Medical Image Analysis 26 (2015) 203–216

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Probabilistic non-linear registration with spatially adaptive regularisation☆

I.J.A. Simpson[a,b,*], M.J. Cardoso[a,b], M. Modat[a,b], D.M. Cash[b], M.W. Woolrich[d,e], J.L.R. Andersson[e], J.A. Schnabel[c], S. Ourselin[a,b], for the Alzheimer's Disease Neuroimaging Initiative

[a] Centre for Medical Image Computing, University College London, United Kingdom
[b] Dementia Research Centre, University College London, United Kingdom
[c] Institute of Biomedical Engineering, University of Oxford, United Kingdom
[d] Oxford Centre for Human Brain Activity, University of Oxford, United Kingdom
[e] Centre for Functional Magnetic Resonance Imaging of the Brain, University of Oxford, United Kingdom

ARTICLE INFO

Article history:
Received 19 October 2014
Revised 9 August 2015
Accepted 20 August 2015
Available online 28 September 2015

Keywords:
Medical image registration
Regularisation
Bayesian inference
Registration uncertainty

ABSTRACT

This paper introduces a novel method for inferring spatially varying regularisation in non-linear registration. This is achieved through full Bayesian inference on a probabilistic registration model, where the prior on the transformation parameters is parameterised as a weighted mixture of spatially localised components. Such an approach has the advantage of allowing the registration to be more flexibly driven by the data than a traditional globally defined regularisation penalty, such as bending energy. The proposed method adaptively determines the influence of the prior in a local region. The strength of the prior may be reduced in areas where the data better support deformations, or can enforce a stronger constraint in less informative areas. Consequently, the use of such a spatially adaptive prior may reduce unwanted impacts of regularisation on the inferred transformation. This is especially important for applications where the deformation field itself is of interest, such as tensor based morphometry. The proposed approach is demonstrated using synthetic images, and with application to tensor based morphometry analysis of subjects with Alzheimer's disease and healthy controls. The results indicate that using the proposed spatially adaptive prior leads to sparser deformations, which provide better localisation of regional volume change. Additionally, the proposed regularisation model leads to more data driven and localised maps of registration uncertainty. This paper also demonstrates for the first time the use of Bayesian model comparison for selecting different types of regularisation.

1. Introduction

Non-linear image registration is a fundamental tool in medical image analysis with a great many applications (Sotiras et al., 2013). One widely explored application of non-linear registration is the analysis of human brain morphology from structural magnetic resonance (MR) images. In this context, non-linear image registration has been used to accurately quantify localised cross-sectional differences between populations, such as subjects with Alzheimer's disease (AD) compared to normal ageing. It has also been used to measure longitudinal changes within individuals. Differences in morphology between populations can be identified using approaches such as tensor based morphometry (TBM) (Ashburner and Friston, 2000; Chung et al., 2001), where statistical analysis is performed on the Jacobian tensor of deformation fields calculated from registering individual subjects to a common space. TBM offers a whole brain approach to statistical analysis, and has the potential to extract rich features that accurately summarise anatomical differences.

TBM features are wholly defined by the registration process, which is complicated by the fact that non-linear registration is an ill-posed problem. In a typical structural MR image there are more than one million voxels in the human brain, where the intensity of a voxel is a noisy surrogate of tissue type. As such, there is a great deal of ambiguity in matching intensities, making it implausible for a unique voxelwise mapping to be determined purely from the image data.

☆ Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

* Corresponding author at: Centre for Medical Image Computing, University College London, United Kingdom. Tel.: +44 (0) 203 549 5530.
E-mail address: ivor.simpson@gmail.com (I.J.A. Simpson).

http://dx.doi.org/10.1016/j.media.2015.08.006

1.1. Regularisation

As no unique mapping can be determined purely from the data, a "reasonable" mapping between images is sought. This is achieved through the use of a data matching term and regularisation, which maximises the similarity of image appearance whilst maintaining a plausible deformation, i.e. with an appropriate magnitude of displacement and spatial smoothness. Regularisation can be considered as a prior on the set of expected deformations, which reduces the space of potential solutions and hence limits the variance of any estimated solution. The form of bias induced by the prior is generally selected based on some physical model of deformation, such as linear elasticity (Miller et al., 1993) or thin-plate spline bending energy (Bookstein, 1997).

Regularisation models are commonly described as having the same effect across the image. However, such models may well be unreasonable in brain registration for two reasons: Firstly, different regions of the image contain different amounts of information. Uninformative image areas should be strongly influenced by the priors as they contain little information, whereas feature-rich regions should be given more freedom. Furthermore, the magnitude of anatomical mis-correspondence is likely to be variable across space, and some regions will require more complex deformations than others to allow an adequate mapping. Therefore, the use of a global spatial regularisation prior may introduce either an unwanted or insufficient bias on the deformation in certain image regions. This could have substantial adverse effects on an application, such as TBM, which directly relies on the interpretability of the deformation field.

1.1.1. Previous approaches to spatially varying regularisation in registration

There have been several previous works on the use of spatially varying regularisation in non-linear registration. These include approaches that vary based on tissues or structures derived from segmentations (Lester et al., 1999; Davatzikos, 1997; Staring et al., 2007; Schmah et al., 2013). These approaches are ideal in cases when an informative deformation prior is known for a specific region or tissue type, which can be robustly defined. However, in the majority of registration applications, this is unlikely to be the case.

More data driven approaches have been proposed, which include anisotropic smoothing of image similarity gradients according to image information (Hermosillo et al., 2002; Papież et al., 2013). Alternative approaches include weighting similarity gradients based on measures of local image reliability (Tang et al., 2010). These approaches allow the image information to affect the local regularisation strength, although are still somewhat ad-hoc, being dependent on the definition of a heuristic weighting between regularisation and data fidelity.

Inference of geometric deviation from an estimated atlas for use as a spatial prior is an alternative approach to define regularisation priors, Allassonniére et al. (2007) proposed a small deformation Bayesian framework for atlas estimation and registration. Gori et al. (2013) proposed a Bayesian approach for estimating an atlas and structure specific regularisation terms for a registration model based on the metric of currents. A recently published approach by Xu et al. (2014) propose a method for deriving an average atlas and a spatial distance metric based on the geometric variability of the atlas. Zhang et al. (2013) proposed a generative registration model using Geodesic shooting for atlas and regularisation estimation, this work was extended to sparsely estimate the principal geodesic modes of variation (Zhang and Fletcher, 2014). Durrleman et al. (2013) also estimate sparse parametrisations of variability from an estimated atlas.

Most similarly to this work, Risholm et al. (2010b, 2013) presented a Bayesian inference scheme that allows linear elastic parameters to be inferred from the data. These parameters can also vary spatially, as demonstrated by Risholm et al. (2011b). This approach does not require the definition of strong heuristics, although informative priors are required for the elastic model parameters. The limitations of the framework lie in the numerical integration inference strategy, which comes with vast computational complexity. Modern sampling techniques may help alleviate this burden (Zhang et al., 2013).

1.2. Contribution of this paper

This paper proposes a novel non-linear registration model and Bayesian inference scheme that allows for data-driven spatially varying regularisation. This approach alleviates the difficulties associated with previous attempts at spatially varying regularisation. Firstly, it is fully data driven, requiring no segmentations or informative priors. Secondly, the trade-off of data fidelity and regularisation is inferred directly from the data and finally, inference is tractable.

This work follows from our previous conference paper (Simpson et al., 2013b), with a second-order inference scheme for the regularisation parameters, a full mathematical derivation and broader validation. Additionally, this paper investigates objective Bayesian model comparison and the effects of the spatially varying prior on registration uncertainty. The proposed framework describes registration using a hierarchical probabilistic model, with a transformation prior that is parameterised by a set of hyper-parameters. Each hyper-parameter influences a spatially localised region of the prior. Through the use of full Bayesian inference, posterior distributions of hyper-parameter weights can be inferred alongside the transformation. This allows the effects of the prior to be locally determined during the registration.

This approach is demonstrated through an application of TBM on synthetic images, as well as comparing subjects with AD to healthy controls. Our results demonstrate the strength of our approach in terms of reducing false positive results, which may improve interpretability. We also highlight additional benefits of the proposed framework including: objective comparison of regularisation models, and more reasonable uncertainty estimates of the deformation fields.

2. Method

2.1. Model

Image registration can be described in a probabilistic manner using a generative model of the target image, $\mathbf{y}$, which is predicted by the deformed source image, $\mathbf{t}(\mathbf{x}, \mathbf{w})$. Here, $\mathbf{t}$ is a transformation model, $\mathbf{x}$ is the source image and $\mathbf{w}$ parametrises the transformation. In this paper, a cubic B-spline free form deformation model (Rueckert et al., 1999; Andersson et al., 2007) is used for $\mathbf{t}$, with $\mathbf{w}$ corresponding to the control point displacement. However, in principle any deformation model could be used.

The generative model also contains an additive noise term, $\mathbf{e}$, which describes the error in model fit. In this work, $\mathbf{e}$, is modelled as independently and identically distributed across voxels and follows a normal distribution:

$$\mathbf{e} \approx \mathcal{N}(0, \mathbf{I}\phi^{-1}\alpha), \tag{1}$$

where $\mathbf{I}$ is an identity matrix the size of the number of voxels, N_v. ϕ corresponds to the noise precision (inverse variance) of the additive Gaussian noise under the assumption of being independently distributed. α corresponds to the virtual decimation factor (Groves et al., 2011), which is a data driven term used to compensate for spatial covariance in the residual, weakening the assumption of independent noise. The assumption of identically distributed noise could also be relaxed in this approach as in Simpson et al. (2012a). The full generative model for registration is therefore given as:

$$\mathbf{y} = \mathbf{t}(\mathbf{x}, \mathbf{w}) + \mathbf{e}. \tag{2}$$

2.2. Prior distributions

Prior information is used to constrain the parameters of the model to plausible values. The noise in model fit, ϕ, is well defined by the data, so an uninformative Gamma distribution prior can be used, $P(\phi) = \mathrm{Ga}(a_0, b_0)$, where $a_0 = 10e^{10}$, $b_0 = 10e^{-10}$. As motivated in Section 1.1, for the problem of non-linear registration an informative prior on the transformation parameters, $p(\mathbf{w})$, is required to ensure a reasonable result.

2.2.1. Priors on transformation parameters

Spatial regularisation for non-linear registration can be encoded as a prior on the transformation parameters. Commonly such priors penalise deviation from the identity transformation, functioning as an elastic type of regularisation. Here, the prior on $\mathbf{w}$ is described using a multivariate normal distribution:

$$p(\mathbf{w}) = \mathcal{N}(0, \boldsymbol{\Sigma}). \tag{3}$$

The mean of the prior is set to 0, representing the identity transformation. $\boldsymbol{\Sigma}$ describes the expected variance, and covariance of the transformation parameters. This definition allows the specification of highly complex and rich priors. Most commonly, bending or linear elastic energy priors have been encoded in such a form (Ashburner and Friston, 1999). Simpler constraints such as penalising the magnitude of the deformation parameters could also be straightforwardly included.

2.2.2. Multiple sparse priors

In this work, the multiple sparse priors (MSP) approach of Friston et al. (2008) is adopted to allow spatially varying regularisation for non-linear registration. The MSP model was previously demonstrated for use in the M/EEG inverse problem. Friston et al. define the prior covariance matrix to be a weighted mixture of n covariance components: $\boldsymbol{\Sigma} = \sum_i^n \exp(\lambda_i)\Sigma_i$, where each Σ_i has a pre-defined form, which is chosen to have limited spatial support, making Σ_i a spatially localised covariance component. The number and form of these components is optional. λ_i is a scalar weight associated with each covariance component that is inferred from the data. λ_i appears within an exponential to ensure a positivity constraint on the weighting factor for each Σ_i.

As in Friston et al. the prior covariance components, Σ_i, are constructed from columns of a spatial coherence prior, G. Here, G is a squared exponential Gaussian process (GP) prior (Rasmussen and Williams, 2006), which can equivalently be considered as the Green's function of a discrete diffusion process (Harrison et al., 2007). The graph encoding the distance between nodes is an adjacency matrix, A, where $A_{ij} = 1$ when transformation parameters $\mathbf{w}_i$ and $\mathbf{w}_j$ are spatially adjacent, and 0 elsewhere. G can be written as:

$$G(\sigma) = \exp(\sigma A) \approx \sum_{m=0}^{m=4} \frac{\sigma^m}{m!} A^m. \tag{4}$$

The parameter σ controls the local coherence between adjacent control points, and takes values between 0 (independence of parameters) and 1 (maximally correlated). This approximation to the Green's function only accounts for 4th order neighbouring control points, as defined by the maximum value of m, which allows sparse priors, with compact spatial support. For non-linear registration, the consideration of 4th order covariance neighbours provides an adequate balance between connectedness and sparsity. For a given prior component: $\Sigma_i = q_i q_i^T$, where q_i corresponds to the ith column in $G(\sigma)$.

Each prior component, Σ_i, strongly controls the variance of a control point displacement, in a given direction, and the covariance with neighbouring control points, with a weaker influence on these neighbours' variance. The scale of this component is dictated by the exponential of its control parameter, λ_i, which is inferred from the data. Fig. 1 illustrates the stages used to create $\boldsymbol{\Sigma}$.

In the present model, there is a univariate normal prior distribution placed on each $\lambda_i \in \{\lambda\}$ where $\{\lambda\} = \{\lambda_1, \lambda_2, \ldots, \lambda_{N_c}\}$ and N_c is the number of transformation parameters. The prior on λ_i is written as:

$$P(\lambda_i) = \mathcal{N}(\eta, \rho^2). \tag{5}$$

Due to the exponential parametrisation of λ_i, this effectively functions as a log-normal hyperprior on the weights of each Σ_i (Friston et al., 2007). The selection of $P(\lambda)$ is discussed in Section 2.5, and the rationale for choosing a normal prior, as opposed to a Gamma distribution, which was used as a prior on a single regularisation parameter, is discussed in Section 2.3.1.

2.3. Model inference

The generative model and priors defined in the previous sections describe a hierarchical probabilistic model that is described graphically in Fig. 2. Bayesian inference is used to infer the unobserved random variables in this hierarchical model. Numerical integration approaches, such as Markov chain Monte Carlo, are often computationally prohibitive in problems with many parameters. For this reason, mean-field variational Bayes (VB) (Attias, 2000) was chosen as the inference strategy. VB allows tractable, approximate full Bayesian

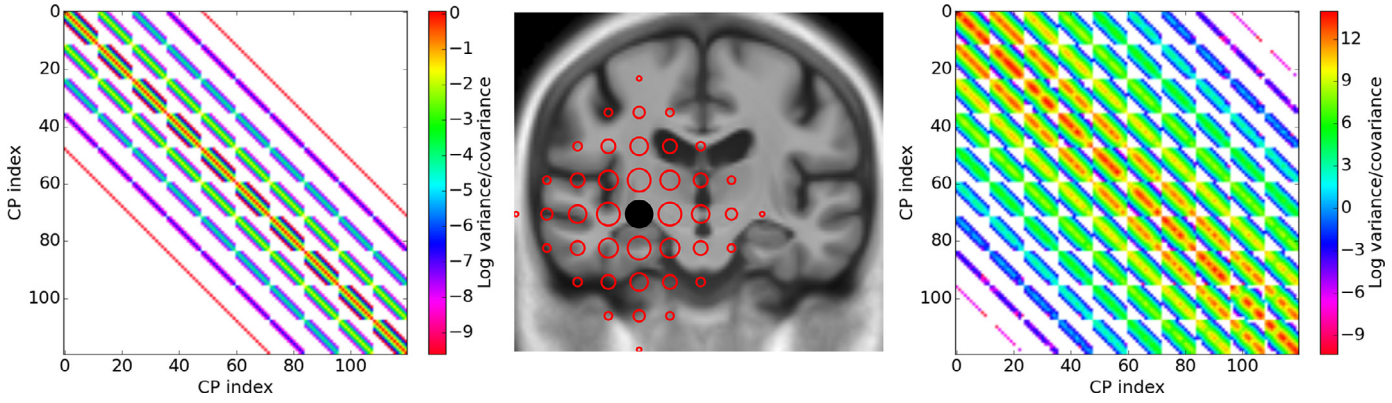

Fig. 1. An illustration of how $\boldsymbol{\Sigma}$ is created. The leftmost plot shows the GP covariance matrix $G(\sigma)$ as calculated from Eq. (4) on a 12 by 10 control point grid. The middle plot illustrates the basis function Σ_i associated with the ith column of $G(\sigma)$, where the black circle indicates the primarily affected control point and the relative size of the red circles illustrates the magnitude of the covariances of the nearby control points. The rightmost plot illustrates how a randomly weighted combination of spatially localised covariance components leads to the complete spatially varying prior covariance matrix, $\boldsymbol{\Sigma}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

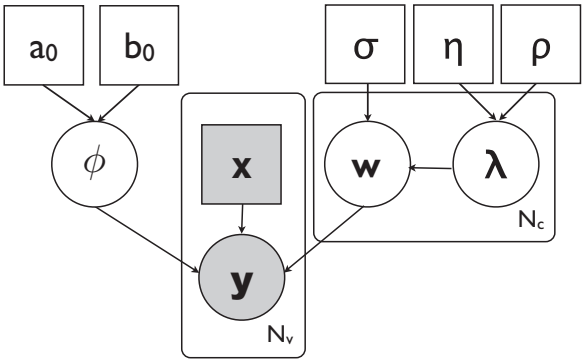

Fig. 2. A graphical description of the probabilistic registration model where the directions of the arrows describe the probabilistic dependencies. Symbols in circles are random variables, those in squares have fixed values. Grey containers are observations. Plates correspond to the dimensionality of the variable.

inference, and has been previously demonstrated for use in high resolution non-linear registration (Simpson et al., 2012b).

VB approximates the posterior distribution of model parameters using parametric distributions. In this work, mean-field VB is used, hence the posterior distribution on the model parameters is approximated as:

$$p(\mathbf{w}, \phi, \lambda|\mathbf{y}) \approx q(\mathbf{w}, \phi, \{\lambda\}) \approx q(\mathbf{w})q(\phi)\prod_{i}^{n} q(\lambda_i). \tag{6}$$

The variational Bayesian cost function is the negative variational free energy, $\mathcal{F}$, which is a lower bound on the log model evidence (Beal, 2003). As $\mathcal{F} = \log P(\mathbf{y}) - \mathcal{KL}$, where $\mathcal{KL}$ is the always positive Kullback–Leibler distance between the unknown true posterior and our approximate posterior distributions, the maximisation of $\mathcal{F}$ leads to the minimisation of $\mathcal{KL}$. The derivation of $\mathcal{F}$ for this model is given in Appendix A, and a condensed form is given in Eq. (14).

Typically, the functional forms of the approximate posterior distributions can be derived algebraically from the model formulation. In this case:

$$q(\mathbf{w}) = \mathcal{N}(\mu, \boldsymbol{\Upsilon}) \tag{7}$$

$$q(\phi) = Ga(a, b), \tag{8}$$

where $\boldsymbol{\mu}$ is the mean of the posterior distribution on the transformation parameters, and $\mathbf{Y}$ describes the posterior covariance of these parameters. a and b are the shape and scale parameters of $q(\phi)$, respectively.

Through the calculus of variations, iterative analytic updates can be found for the parameters of the approximate posterior distributions $q(\mathbf{w})$ and $q(\phi)$. Briefly, the nature of these updates involves finding the zero-derivative of the functional $\mathcal{F}$ with respect to a particular parameter group. As an example, the optimal value of $q(\mathbf{w})$ would be found conditional on the approximate posterior distribution of the other model parameters $q(\phi)\prod_i q(\lambda_i)$.

2.3.1. Regularisation parameters

Unlike the single regularisation hyper-parameter case described in previous work (Simpson et al., 2012b), where $q(\lambda)$ can also be derived as following a Gamma distribution, the spatially localised hyper-parameters cannot be algebraically determined as following a particular distribution. This is because λ_i appears within a matrix inverse in $\mathcal{F}$ (see Appendix B), which also complicates the marginalisation of these parameters.

To allow inference, and marginalisation, of these parameters within a tractable framework, two further approximations are required. Firstly, the Laplace approximation is used to assume a normal posterior form for $q(\lambda_i) = \mathcal{N}(\hat{\lambda}_i, \sigma_i^2)$. Secondly, it is assumed that the prior covariance matrix only depends on the first order moments of λ_i, which greatly simplifies the marginalisation of $q(\lambda_i)$ and the estimation of σ_i^2. The expectation of the prior covariance matrix, $\boldsymbol{\Sigma}$, can now be written as:

$$\langle \boldsymbol{\Sigma} \rangle_{\prod_i^{N_c} q(\lambda_i)} = \sum_{i}^{N_c} \exp(\hat{\lambda}_i)\Sigma_i, \tag{9}$$

where the angular brackets correspond to an expectation of the encompassed term with respect to the subscript.

2.3.2. Inference of transformation and noise parameters

The updates for the transformation and noise parameters are derived in the same way as (Simpson et al., 2012b), taking the expectation of the prior covariance matrix with respect to $\prod_i q(\lambda_i)$ as given in Eq. (9). As $\mathbf{t}(\mathbf{x}, \mathbf{w})$ is non-linear with respect to the transformation parameters, $\mathbf{w}$, a first order Taylor series approximation is used to locally linearise the function about the current mean estimate. This requires the calculation of the matrix of partial derivatives, $\mathbf{J}$, of $\mathbf{t}(\mathbf{x}, \mathbf{w})$ with respect to $\mathbf{w}$ about the current mean μ_{old}, $\mathbf{J}_{ij} = \frac{\partial \mathbf{t}(\mathbf{x},\mathbf{w})_i}{\partial \mathbf{w}_j}|_{\mathbf{w}=\mu_{old}}$. The transformation mean, μ, and covariance $\mathbf{Y}$ are updated by:

$$\boldsymbol{\Upsilon} = (\alpha\bar{\phi}\mathbf{J}^{\mathsf{T}}\mathbf{J} + \boldsymbol{\Sigma}^{-1})^{-1} \tag{10}$$

$$\mu_{new} = \boldsymbol{\Upsilon}\left[\alpha\bar{\phi}\mathbf{J}^{\mathsf{T}}(\mathbf{J}\mu_{old} + \mathbf{k})\right], \tag{11}$$

where $\mathbf{k}$ is the vector representing the residual image $\mathbf{y} - \mathbf{t}(\mathbf{x}, \mathbf{w})$. μ_{new} describes the current estimated transformation parameters, and is dependent on the old estimated values, μ_{old}. $\bar{\phi} = ab$, which is the expectation of the estimated noise precision.

The posterior parameters of $q(\phi)$ are updated by:

$$b = b_0 + \frac{N_v\alpha}{2} \tag{12}$$

$$\frac{1}{a} = \frac{1}{a_0} + \frac{1}{2}\alpha(\mathbf{k}^{\mathsf{T}}\mathbf{k} + \mathrm{Trace}(\boldsymbol{\Upsilon}\mathbf{J}^{\mathsf{T}}\mathbf{J})) \tag{13}$$

where N_v is the count of voxels within the masked region.

2.3.3. Inference of regularisation parameters

A different but consistent inference mechanism is required to infer the spatial prior parameters, $\{\lambda\}$, from the data. As described in Section 2.3.1, the Laplace approximation uses a Taylor series expansion of $\mathcal{F}$ to estimate a normal distribution for $q(\lambda)$. Based on this approximation, these parameters can be inferred through Newton's method updates with respect to the variational Bayesian cost function, $\mathcal{F}$. Given the mean-field approximation in Eq. (6), and the resulting $\mathcal{F}$ described in Appendix A, the optimistion of $\{\lambda\}$ purely involves terms from the minimisation of the Kullback–Liebler distance between the prior and posterior distributions of $\mathbf{w}$, as $\{\lambda\}$ is a component of the prior on $\mathbf{w}$ (see Eq. (9)), and the prior and posterior of λ_i. The terms from $\mathcal{F}$ that contain $\{\hat{\lambda}\}$, or $\boldsymbol{\Sigma}$, are:

$$\mathcal{F} = \frac{1}{2}\left(-\log|\boldsymbol{\Sigma}| - \mathrm{Trace}(\boldsymbol{\Upsilon}\boldsymbol{\Sigma}^{-1}) - \mu\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{\rho^2}\sum_i(\hat{\lambda}_i - \eta)^2\right) + \mathrm{const}[\{\hat{\lambda}\}, \boldsymbol{\Sigma}], \tag{14}$$

where $\mathrm{const}[\{\hat{\lambda}\}, \boldsymbol{\Sigma}]$ contains all terms that are constant with $\{\hat{\lambda}\}$ and $\boldsymbol{\Sigma}$.

The derivation of the 1st and 2nd order partial derivatives of Eq. (14) are given in full in Appendix B. The derivative of $\mathcal{F}$ with respect to the mean of each local regularisation control parameter, $\hat{\lambda}_i$,

can be expressed as:

$$\frac{\partial \mathcal{F}}{\partial \hat{\lambda}_i} = \frac{1}{2}\left[-\text{Trace}\left(\boldsymbol{\Upsilon}\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \hat{\lambda}_i}\right) + \text{Trace}\left(\boldsymbol{\Sigma}\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \hat{\lambda}_i}\right) - \mu^T \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \hat{\lambda}_i}\mu\right] - \frac{\hat{\lambda}_i - \eta}{\rho^2} \tag{15}$$

where

$$\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \hat{\lambda}_i} = -\boldsymbol{\Sigma}^{-1}\exp(\hat{\lambda}_i)\Sigma_i\boldsymbol{\Sigma}^{-1} \tag{16}$$

The second partial derivative, taking advantage of the approximation that $\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \hat{\lambda}^2} = 0$, is simply written as:

$$\frac{\partial^2 \mathcal{F}}{\partial \hat{\lambda}_i^2} = \text{Trace}\left(\exp(\hat{\lambda}_i)\Sigma_i\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \hat{\lambda}_i} - \frac{1}{\rho}\right). \tag{17}$$

As such, $q(\lambda_i)$ can be updated according to the derivatives in Eqs. (15) and (17), where

$$\frac{1}{\sigma_i^2} = -\frac{\partial^2 \mathcal{F}}{\partial \lambda_i^2}, \tag{18}$$

and the posterior mean $\hat{\lambda}$ is updated by:

$$\hat{\lambda} = \hat{\lambda} + \frac{\partial \mathcal{F}}{\partial \lambda_i}\sigma_i^2. \tag{19}$$

2.4. Model comparison

The negative variational free energy, $\mathcal{F}$, is an objective means for allowing comparison of models without requiring ground truth, or gold standard information. $\mathcal{F}$ summarises the fit of the data, and the deviation of the model parameters from their prior distributions. Unlike the Bayesian information criteria, $\mathcal{F}$ only penalises model parameters that deviate from the prior, and the cost of a parameter that retains the same distribution as the prior is zero. In the case of the proposed model, this means that the complexity of having additional λ parameters that only take the prior distribution, have no additional cost.

Although $\mathcal{F}$ has been previously used for model comparison in the medical image analysis domain (Groves et al., 2009; Penny et al., 2005; Friston et al., 2008), to the best of the authors' knowledge it has never been used in medical image registration. However, previous attempts at probabilistic model selection have appeared using the minimum description length in Van Leemput (2009) and Marsland et al. (2008) and information theoretic model selection approaches include Schnabel et al. (2001), Rohde et al. (2003) and Hansen et al. (2008).

2.5. Selection of p(λ)

The prior on the regularisation control parameters, $p(\lambda)$, has an important effect. If there is little information from the data to suggest a value for these parameters, then they will tend to take the values of the prior. As described previously, $p(\lambda) = \mathcal{N}(\eta, \rho)$. As our interests lie in a more interpretable formulation of registration, we therefore only wish to see deformations that are reliably driven by the data. As such, a low value for η would be preferable, such that in the absence of information to suggest otherwise, transformation parameters would tend towards the identity transformation. Conversely, we want the value of λ to be strongly driven by the data, hence, we choose a large value for ρ. The influence of $p(\lambda)$ can be thought of as selecting the prior probability of different scales of deformations being allowable. In this work a weakly informative prior is chosen, where $\eta = -6$ and $\rho = 40$.

2.6. Implementation and initialisations

This algorithm was implemented within the FMRIB Non-linear Image Registration Tool (FNIRT) (Andersson et al., 2007), which provides the facility for efficient calculation of the Hessian of the transformation parameters, $\mathbf{J}^T\mathbf{J}$. The algorithm uses a 3 level multi-resolution scheme where the image is down-sampled, initially by a factor of 4, then 2, then full-resolution. The B-spline knots are super-sampled through interpolation at each new level to yield a higher resolution grid. The final spacing is given in the experimental description. The original regularisation model is bending energy, described as an inverse covariance matrix, the scale of which is either adaptively inferred, as in Simpson et al. (2012b), or manually selected.

In terms of initialisation, at the first multi-resolution level, $\{\hat{\lambda}\}$ are set to give an initial control point variance of 2 mm. The first three updates at the first level perform a global scaling of the initial prior matrix. Subsequent iterations treat each λ independently.

Between multi-resolution levels, $\{\hat{\lambda}\}$ is interpolated using trilinear interpolation. A maximum of 20 iterations was run for each multi-resolution level, with convergence defined by: $\mathbf{k}^T\mathbf{k} + \mu\boldsymbol{\Sigma}^{-1}\mu$, which is the sum of squared differences plus the deviation of the transformation mean, from the prior instead of $\mathcal{F}$ for computational convenience.

3. Synthetic experiments

Synthetic 2D images were created to demonstrate the effects of this algorithm, see top row of Fig. 3. 10 instances of two 2D phantom images, 30 × 30 pixels, were created with varying SNR. As reference image, a circle with a radius of 10 pixels, and a floating image, which is two pixels thinner on one side. An ideal transformation that links these two images should be spatially localised to the area of shrinkage and have very high confidence in the transformation parameters at all other locations.

3.1. Visualisation of uncertainty

The distributions of the posterior transformation parameters $q(\mathbf{w})$ and of the transformation prior $p(\mathbf{w})$ are multivariate normal. In order to display the uncertainty of the posterior, or the support of the prior, in this work the sum of the variance in each direction is summed and the result is square rooted to give an uncertainty value in pixels/mm. This is approximated as the variance at each of the knot points and interpolated over the image using the B-spline basis set.

3.2. Example registration

An example set of images and registration results at two SNRs is given in Fig. 3. The log Jacobian maps show that when using the proposed prior the deformation is well localised to the region of change, as opposed to using an adaptive bending energy prior as in Simpson et al. (2012b), where the deformation propagates across the entire circle, despite there being no local image information to support this. The reason for this localisation is that the spatial prior only supports deformation within certain areas. Consequently, this provides a more interpretable estimation of registration uncertainty, where the uncertain regions are only in the areas of change rather than across the image.

3.3. Model comparison

Bayesian model selection can be used to objectively choose model parameters that cannot be inferred directly from the data. Here, we investigate the effects of the number of transformation parameters, in terms of B-spline knot spacing, as well as the form of the spatial prior on $\mathcal{F}$ at two SNRs. This is plotted in Fig. 4. For both SNRs, using the proposed prior leads to an improvement in $\mathcal{F}$ over bending

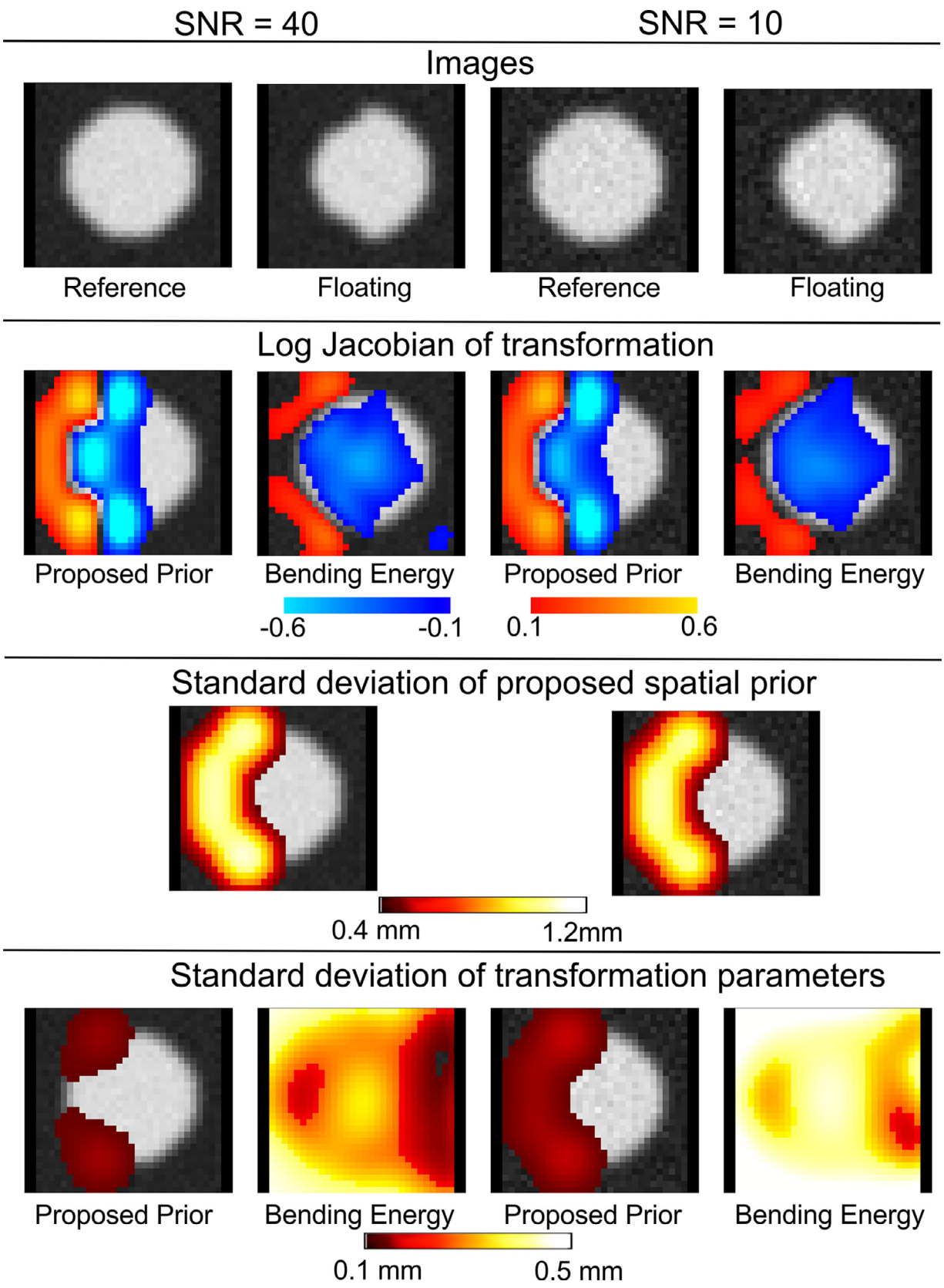

Fig. 3. Illustrative simulated registration examples. The results were calculated using a B-spline knot spacing of 5 pixels, for the proposed prior $\sigma = 0.1$. These parameters values were chosen as they provide relatively good results at both SNRs in terms of $\mathcal{F}$, see Fig. 4. The top row shows the synthetic reference and floating image at two signal to noise ratios (SNRs). The second row shows the resulting log Jacobian map, illustrating expansion or contraction, when using the proposed prior or an adaptive level of bending energy. The third row illustrates the standard deviation of the proposed spatial prior, which is well localised to the region of deformation. The final row shows the uncertainty of the posterior distribution of transformation parameters using either the proposed prior or bending energy.

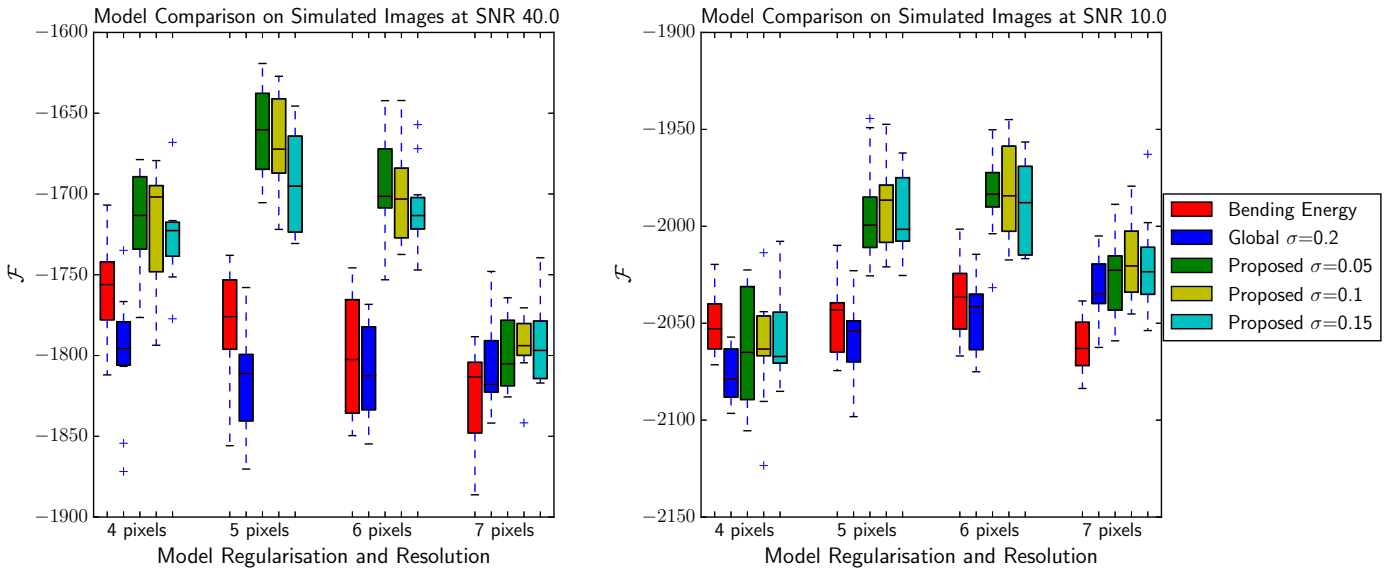

Fig. 4. Bayesian model comparison, using the negative variational free energy $\mathcal{F}$, comparing regularisation strategy and B-spline knot spacing using simulated images. The legend describes the regularisation strategy, where σ is the parameter of the GP prior in Eq. (4). Global refers to the use of a global weight for the GP prior.

energy and a global version of the Gaussian process prior henceforth GP prior, where $\sigma = 0.2$ is shown as it gave the best average values for $\mathcal{F}$, despite the increased number of parameters. The exception to this is where a 4 pixel B-spline knot spacing resolution was used with low SNR data, where bending energy fares slightly better. Interestingly, a slightly higher value of σ is preferable at lower SNR, which leads to greater spatial covariance in the prior. A 5 pixel knot spacing seems to provide the best balance of complexity and data fitting at both SNRs for this example.

4. Real data experiments

4.1. Materials

Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the FDA, private pharmaceutical companies and non-profit organisations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada.

60 structural MR images acquired on 3 T scanners were taken from the ADNI database, 30 of these subjects suffered from AD, the other 30 are healthy controls (HC). There were 18 males with AD and 12 male HC. The age means and standard deviations were 74.3 (8.4) for AD and 70.1 (13.95) for HC. The AD subjects were taken from 10 different sites and the HC from 7.

4.2. Cross sectional TBM

A single high-resolution representative atlas was constructed for use in the tensor based morphometry experiments. Having a common atlas allows direct comparison of the TBM results from the different regularisation approaches. To prevent bias towards a particular regularisation strategy, an entirely different approach was used to create the atlas. The atlas was created by first probabilistically segmenting the images into grey and white matter, followed by co-registering these probability maps into a common space using the geodesic shooting approach (Ashburner and Friston, 2011) within SPM12 beta. The bias corrected images were then resampled into the atlas space and averaged to create the atlas.

Each of the bias field corrected subject images was rigidly registered to the template image using FLIRT (Jenkinson and Smith, 2001). Subsequently, each image was non-linearly registered to the atlas space using one of six regularisation strategies: a fixed level of bending energy (Andersson et al., 2007), a globally adaptive level of bending energy, where the level is inferred from the data as in Simpson et al. (2012b), a global GP prior and the proposed prior where $\sigma = \{0.05, 0.1, 0.15\}$. All registrations were run to a 10 mm B-spline knot spacing. 10 mm was selected for computational reasons, as the current implementation does not provide an efficient mechanism for the inversion of sparse matrices. Following registration, the logarithm of the voxelwise determinant of the Jacobian of the mean transformation, μ, is calculated. This provides a measure of local expansion or contraction.

For the proposed method, $p(\lambda) = \mathcal{N}(-6, 40)$. For the proposed model σ was selected to be 0.1 based on the model comparison described in Section 4.2.1. For the global GP prior, different σ values were tested, but $\sigma = 0.1$ gave the highest score in terms of $\mathcal{F}$ so is presented in all experiments. Two example registrations are given in Figs. 5 and 6.

4.2.1. Model comparison

Model comparison can be used to find the ideal value of σ. In this case we compared the $\mathcal{F}$ for an adaptive level of bending energy, a global GP prior with $\sigma = 0.1$, and the proposed prior with $\sigma = 0.05$, $\sigma = 0.1$ and $\sigma = 0.15$. The results of this model comparison are illustrated in Fig. 7. $\sigma = 0.1$ was chosen for illustration as it generally outperformed $\sigma = 0.15$ and adaptive bending energy, with less variability than $\sigma = 0.05$.

4.2.2. Jacobian analysis

The distinction between the proposed prior, and a global prior can be seen in terms of the distribution of local volume change as given by the log Jacobian, an example histogram of which is given in

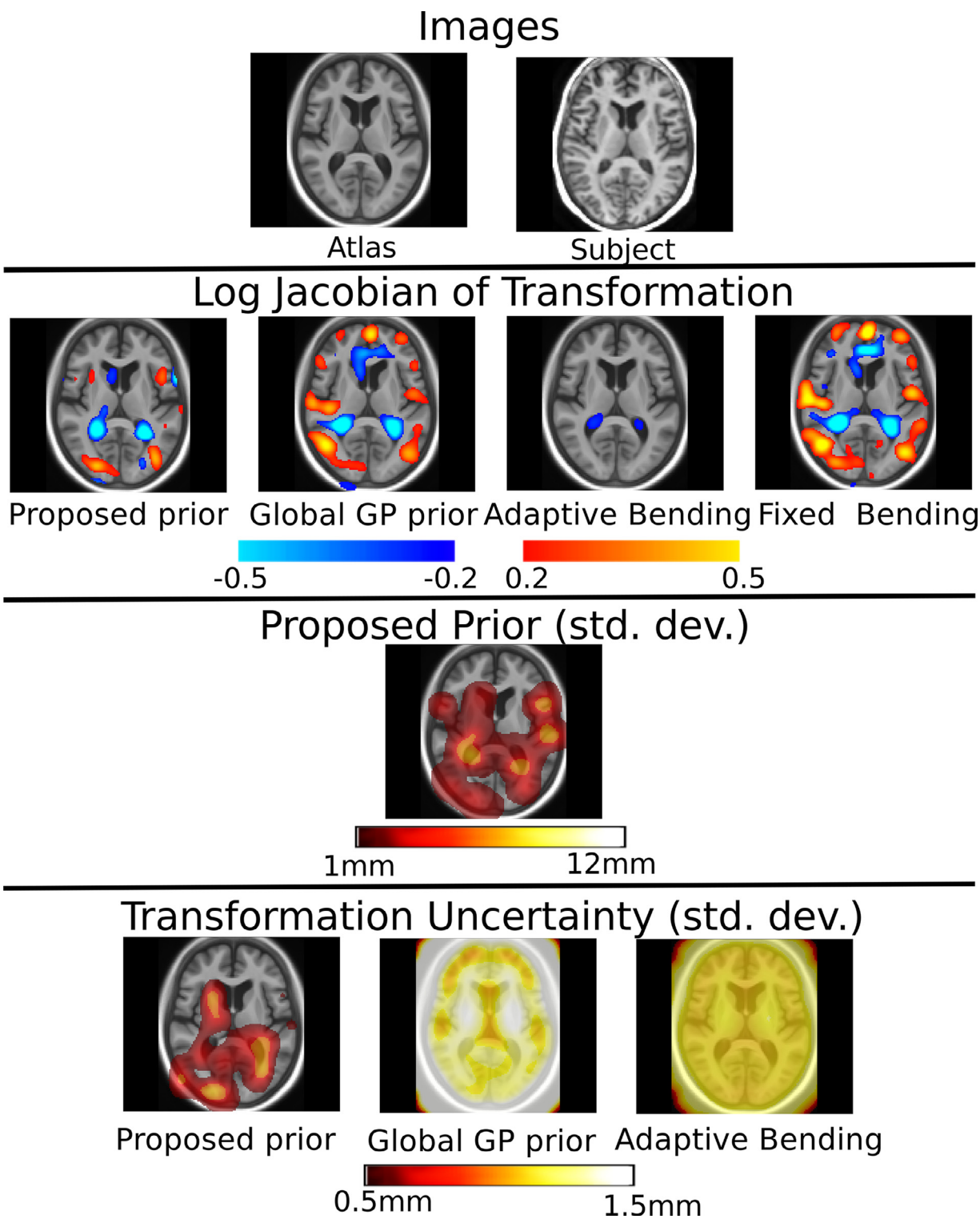

Fig. 5. An example slice illustrating a 3D registration where the substantial volume changes are quite sparsely distributed. In this case, the three methods produce quite different log Jacobian maps. The adaptive global bending energy infers an inflexible transformation prior, as insufficient information globally suggests more flexibility is needed. The fixed level of bending energy produces a lot of changes across the brain, the causes of some are not immediately apparent from visual inspection of the data. The global GP prior which does not encourage particularly strong spatial smoothness performs similarly. Conversely, using the proposed prior leads to a sparser set of volume changes that subjectively seem more reasonable, and contain less false positives.

Fig. 8. The proposed prior prohibits much displacement in uninformative regions, thus leads to large regions of no volume change. Furthermore, in informative regions the registration is free to follow the data completely leading to more substantial volume changes, which are seen in the tails of the distributions. This emphasises the well supported signal from the data, and reduces other effects. This can be measured using the kurtosis of the log Jacobian distribution, where higher kurtosis implies a more peaked distribution, with heavier tails. Fig. 9 shows a boxplot of the kurtosis of the log Jacobian maps across the population.

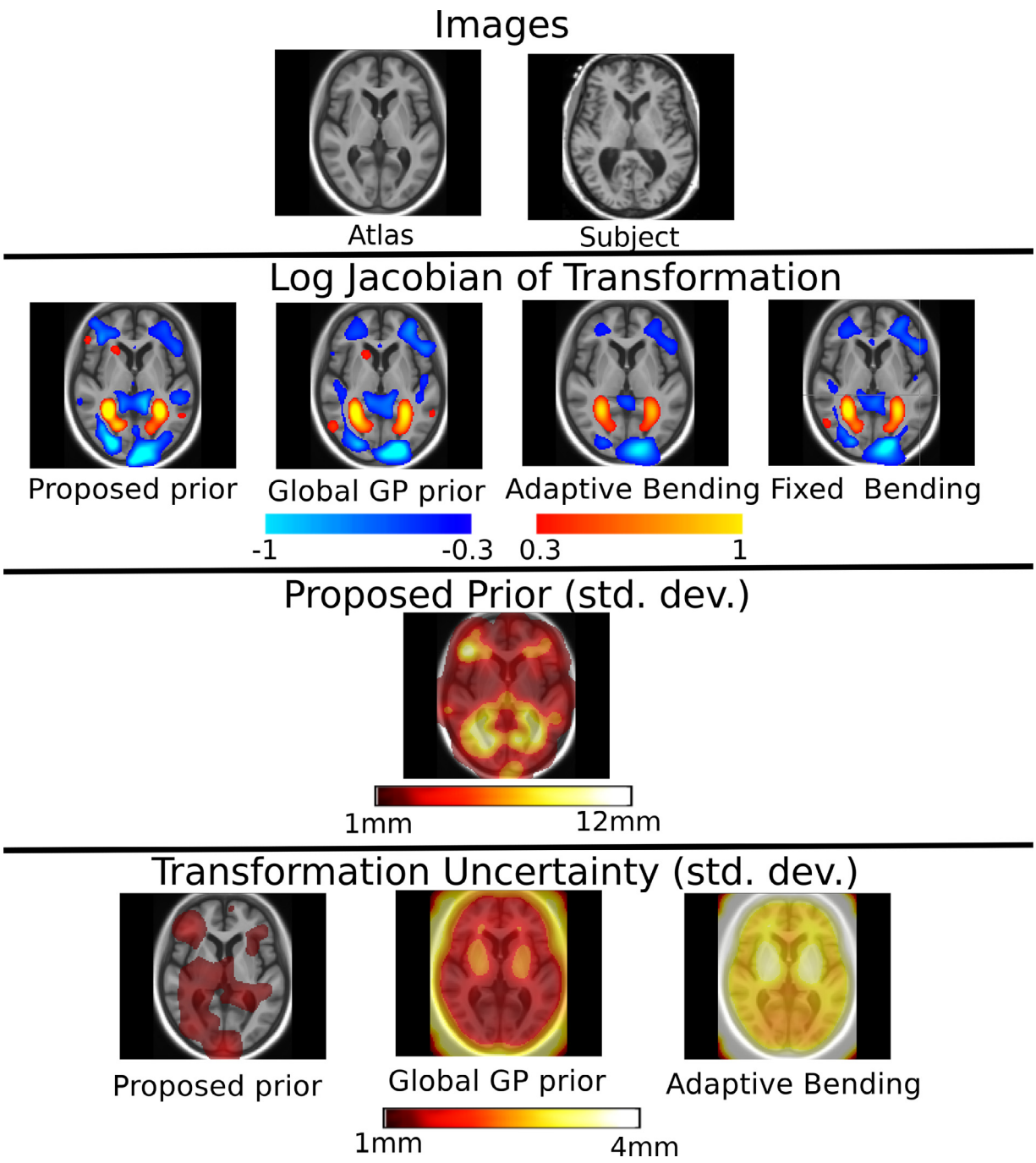

Fig. 6. An example slice illustrating a 3D registration where there are changes distributed across the whole brain. As can be seen, all four methods produce similar log Jacobian maps. The proposed spatial prior shows fairly wide flexibility across the image with more flexibility in the anterior, as there are more substantial changes there. This illustrates that the proposed prior is appropriate even in cases where the changes are widely distributed. The spatial uncertainty is much lower and more focal than when using either of the adaptive global priors.

4.2.3. *Population statistics*

The log Jacobian maps were analysed using a general linear model, where statistical differences were evaluated between subject groups. The Jacobian maps were not smoothed prior to analysis. All the analyses were performed using tools from the FSL library.[1] Age and total intracranial volume (TIV), as estimated by combining the white matter, grey matter and CSF maps from SPM, were used as co-regressors. Fig. 10 shows the results of these statistical analyses.

[1] www.fmrib.ox.ac.uk/fsl/ .

5. Discussion

This paper has demonstrated that a spatially adaptive transformation prior can be estimated alongside the non-linear registration parameters from a pair of images. The current framework was implemented using a B-spline FFD transformation model but the method itself is independent of the transformation model. The inferred spatial prior aims to reduce the Kullback–Leibler distance between the prior and posterior distributions of the transformation parameters and consequently derives information from the data in terms of the level of local image information, and areas where

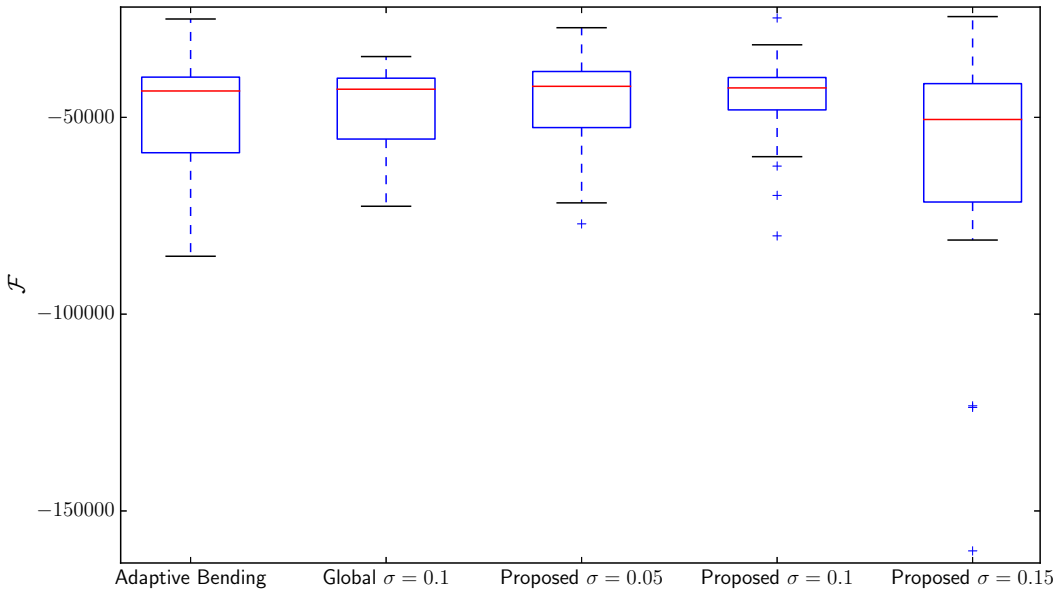

Fig. 7. Bayesian model comparison of the different regularisation strategies for population to atlas registration. $\mathcal{F}$ was significantly lower for $\sigma = 0.15$ than all other methods (paired t-test, $p < 0.05$). $\sigma = 0.05$ and $\sigma = 0.1$ are fairly similar, and weakly significantly better than the adaptive bending energy regulariser (paired t-test, $p < 0.06$) and the global GP prior (paired t-test, $p < 0.12$). As $\sigma = 0.1$ has a smaller inter-quartile range, and similar median to $\sigma = 0.05$, this was used in future experiments.

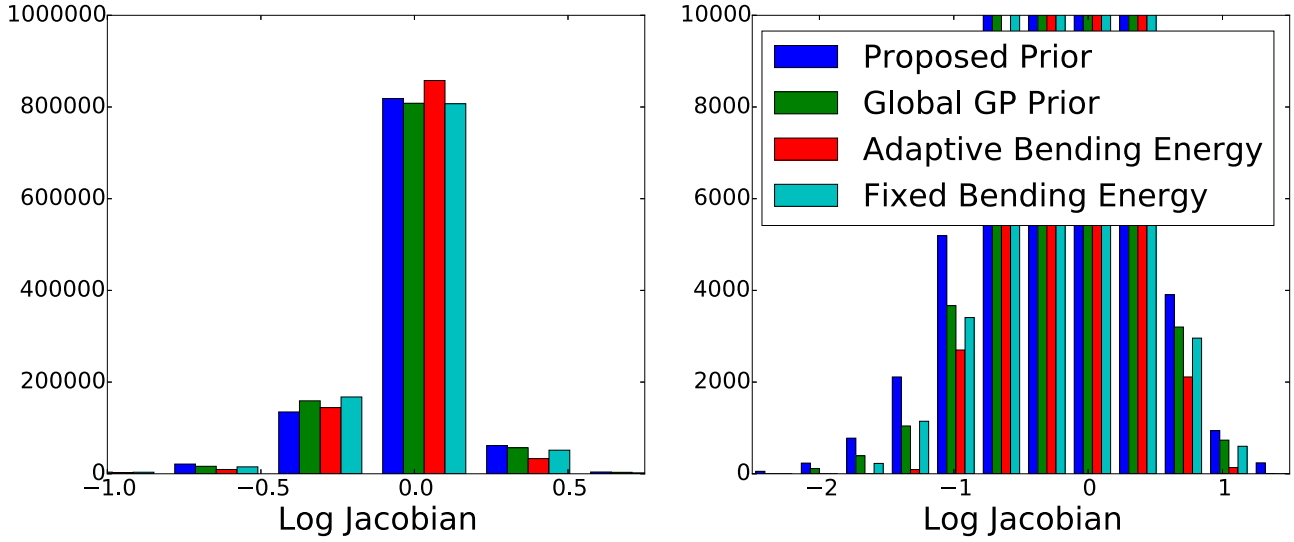

Fig. 8. Histograms of the log Jacobian values from the registrations in Fig. 6. The left image shows the overall distributions, whereas the right plot focuses on the tails of the same distributions. As can be seen, using the proposed prior leads to substantially heavier tails. In this case, the kurtosis varies from 7.7, for adaptive bending energy, 8.6, for the fixed level of bending energy, 9.1 for the global GP prior, which encourages less smooth deformations than bending energy, and 15.0 for the proposed regularisation prior.

deformations occur. In other areas, the spatial prior has very low variance allowing little displacement to occur. This can lead to sparse deformations, as shown in Fig. 5, where the registration is very free in informative areas allowing larger volume changes, and constrained in other areas prohibiting volume change. This leads to distributions of log Jacobians that have higher kurtosis. We postulate that this may lead to a reduction is weaker false positives, and emphasises true volume changes in the data.

This model can be thought of as equivalent to a sparse deformation model, where the hyper-parameters controlling regularisation $\{\lambda\}$ can effectively switch off transformation parameters in non-informative regions, therefore the deformation in those locations cannot be uncertain, as it not being estimated. For alternative applications to TBM, a map of inactive regions may be useful, as the alignment of these regions cannot be deemed trustworthy, an intuition for these locations can be seen in the proposed prior maps in Figs. 5 and 6.

In computational terms, the current implementation is quite expensive, which limits the B-spline knot resolution that this method has been tested on. The computational bottleneck lies in the numerical inverse of the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Upsilon}^{-1}$. Future work would seek to find efficient means of inverting the matrices, possibly using a sparse Cholesky decomposition that allows updating, or through separating the matrix into blocks as in Harrison et al. (2008).

Ideally, a regularisation strategy would not enforce sparsity on the covariance matrix. Instead, it may be more appropriate to have a spatially adaptive prior as a mixture of precision, rather than covariance components. This would permit longer range covariance in the prior, which cannot occur in the proposed work. A difficulty with such an approach is learning a suitable set of prior

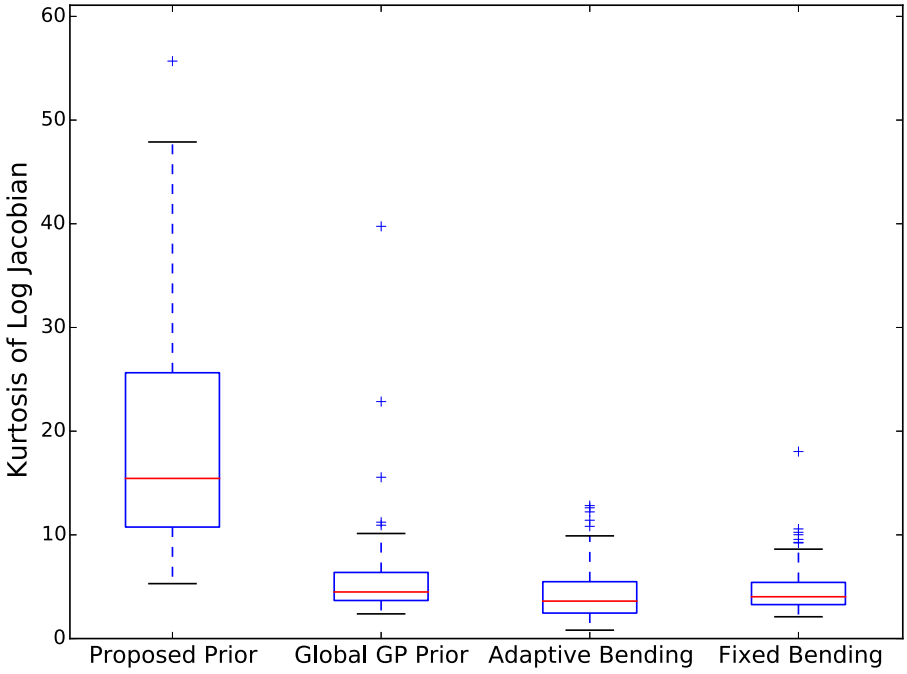

Fig. 9. Boxplots illustrating the distribution of kurtosis in the log Jacobian maps between the different priors across the 60 registrations. The proposed prior has significantly higher kurtosis then the other methods ($p < 0.05$ paired t-test).

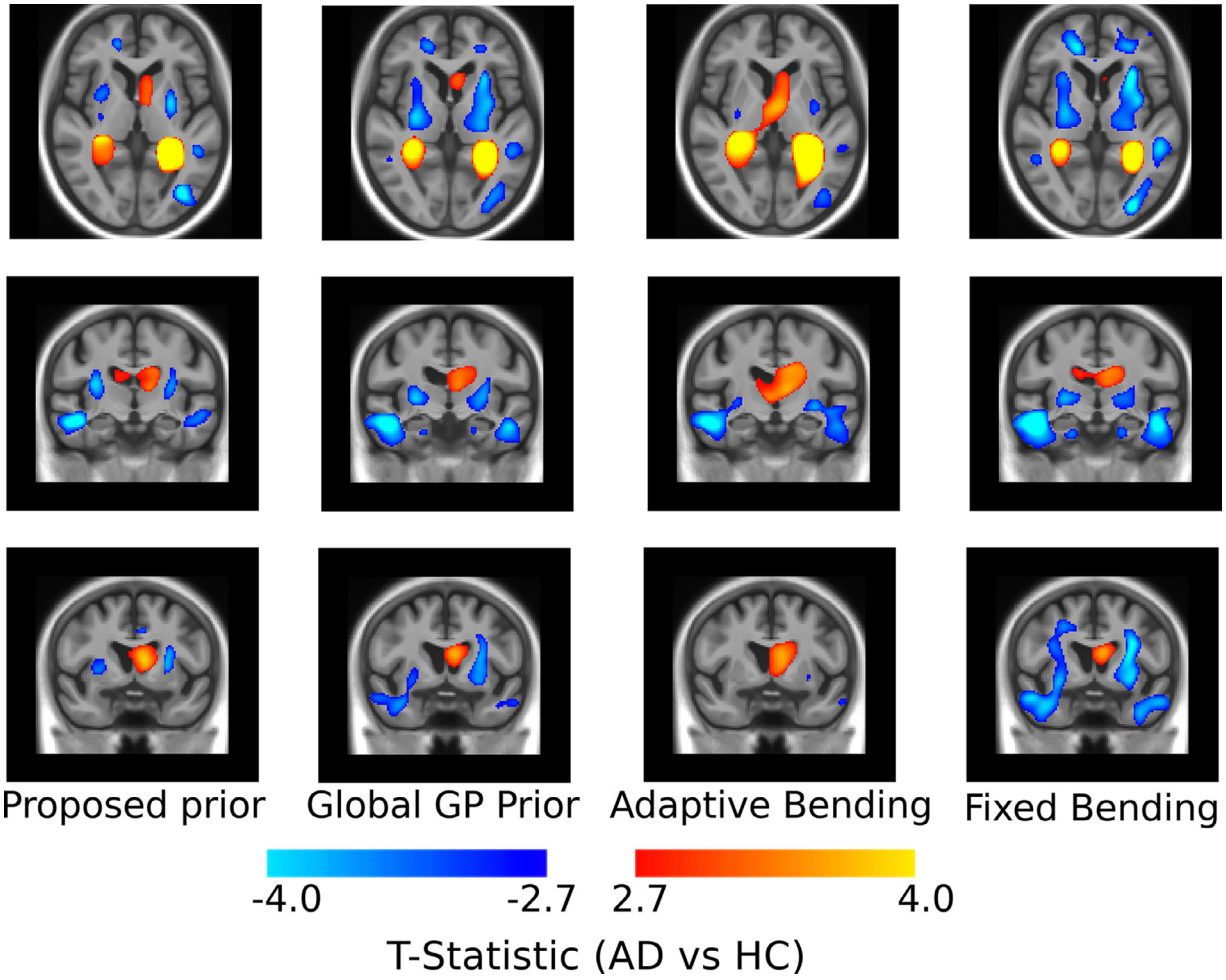

Fig. 10. Population t-statistics (uncorrected) comparing the population with AD and HC. As can be seen, the fixed level of bending energy and global GP prior leads to more widespread changes, particularly in the white matter visible in the bottom row. These may be false positive effects caused by higher global variance than the other methods, or lower spatial smoothness in the case of the global GP prior. The proposed prior leads to focal contractions of high significance in the grey matter and expansion of the ventricles, which may be more plausible.

components to use, and ensuring that the resulting prior matrix is positive-definite.

In the current implementation, the subject images were registered to the atlas to allow the deformation fields (and therefore the Jacobian maps) to be in a common space. However, in a generative model such as this, it would be more appropriate to register the smooth atlas image to the subject for estimating the deformation field. As we are currently using a small deformation model, the inverse is not always well defined and therefore such an approach may not be ideal. Future work will implement this model within a large deformation transformation model, such as a stationary velocity field.

A straightforward extension of this work would investigate the use of a population prior distribution of $p(\lambda)$ that has a variable mean and variance across the image. Furthermore, local covariance components could be merged together where appropriate as in Friston et al. (2008).

Registration uncertainty has been demonstrated to be useful in improving hippocampal subfield segmentations (Iglesias et al., 2013), estimating dose delivery in radiotherapy (Risholm et al., 2011a), assisting neurosurgical decision making (Risholm et al., 2010a) and improving classification (Simpson et al., 2013a). Future work could also investigate the use of posterior deformation distributions to identify whether an individual belongs to a sub-population of the data, either globally or for a specific structure. This work demonstrates how strongly the registration uncertainty depends on the prior information, as well as the local image information. The use of a global spatial prior leads to a global variance contribution, which is modified based on the local image information. Conversely, with an adaptive spatial prior, areas that are informative are given freedom to move, but because they are informative regions, they consequently lead to low variance. As opposed to areas that are uninformative, which are given little freedom in the prior and therefore have a tight posterior distribution as there is no evidence to suggest that they should move.

We believe that this paper provides the first example of Bayesian model comparison for non-linear registration, as demonstrated for choosing the form of the regularisation model. Future work will also investigate finding an optimal B-spline knot spacing or transformation model for a given application.

6. Conclusions

This paper has described a spatially adaptive regularisation prior model and inference scheme for non-linear registration. The components are optimised using the variational Bayesian cost function, which aims to reduce the Kullback–Leibler distance between the prior and posterior distribution of transformation parameters. This approach leads to better feature localisation and a reduction of false positives in tensor based morphometry, through having a spatial prior that adapts to the local data. Further advantages are Bayesian model comparison and allowing for more plausible measures of registration uncertainty.

Acknowledgements

I. Simpson was supported by the NIHR Queen's Square Dementia BRU. M.J. Cardoso was supported by EPSRC (EP/H046410/1). Marc Modat is supported by the UCL Leonard Wolfson Experimental Neurology Centre. Sebastien Ourselin receives funding from the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense Award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Derivation of the variational free energy

The negative variational free energy, $\mathcal{F}$, is a lower bound of the log model evidence, and is the measure that VB seeks to maximise (Beal, 2003). Maximisation of $\mathcal{F}$ is equivalent to minimisation of the Kullback–Leibler distance between the true and approximate posterior distributions. For a model with parameters Θ, $\mathcal{F}$ is composed of two terms:

$$\mathcal{F} = \int q(\Theta)\log P(\mathbf{y}|\Theta)d\Theta + \int q(\Theta)(\log P(\Theta) - \log q(\Theta))d\Theta \tag{A.1}$$

$$= \mathcal{L}_{av} - D_{KL}(q(\Theta)||P(\Theta)) \tag{A.2}$$

where $\mathcal{L}_{av}$ is the marginal value of the log likelihood with respect to the approximate posterior distribution, $q(\Theta)$, and D_{KL} is the Kullback–Leibler distance between the approximate posterior and prior distributions.

The mean-field approximation assumes independence of groups of parameters, and for the model in question: $q(\Theta) = q(\mathbf{w})q(\phi)\prod_i q(\lambda_i)$. Therefore, for the proposed model $\mathcal{L}_{av}$ is calculated as the expectation of the likelihood with respect to the approximate posterior distributions:

$$\mathcal{L}_{av} = \int q(\mathbf{w})q(\phi)\prod_i^I q(\lambda_i)(\log P(\mathbf{y}|\Theta)d\mathbf{w}\,d\phi\,d\lambda_i \tag{A.3}$$

This results in the following expression for the marginal likelihood:

$$\mathcal{L}_{av} = \frac{\alpha N_v}{2}(\log(a) + \psi(b)) - \frac{\alpha\bar{\phi}}{2}(\mathbf{k}^T\mathbf{k} + \mathrm{Tr}(\boldsymbol{\Upsilon}\mathbf{J}^T\mathbf{J})) \tag{A.4}$$

where ψ is the digamma function.

Similarly, D_{KL} comprises the integral of the second term of Eq. (A.1). Due to the mean-field approximation, $D_{KL}(\Theta)$ is split into approximate posterior parameter groups:

$$D_{KL}(q(\Theta)||P(\Theta)) = D_{KL}(q(\mathbf{w})||p(\mathbf{w})) + D_{KL}(q(\phi)||P(\phi)) + \sum_i^I D_{KL}(q(\lambda_i)||P(\lambda_i)) \tag{A.5}$$

These are the standard Kullback–Leibler distances between either normal, or Gamma distributions and can be found in the literature (Roberts and Penny, 2002).

Closed form updates for the parameters of the approximate posterior distributions can be derived using the calculus of variations. This involves finding the derivative of the functional $\mathcal{F}$ with respect to a set of model parameters, given the current posterior distribution on the conditionally independent model parameters. In practical terms, this involves equating the log-likelihood and prior probabilities, marginalised over the independent posterior distributions, with the approximate log posterior distribution. For example, if:

$$\mathcal{M} = \log p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \phi) + \log p(\mathbf{w}) + \log P(\phi) + \sum \log P(\lambda) \quad \text{(A.6)}$$

then the updated distribution for $q(\mathbf{w})$ can be found as:

$$\log q(\mathbf{w}) = \langle \mathcal{M} \rangle_{q(\phi)\prod_i q(\lambda_i)} \quad \text{(A.7)}$$

where the angled brackets correspond to taking an expectation of the bracketed term with respect to the sub-scripted terms. The full derivation of the updates for $q(\mathbf{w})$ and $q(\phi)$ are not given here, but can be found in previous work (Simpson et al., 2012b).

Appendix B. Regularisation parameters

The terms of $\mathcal{F}$ that relate to the prior covariance matrix, $\boldsymbol{\Sigma}$ are given as:

$$\mathcal{F} = \frac{1}{2}\left(-\log|\boldsymbol{\Sigma}| - \text{Trace}(\boldsymbol{\Upsilon}\boldsymbol{\Sigma}^{-1}) - \mu\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{\rho^2}\sum_i (\hat{\lambda}_i - \eta)^2 \right) + \text{const}\{\hat{\lambda}_i\} \quad \text{(B.1)}$$

As can be seen, Σ appears twice within a matrix inverse. As $\{\lambda\}$ parameterises Σ, rather than Σ^{-1}, $q(\lambda)$ does not have an algebraically defined posterior distribution. Instead, the Laplace approximation is used to assume a normal posterior distribution, by taking a Taylor series expansion of $\mathcal{F}$ around the current mean. Furthermore, it is assumed that Σ only depends on the first order moments of λ, as described in Eq. (9).

B.1. First order derivative

Each of these terms can be analytically differentiated with respect to the posterior mean of a given regularisation parameter, $\hat{\lambda}_i$:

$$-\frac{\partial}{\partial\hat{\lambda}_i}\log|\boldsymbol{\Sigma}| = \frac{\partial}{\partial\hat{\lambda}_i}\log|\boldsymbol{\Sigma}^{-1}| = \frac{1}{|\boldsymbol{\Sigma}^{-1}|}\frac{\partial|\boldsymbol{\Sigma}^{-1}|}{\partial\hat{\lambda}_i} \quad \text{(B.2)}$$

$$= \text{Trace}\left(\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\right) \quad \text{(B.3)}$$

where the identity $\frac{\partial\log|\mathbf{X}|}{\partial\mathbf{X}} = \text{Trace}(\mathbf{X}^{-1}\partial\mathbf{X})$ has been used.

The quantity $\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}$ can be analytically calculated as:

$$\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i} = -\boldsymbol{\Sigma}^{-1}\exp(\hat{\lambda}_i)\Sigma_i\boldsymbol{\Sigma}^{-1} \quad \text{(B.4)}$$

where the identity $\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}A^{-1}$ has been used.

The next term is simply:

$$-\frac{\partial}{\partial\hat{\lambda}_i}\text{Trace}(\boldsymbol{\Upsilon}\boldsymbol{\Sigma}^{-1}) = -\text{Trace}\left(\boldsymbol{\Upsilon}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\right) \quad \text{(B.5)}$$

The derivative of the third term is:

$$-\mu^T\boldsymbol{\Sigma}^{-1}\mu = -\mu^T\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\mu \quad \text{(B.6)}$$

The derivative of the final term is:

$$\frac{\partial}{\partial\hat{\lambda}_i}\frac{(\hat{\lambda}_i - \eta)^2}{2\rho^2} = \frac{2\hat{\lambda}_i - 2\eta}{2\rho^2} = \frac{\hat{\lambda}_i - \mu_\lambda}{\rho^2} \quad \text{(B.7)}$$

This gives the complete derivative of $\mathcal{F}$ with respect to $\hat{\lambda}_i$ as:

$$\frac{\partial\mathcal{F}}{\partial\hat{\lambda}_i} = \frac{1}{2}\left[-\text{Trace}\left(\boldsymbol{\Upsilon}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\right) + \text{Trace}\left(\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\right) - \mu^T\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\mu\right] - \frac{\hat{\lambda}_i - \eta}{\rho^2} \quad \text{(B.8)}$$

B.2. Second order derivatives

The second order derivatives of $\mathcal{F}$ wrt. $\hat{\lambda}_i$ can be used to estimate the step size for the parameter updates. To get the step size of each parameter update, the second derivative of $\hat{\lambda}_i$ w.r.t. $\mathcal{F}$ can be calculated:

$$\begin{aligned}\frac{\partial^2\mathcal{F}}{\partial\hat{\lambda}_i^2} &= \frac{\partial}{\partial\hat{\lambda}_i}\frac{1}{2}\text{Trace}\left(\left[\boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i} - \boldsymbol{\Upsilon}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i} - \mu\mu^T\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\right]\right)\\ &= \frac{1}{2}\text{Trace}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\hat{\lambda}_i}\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i} + \boldsymbol{\Sigma}\frac{\partial\boldsymbol{\Sigma}^{-2}}{\partial\hat{\lambda}_i^2} - \boldsymbol{\Upsilon}\frac{\partial\boldsymbol{\Sigma}^{-2}}{\partial\hat{\lambda}_i^2} - \mu\mu^T\frac{\partial\boldsymbol{\Sigma}^{-2}}{\partial\hat{\lambda}_i^2}\right)\\ &= \frac{1}{2}\text{Trace}\left(\exp(\hat{\lambda}_i)\Sigma_i\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i} + \left(\boldsymbol{\Sigma} - \boldsymbol{\Upsilon} - \mu\mu^T\right)\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i^2} - \frac{1}{\rho^2}\right)\end{aligned} \quad \text{(B.9)}$$

where

$$\begin{aligned}\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i^2} &= -\frac{\partial}{\partial\hat{\lambda}_i}(\boldsymbol{\Sigma}^{-1}\exp(\hat{\lambda}_i)\Sigma_i\boldsymbol{\Sigma}^{-1})\\ &= -\left(\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\exp(\hat{\lambda}_i)\Sigma_i\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\exp(\hat{\lambda}_i)\Sigma_i\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\exp(\hat{\lambda}_i)\Sigma_i\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i}\right)\end{aligned} \quad \text{(B.10)}$$

and the identity $\partial\mathbf{XY} = (\partial\mathbf{X})\mathbf{Y} + \mathbf{X}(\partial\mathbf{Y})$ has been used.

This work makes the assumption that $\boldsymbol{\Sigma}$ only depends on the first order moment of $\hat{\lambda}_i$. This means that $\frac{\partial^2\boldsymbol{\Sigma}^{-1}}{\partial\hat{\lambda}_i} = 0$, which leads to a simplification of Eq. (B.9) as given in Eq. (17).

References

Allassonniére, S., Amit, Y., Trouvè, A., 2007. Toward a coherent statistical framework for dense deformable template estimation. J. R. Stat. Soc. Ser. B 69 (2), 3–29.

Andersson, J., Jenkinson, M., Smith., S., 2007. Non-linear Registration, aka Spatial Normalisation. FMRIB Technical Report TR07JA1. Available from: www.fmrib.ox.ac.uk/analysis/techrep.

Ashburner, J., Friston, K.J., 1999. Nonlinear spatial normalization using basis functions. Hum. Brain Mapp. 7 (4), 254–266.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry – the methods. NeuroImage 11 (6), 805–821.

Ashburner, J., Friston, K.J., 2011. Diffeomorphic registration using geodesic shooting and gauss-newton optimisation. NeuroImage 55 (3), 954–967.

Attias, H., 2000. A variational Bayesian framework for graphical models. Adv. Neural Inf. Process. Syst. 12 (1–2), 209–215.

Beal, M.J., 2003. Variational Algorithms for Approximate Bayesian Inference (Unpublished Doctoral dissertation). University College London.

Bookstein, F.L., 1997. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med. Image Anal. 1 (3), 225–243.

Chung, M.K., Worsley, K.J., Paus, T., Cherif, C., Collins, D.L., Giedd, J.N., Rapoport, J.L., Evans, A.C., 2001. A unified statistical approach to deformation-based morphometry. NeuroImage 14 (3), 595–606.

Davatzikos, C., 1997. Spatial transformation and registration of brain images using elastically deformable models. Comput. Vis. Image Underst. 66 (2), 207–222.

Durrleman, S., Allassonnière, S., Joshi, S., 2013. Sparse adaptive parameterization of variability in image ensembles. Int. J. Comput. Vis. 101 (1), 161–183.

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., Mattout, J., 2008. Multiple sparse priors for the M/EEG inverse problem. NeuroImage 39 (3), 1104–1120.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34 (1), 220–234.

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: Proceedings of the 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 (3), 2198–2217.
Groves, A.R., Chappell, M.A., Woolrich, M.W., 2009. Combined spatial and non-spatial prior for inference on MRI time-series. NeuroImage 45 (3), 795–809.
Hansen, M.S., Larsen, R., Glocker, B., Navab, N., 2008. Adaptive parametrization of multivariate b-splines for image registration. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008. IEEE, pp. 1–8.
Harrison, L., Penny, W., Ashburner, J., Trujillo-Barreto, N., Friston, K., 2007. Diffusion-based spatial priors for imaging. NeuroImage 38 (4–3), 677.
Harrison, L.M., Penny, W., Flandin, G., Ruff, C.C., Weiskopf, N., Friston, K.J., 2008. Graph-partitioned spatial priors for functional magnetic resonance images. NeuroImage 43 (4), 694–707.
Hermosillo, G., Chefd'Hotel, C., Faugeras, O., 2002. Variational methods for multimodal image matching. Int. J. Comput. Vis. 50 (3), 329–343.
Iglesias, J.E., Sabuncu, M.R., Van Leemput, K., 2013. Improved inference in Bayesian segmentation using monte carlo sampling: application to hippocampal subfield volumetry. Med. Image Anal. 17 (7), 766–778.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.
Lester, H., Arridge, S., Jansons, K., Lemieux, L., Hajnal, J., Oatridge, A., 1999. Non-linear registration with the variable viscosity fluid algorithm. In: Kuba, A., Samal, M., Todd-Pokropek, A. (Eds.), Proceedings of International Conference on Information Processing in Medical Imaging, IPMI. In: Volume 1613 of Lecture Notes in Computer Science, pp. 238–251.
Marsland, S., Twining, C.J., Taylor, C.J., 2008. A minimum description length objective function for groupwise non-rigid image registration. Image Vis. Comput. 26 (3), 333–346.
Miller, M.I., Christensen, G.E., Amit, Y., Grenander, U., 1993. Mathematical textbook of deformable neuroanatomies. Proc. Natl. Acad. Sci. USA 90 (24), 11944–11948.
Papież, B.W., Heinrich, M.P., Risser, L., Schnabel, J.A., 2013. Complex lung motion estimation via adaptive bilateral filtering of the deformation field. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 25–32.
Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24 (2), 350–362.
Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning, vol. 1. MIT Press, Cambridge, MA.
Risholm, P., Balter, J., Wells, W., 2011a. Estimation of delivered dose in radiotherapy: the influence of registration uncertainty. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 548–555.
Risholm, P., Ross, J., Washko, G., Wells, W., 2011b. Probabilistic elastography: estimating lung elasticity. In: Information Processing in Medical Imaging, vol. 6801. Springer, pp. 699–710.
Risholm, P., Janoos, F., Norton, I., Golby, A.J., Wells III, W., 2013. Bayesian characterization of uncertainty in intra-subject non-rigid registration. Med. Image Anal 17 (5), 538–555.
Risholm, P., Pieper, S., Samset, E., Wells, W., 2010a. Summarizing and visualizing uncertainty in non-rigid registration. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 6362, pp. 554–561.
Risholm, P., Samset, E., Wells, W., 2010b. Bayesian estimation of deformation and elastic parameters in non-rigid registration. In: Proceedings of Workshop on Biomedical Image Registration. Springer, pp. 104–115.
Roberts, S.J., Penny, W.D., 2002. Variational Bayes for generalized autoregressive models. IEEE Trans. Signal Process. 50 (9), 2245–2257.
Rohde, G.K., Aldroubi, A., Dawant, B.M., 2003. The adaptive bases algorithm for intensity-based nonrigid image registration. IEEE Trans. Med. Imaging 22 (11), 1470–1479.
Rueckert, D., Sonoda, L., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging 18 (8), 712–721.
Schmah, T., Risser, L., Vialard, F., 2013. Left-invariant metrics for diffeomorphic image registration with spatially-varying regularisation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 203–210.
Schnabel, J., Rueckert, D., Quist, M., Blackall, J., Castellano-Smith, A., Hartkens, T., Penney, G., Hall, W., Liu, H., Truwit, C., 2001. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 573–581.
Simpson, I.J.A., Schnabel, J.A., Andersson, J.L.R., Groves, A.R., Woolrich, M.W., 2012a. A probabilistic non-rigid registration framework using local noise estimates. In: Proceedings of IEEE International Symposium on Biomedical Imaging 2012, pp. 688–691.
Simpson, I.J.A., Schnabel, J.A., Groves, A.R., Andersson, J.L.R., Woolrich, M.W., 2012b. Probabilistic inference of regularisation in non-rigid registration. NeuroImage 59, 2438–2451.
Simpson, I.J.A., Woolrich, M.W., Andersson, J.L.R., Groves, A.R., Schnabel, J.A., 2013a. Ensemble learning incorporating uncertain registration. IEEE Trans. Med. Imaging 32 (4), 748–756.
Simpson, I.J.A., Woolrich, M.W., Cardoso, M.J., Cash, D.M., Modat, M., Schnabel, J.A., Ourselin, S., 2013b. A Bayesian approach for spatially adaptive regularisation in non-rigid registration. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 10–18.
Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: a survey. IEEE Trans. Med. Imaging 32 (7), 1153–1190.
Staring, M., Klein, S., Pluim, J.P.W., 2007. Nonrigid registration with tissue-dependent filtering of the deformation field. Phys. Med. Biol. 52, 6879.
Tang, L., Hamarneh, G., Abugharbieh, R., 2010. Reliability-driven, spatially-adaptive regularization for deformable registration. In: Proceedings of International Workshop on Biomedical Image Registration, pp. 173–185.
Van Leemput, K., 2009. Encoding probabilistic brain atlases using Bayesian inference. IEEE Trans. Med. Imaging 28 (6), 822–837.
Xu, H., Thirion, B., Allassonnière, S., 2014. Probabilistic atlas and geometric variability estimation to drive tissue segmentation. Stat. Med. 33 (20), 3576–3599.
Zhang, M., Fletcher, P.T., 2014. Bayesian principal geodesic analysis in diffeomorphic image registration. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2014. Springer, pp. 121–128.
Zhang, M., Singh, N., Fletcher, P.T., 2013. Bayesian estimation of regularization and atlas building in diffeomorphic image registration. In: Proceedings of the International Conference on Information Processing in Medical Imaging. Springer, pp. 37–48.