}_k = \boldsymbol{r}_k - \left[\Pi_k^\intercal \tilde{\Sigma}_l\right] \boldsymbol{r}_l \tag{G.8}$$

where $\tilde{\boldsymbol{r}}_k$ and $\tilde{\boldsymbol{t}}_k$ denote updated quantities, as opposed to quantities prior to the update $\boldsymbol{r}_k$ and $\boldsymbol{t}_k$. The quantities indexed by l are computed (once for all bases) following AppendixF. Similarly we maintain $\lambda \mathbf{R}_{kk} - (\lambda \mathrm{R}_k)^\intercal \mathbf{L} (\lambda \mathrm{R}_k)$ instead of $\boldsymbol{\kappa}_k$ (index $-k$ dropped from $\mathbf{L}_{-k}$).

AppendixH. Basis functions: Computational complexity with translation invariant kernels

The section discusses the generality of the proposed approach w.r.t. the basis functions ϕ_k parametrizing the displacement field. Smooth radial basis functions as well as many types of splines can be used without affecting the computational complexity. We show how to efficiently compute elements $\mathbf{R}_{kl}$ of the quadratic regularizer and inner products $\phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\Phi$, which are required in the scheme (see for instance AppendixF and AppendixG).

Regularization and Fourier analysis. Let $K(x, y) = \kappa(x - y)$ be a translation invariant positive definite kernel such that κ is integrable on $\mathbb{R}^d$ as well as its Fourier transform $\hat{\kappa}$. We consider

[9] Indeed, $\boldsymbol{q}_k = \phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})(\boldsymbol{t} - \Phi\boldsymbol{\mu}_{-k}) - \lambda \mathbf{R}_k^\intercal \boldsymbol{\mu}_{-k}$.

the space $\mathcal{H}$ of integrable d-vector fields $f : \mathbb{R}^d \to \mathbb{R}^d$ such that $\|f\|^2 = \langle f|f\rangle < +\infty$, where

$$\langle f|g\rangle = \frac{1}{(2\pi)^d} \int_\xi \widehat{f(\xi)}^\dagger \widehat{g}(\xi)\, \hat{\kappa}^{-1}(\xi) d\xi\,. \tag{H.1}$$

$\hat{f}(\xi) \triangleq \mathcal{F}[f](\xi) \triangleq \int_x \exp\{-\iota x^\intercal \xi\} f(x) dx$ is the Fourier transform of f. If κ is sufficiently smooth, the successive partial derivatives $\partial_{x^{i_1}\cdots x^{i_p}} f$ of elements $f \in \mathcal{H}$ exist and all lie in $\mathcal{H}$ (Zhou, 2008). As such, we can consider families of regularizers of the form $\mathcal{R}_D(f) = \|Df\|^2$, where D is a linear differential operator. κ enables additional filtering in the frequency domain and can be set to $\hat{\kappa} = 1$ if no such penalty is desired.

Representation on a finite dictionary. The displacement field $u(x) = \sum_{1\le k\le M} \phi_k(x) \boldsymbol{w}_k$ is assumed to be expressed as a linear combination of bases $\phi_k \in \mathcal{H}$ (with associated weight $\boldsymbol{w}_k \in \mathbb{R}^d$). By linearity of the representation, $\mathcal{R}_D(f) = \boldsymbol{w}^\intercal \mathbf{R} \boldsymbol{w}$ where the k,lth block element $\mathbf{R}_{kl}$ is a $d\times d$ matrix, whose i,jth coefficient is $\langle D(\phi_k \boldsymbol{e}_i)\,|\,D(\phi_l \boldsymbol{e}_j)\rangle$, with $\boldsymbol{e}_1 \cdots \boldsymbol{e}_d$ the canonical frame. We wish to compute these coefficients efficiently.

Computation of $\mathbf{R}_{kl}$. If the basis functions are regularly translated versions of a single kernel ϕ, *i.e.* $\phi_k(x) = \phi(x - x_k)$, then we have from the properties of the Fourier transform and Eq. (H.1):

$$\begin{aligned}\mathbf{R}_{kl} &= \langle D(\phi_k \boldsymbol{e}_i)\,|\,D(\phi_l \boldsymbol{e}_j)\rangle \\ &= \frac{1}{(2\pi)^d} \int_\xi e^{\iota \delta x^\intercal \xi}\, \widehat{D(\phi \boldsymbol{e}_i)}(\xi)^\dagger \widehat{D(\phi \boldsymbol{e}_j)}(\xi)\, \hat{\kappa}^{-1}(\xi) d\xi \\ &= \mathcal{F}^{-1}[\widehat{D(\phi \boldsymbol{e}_i)}(\xi)^\dagger \widehat{D(\phi \boldsymbol{e}_j)}(\xi) \hat{\kappa}^{-1}(\xi)](\delta x)\,.\end{aligned} \tag{H.2}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform and $\delta x = x_k - x_l$. Values for all offsets can be computed at once using the fast Fourier transform. If the basis offset is the voxel size, this runs in $O(N \log N)$. The values can be stored in an image, so that retrieving a coefficient $\mathbf{R}_{kl}$ at a later stage is done in $O(1)$, after recomputing the corresponding offset. If the dictionary of basis functions is generated from a few (say, K) such translation invariant kernels, the same process is repeated K^2 times, once for each couple of generating kernels.

Case of the Gaussian kernel. For Gaussian basis functions and standard differential operators D, closed form expressions can be obtained instead of relying on the fast Fourier transform. We illustrate this on the bending energy ($D = \Delta$). Recall that for any p-uplet of integers $i_1 \cdots i_p \in \{1, \cdots, d\}$, $\mathcal{F}[\partial_{x^{i_1}\cdots x^{i_p}} f](\xi) = \iota^p \xi^{i_1} \cdots \xi^{i_p} \mathcal{F}[f](\xi)$. Thus for any $f \in \mathcal{H}$,

$$\mathcal{F}[\Delta f](\xi) = -\|\xi\|^2 \mathcal{F}[f](\xi)\,, \tag{H.3}$$

$$\mathcal{F}[\Delta^2 f](\xi) = \|\xi\|^4 \mathcal{F}[f](\xi)\,. \tag{H.4}$$

The Fourier transform of a Gaussian basis

$$\phi_k(x) = \exp -\frac{1}{2}(x - x_k)^\intercal S_k^{-1} (x - x_k) \tag{H.5}$$

is given by $\mathcal{F}[\phi_k](\xi) = |2\pi S_k|^{1/2} e^{-\iota \xi^\intercal x_k} \exp\{-\frac{1}{2}\xi^\intercal S_k \xi\}$, which is again Gaussian. Using Eq. (H.3), regrouping the exponential factors and using Eq. (H.4), we finally obtain for two Gaussian bases ϕ_k and ϕ_l (with respective variance S_k and S_l and centered at x_k and x_l):

$$\mathbf{R}_{k,l} = \left(\frac{|S_k| \cdot |S_l|}{|S_{k,l}|/(2\pi)^d}\right)^{1/2} (-\Delta)^2 [K_{S_{k,l}}](\delta x)\, \mathbf{I} \tag{H.6}$$

with $\mathbf{I}$ the $d \times d$ identity matrix, $\delta x = x_l - x_k$, $S_{k,l} = S_k + S_l - S$ and $K_{S_{k,l}}(x) = \exp -\frac{1}{2} x^\intercal S_{k,l}^{-1} x$. The derivative $(-\Delta)^2[K_{S_{k,l}}]$ of the Gaussian kernel at any point is known in closed form.

Inner products $\phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\Phi$. For a block-diagonal matrix $\hat{\boldsymbol{\beta}}\mathbf{H}$ (with N blocks) and an arbitrary set of M basis functions, computing the inner product of ϕ_k with all other bases functions is $O(MN)$. If the dictionary of bases is generated from K translation invariant kernels, these inner products can be computed in $O(KN \log N)$. $\phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})$ is computed in $O(N)$, then K convolutions (one with each generating kernel, in $O(N \log N)$ each) yield the desired result. Similarly all of the $\phi_k^\intercal(\hat{\boldsymbol{\beta}}\mathbf{H})\phi_k$, $k = 1 \ldots M$, can be computed by convolution of the image whose voxels are the $d \times d$ tensors $\hat{\beta}_i \mathbf{H}_i$ with the square of the K generating kernels.

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.

Argyriou, A., Foygel, R., Srebro, N., 2012. Sparse prediction with the k-support norm, in: Advances in Neural Information Processing Systems, pp. 1457–1465.

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.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012. Optimization with sparsity-inducing penalties. Foundations and Trends® in Machine Learning 4, 1–106.

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.

Belilovsky, E., Argyriou, A., Varoquaux, G., Blaschko, M.B., 2015a. Convex relaxations of penalties for sparse correlated variables with bounded total variation . Machine Learning , 1–21.

Belilovsky, E., Gkirtzou, K., Misyrlis, M., Konova, A., Honorio, J., Alia-Klein, N., Goldstein, R., Samaras, D., Blaschko, M., 2015b. Predictive sparse modeling of fMRI data for improved classification, regression, and visualization using the k-support norm. Computerized Medical Imaging and Graphics , 1.

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.

Ganz, M., Sabuncu, M.R., Van Leemput, K., 2013. An improved optimization method for the relevance voxel machine, in: Machine Learning in Medical Imaging. Springer, pp. 147–154.

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

Gerig, T., Shahim, K., Reyes, M., Vetter, T., Lüthi, M., 2014. Spatially varying registration using gaussian processes, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 413–420.

Glocker, B., Paragios, N., Komodakis, N., Tziritas, G., Navab, N., 2008. Optical flow estimation with uncertainties through dynamic mrfs, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

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.

Guimond, A., Roche, A., Ayache, N., Meunier, J., 2001. Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections. Medical Imaging, IEEE Transactions on 20, 58–69.

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.

Heinrich, M.P., Simpson, I.J., Papież, B.W., Brady, M., Schnabel, J.A., 2016. Deformable image registration by combining uncertainty estimates from supervoxel belief propagation. Medical image analysis 27, 57–71.

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

Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., Thirion, B., 2012. Multi-scale Mining of fMRI data with Hierarchical Structured Sparsity. SIAM Journal on Imaging Sciences 5, 835–856.

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.

Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N.D., Scholkopf, B., 2009. Learning similarity measure for multi-modal 3d image registration, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 186–193.

Leventon, M.E., Grimson, W.E.L., 1998. Multi-modal volume registration using joint intensity distributions, in: Medical Image Computing and Computer-Assisted InterventationMICCAI98. Springer, pp. 1057–1066.

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 .

Parisot, S., Wells, W., Chemouny, S., Duffau, H., Paragios, N., 2014. Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs. Medical image analysis 18, 647–659.

Penny, W., Kilner, J., Blankenburg, F., 2007. Robust bayesian general linear models. Neuroimage 36, 661–671.

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.

Roberts, S.J., Penny, W.D., 2002. Variational bayes for generalized autoregressive models. Signal Processing, IEEE Transactions on 50, 2245–2257.

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.

Rohr, K., Fornefett, M., Stiehl, H., 2003. Spline-based elastic image registration: integration of landmark errors and orientation attributes. Computer Vision and Image Understanding 90, 153 – 168.

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., 2012. The relevance voxel machine (rvoxm): A self-tuning bayesian model for informative image-based prediction. Medical Imaging, IEEE Transactions on 31, 2290–2306.

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., Cardoso, M., Modat, M., Cash, D., Woolrich, M., Andersson, J., Schnabel, J., Ourselin, S., Initiative, A.D.N., et al., 2015. Probabilistic non-linear registration with spatially adaptive regularisation. Medical image analysis .

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.

Stewart, C.V., Tsai, C.L., Roysam, B., 2003. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. Medical Imaging, IEEE Transactions on 22, 1379–1394.

Tang, L., Hero, A., Hamarneh, G., 2012. Locally-adaptive similarity metric for deformable medical image registration, in: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on, IEEE. pp. 728–731.

Tang, L.Y., Hamarneh, G., 2013. Random walks with efficient search and contextually adapted image similarity for deformable registration, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 43–50.

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

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 .

Wells III, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal 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 de-

termination, 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.

Zhou, S.K., Georgescu, B., Comaniciu, D., Shao, J., 2006. Boostmotion: Boosting a discriminative similarity function for motion estimation, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE. pp. 1761–1768.