Probabilistic inference of regularisation in non-rigid registration

Ivor J.A. Simpson a,b,⁎, Julia A. Schnabel a, Adrian R. Groves b, Jesper L.R. Andersson b, Mark W. Woolrich b,c

a Institute of Biomedical Engineering, Department of Engineering Science (IBME), Old Road Campus Research Building, University of Oxford, Headington, Oxford OX3 7DQ, UK
b Oxford Centre for Functional MRI of the Brain (FMRIB), University of Oxford, Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK
c Oxford Centre for Human Brain Activity (OHBA), University Department of Psychiatry, Warneford Hospital, Oxford OX3 7JX, UK

ARTICLE INFO
Article history:
Received 18 April 2011
Revised 26 August 2011
Accepted 2 September 2011
Available online 10 September 2011

Keywords:
Registration
Bayesian modelling
Regularisation

ABSTRACT

A long-standing issue in non-rigid image registration is the choice of the level of regularisation. Regularisation is necessary to preserve the smoothness of the registration and penalise against unnecessary complexity. The vast majority of existing registration methods use a fixed level of regularisation, which is typically hand-tuned by a user to provide “nice” results. However, the optimal level of regularisation will depend on the data which is being processed: lower signal-to-noise ratios require higher regularisation to avoid registering image noise as well as features, and different pairs of images require registrations of varying complexity depending on their anatomical similarity. In this paper we present a probabilistic registration framework that infers the level of regularisation from the data. An additional benefit of this proposed probabilistic framework is that estimates of the registration uncertainty are obtained. This framework has been implemented using a free-form deformation transformation model, although it would be generically applicable to a range of transformation models. We demonstrate our registration framework on the application of inter-subject brain registration of healthy control subjects from the NIREP database. In our results we show that our framework appropriately adapts the level of regularisation in the presence of noise, and that inferring regularisation on an individual basis leads to a reduction in model over-fitting as measured by image folding while providing a similar level of overlap.

© 2011 Elsevier Inc. All rights reserved.

Introduction

Medical image registration is an important stage in scientific and clinical, group and longitudinal studies. It provides an estimate of the mapping between one image and another. In order to maximise anatomical or functional correspondence between images, non-rigid registration methods provide a mechanism for high-resolution anatomical alignment (Crum et al., 2004). These algorithms allow flexible and localised mappings between images. An issue present in all approaches to non-rigid registration is how to regularise the inferred model parameters. In the field of non-rigid registration, regularisation is commonly used to provide a penalty against rough deformations, to ensure that the estimated transformations are spatially smooth. In some regularisation approaches, this has the additional effect of penalising the inference of complex mappings.

It is a necessary, but not sufficient condition that the transformation is spatially smooth to maintain the topology of the original image after transformation. The preservation of topology encourages spatially adjacent features in the original image to remain adjacent in the transformed image. It is also appropriate to penalise the complexity of a registration to ensure the plausibility of a mapping.

Purely maximising a similarity measure can produce very large, complicated and noisy deformations as there is no restriction on the complexity of the mapping that is required to improve the model fit (Ashburner and Friston, 1999). This approach of penalising the path length, or deviation from the identity transformation of the inferred mapping, is used in several recent diffeomorphic works on registration (Ashburner, 2007; Ashburner and Friston, 2011; Avants et al., 2008; Beg et al., 2005), and it is clear that the smoothest, shortest mapping which leads to an equivalent model fit is preferable. When using a small deformation framework, such as a free-form deformation (FFD) model (Rueckert et al., 1999) regularisation is also used to reduce any folding of the image which may occur in complex or noisy transformations.

Regularisation often takes the form of membrane, or thin-plate spline bending energy. These simple models penalise deviation from the identity transformation in the second derivative, or bending in the transformation, respectively. In current approaches, these models have a fixed regularisation coefficient which controls the strength of the regularisation. Regularisation parameter values have traditionally been selected using a trial and improvement strategy, where a user finds an appropriate set of parameters which provide qualitatively reasonable results over a specific set of data. As multi-resolution schemes are commonly utilised in non-rigid registration, this would require a user to hand-tune several parameters. Alternatively, regularisation parameters could be selected by testing a range of values and assessing registration performance according to some external metric, such as...
segmentation accuracy (Yeo et al., 2010) which requires manually labelled representative data to train on. However it is derived, a fixed level of regularisation makes the assumption that all data require a similar level of regularisation, whereas the optimal level of regularisation will have a dependence on the data presented to it. For example, it would seem apparent that the anatomy of a particular individual would be more similar to some than others. Therefore, a “one size fits all” approach to penalising the complexity in registration will naturally lead to over- or under-constraining the transformation in some circumstances. Furthermore, higher regularisation may be required when there is a low signal-to-noise ratio (SNR) to constrain the optimisation against noise.

In this work we propose a novel, principled approach for inferring the regularisation parameter required for non-rigid registration in a data-driven way. This is achieved by modelling the regularisation parameter within a hierarchical Bayesian model. This adaptivity in the registration approach allows flexible treatment for different data and multi-resolution optimisation schemes without necessitating any hand-tuning of the regularisation.

In Methods we describe our novel probabilistic framework for non-rigid image registration with an inferred level of spatial regularisation, and demonstrate its application to inter-subject registration of MR images of the human brain. In the following section we describe our novel probabilistic model which is used to drive the registration process and how we can introduce prior information onto the model parameters to provide regularisation. We demonstrate how Variational Bayes can be used to define approximate posterior distributions for our model parameters. As this is a generic framework, any parametrisable transformation model can be used with this probabilistic inference scheme, but for demonstration purposes we implement this framework using a free-form deformation transformation model. We demonstrate that inferring individual regularisation parameters for inter-subject brain registration provides adaptivity to a wide range of SNRs. We also show that individual adaptivity in regularisation yields similarly accurate registrations compared to fixed regularisation, with less image folding as the transformation is more appropriately constrained. Finally, we illustrate the registration uncertainty which arises naturally from this probabilistic framework.

Methods

The process of image registration can be described probabilistically by using a generative model. The majority of generative models for registration use an image similarity term based on the sum-of-squared differences (SSD), which has been previously demonstrated as being appropriate to single-modal brain registration (Ashburner and Friston, 1999). This simple model can be improved upon by explicitly modelling a spatially varying non-linear intensity mapping between images (Andersson et al., 2007). Alternative models have also been proposed which permit different image similarity measures, for example the minimisation of the entropy of the joint image distribution (Zöllei et al., 2007). However, in our work we choose to use a Gaussian likelihood due to its simple and efficient formulation.

It is assumed that the target image data $y$ can be generated from a source image $x$, when it is deformed using some transformation model, where $t(x, w)$ is the transformed source image, and where $w$ parameterises the transformation. The specific form of $t$ used in this implementation is a free-form deformation (FFD), where $w$ is the set of control point displacements in each direction; the rationale for this choice is discussed in Transformation model.

This model will contain some residual error throughout the registration process, either due to misalignment of the images, or because of noise in the image formation process. This error needs to be included within the generative model. The model noise is assumed to be independent and identically distributed (i.i.d.) across image voxels. In this case, the noise is assumed to be a normally distributed error $e \sim N(0, \phi^{-1}I)$, where $I$ represents the identity matrix and $\phi$ is a global precision (inverse variance) of the error across the image.

Therefore, the generic generative model for registration can be formulated as:

$$y = t(x, w) + e$$

(1)

Hence, any given voxel, indexed by $i$, has a likelihood of:

$$P(y_i|x, w, \phi) = \left(\frac{\phi}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\phi} y_i^2\right)$$

(2)

From Eq. (2) the log-likelihood of the target image data $y$ given the set of model parameters can be defined:

$$\log P(y|x, w, \phi) = -\frac{1}{2} \log \phi - \frac{\phi}{2} (y - t(x, w))^T(y - t(x, w)) + \text{const}(w, \phi)$$

(3)

where $N_v$ is the number of voxels containing foreground information in either image, which are also present in the overlap region between the two image domains $t(x, w)$ and $y$ and $\text{const}(w, \phi)$ contains all terms which are constant with respect to $w$ and $\phi$.

In order to regularise the registration, prior distributions over the parameters are included. This provides a full probabilistic model which is graphically described in Fig. 1. The model parameters are inferred to find the mapping between images.

Priors

Transformation parameters

The set of transformation parameters $w$ needs to be spatially regularised. Regularisation is used to preserve the topology of the source image, by enforcing spatial covariance in the transformation parameters and penalising the registration complexity such that complex transformations are not inferred without sufficient image information. In this case transformation complexity is a measurement of deviation from the spatial prior, which in the case of an elastic approach to regularisation, is the identity transformation with some covariance form as described below. We propose to incorporate regularisation into this probabilistic framework by assigning an appropriate prior distribution to the transformation parameters. In this

![Fig. 1.](image-url)
The prior knowledge of \( \mathbf{w} \) is described in Eq. (4) where \( \mathbf{A} \) encodes our regularisation as a \( N \times N \) spatial kernel matrix, details of which are described in Regularisation model. \( N \) is the count of all the transformation parameters, in this case we count every control point in all three directions of motion. In the case of FFDs, this is the number of control points in the image for all three deformation directions. \( \lambda \) is the scalar spatial precision parameter which controls the level of regularisation. \( \lambda \) is modelled as an unknown parameter, therefore it can be determined adaptively from the data resulting in an automated approach to regularisation. Where \( \lambda \) has a constant value, this approach to regularisation is seen in other generative approaches to registration, e.g. in DARTEL (Ashburner, 2007) and FNIRT (Andersson et al., 2007). The novelty of the proposed approach lies in the inference of \( \lambda \) based on the data, and the inference of a posterior density function for \( \mathbf{w} \).

**Spatial precision**

As the spatial precision parameter \( \lambda \) is probabilistically modelled, a prior distribution of \( \lambda \) must be specified. The prior on \( \lambda \) is modelled using a Gamma (Ga) distribution.

\[
P(\lambda) = Ga(\lambda; s_0, c_0) = \lambda^{s_0 - 1} \exp(-\frac{\lambda}{c_0}) \Gamma(c_0, s_0)
\]  

Eq. (5) shows the definition of the prior over \( \lambda \), with initial scale \( (s_0) \) and shape \( (c_0) \) parameters. A Gamma distribution allows us to provide a wide, uninformative prior over the possible values of \( \lambda \), in this case we use values of \( s_0 = 10^{10}, c_0 = 10^{-10} \).

**Noise precision**

In order to evaluate the optimal regularisation of our parameters, we need to accurately estimate the level of noise in our model fit. This is because of the inherent trade-off between regularisation and maximising the likelihood. We therefore also infer \( \phi \) during the registration. We assume that the residual image has a global noise level that has zero mean and is independent and identically distributed Gaussian noise for each voxel with variance \( \phi^{-1} \). We model the prior on \( \phi \) as being Gamma distributed:

\[
P(\phi) = Ga(\phi; a_0, b_0) = \phi^{a_0 - 1} \exp(-\frac{\phi}{b_0}) \Gamma(b_0, a_0)
\]  

where \( a_0 \) and \( b_0 \) are the initial scale and shape prior hyper-parameter estimates of the distribution. These are again chosen to give a wide, non-informative prior distribution with \( a_0 = 10^{10}, b_0 = 10^{-10} \).

**Transformation model**

A key feature of any non-rigid registration algorithm is the choice of the transformation model. Many transformation methods have been proposed for use in the field of medical image registration (Holden, 2008). The proposed regularisation solution penalises the complexity of the deformation parameters, rather than just enforcing smoothness. Therefore an elastic registration model, which penalises deviation from the identity transformation, rather than a viscous fluid registration model (Christensen et al., 1996) is required.

Parametric transformation models describe the complete image transformation using a linear combination of basis functions. The selected basis set can have either global (Ashburner and Friston, 1999), or local (Rueckert et al., 1999) support within the image. These simple methods allow the transformation to be fully described by a single displacement field with three directional components; however, this constrains the transformation to only be capable of modelling small deformations, hence these are referred to as small deformation-models. These can be improved by using multi-resolution frameworks to limit image folding (Schnabel et al., 2001). Although these transformation models have difficulty modelling large deformations, they have substantially fewer parameters to optimise than a time-varying (Beg et al., 2005) or stationary (Ashburner, 2007; Vercauteren et al., 2008) velocity field approach. Therefore, a FFD transformation model using B-splines to interpolate between control points (Rueckert et al., 1999) was deemed an appropriate choice for demonstrating the inference of regularisation in non-rigid registration.

**Noise model**

As described at the beginning of Methods, the noise in model fit is approximated to be zero mean, independently and identically distributed Gaussian noise. We illustrate the reasonableness of this approximation in Fig. 2 using histograms of the residual image \( y - f(x, \mathbf{w}) \) from fitted registration models, overlaid with the inferred Gaussian noise distribution. In inter-subject brain registration, misalignment of anatomy, rather than image formation noise will be a large cause of error in model fit. This may add heavier tails to the residual image distribution, although the centre of the distribution is still well approximated as a Gaussian.

Unfortunately, the assumption of spatial independence in the noise model is largely incorrect. There are two primary sources of spatial noise covariance to consider; firstly, image data is often pre-smoothed using a Gaussian filter to increase the SNR of the data and preserve features of a specific scale. This is performed with different full width at half maximum (FWHM) values at different multi-resolution levels. Image smoothing introduces additional covariance in the image data, and therefore also in the noise. More importantly,
the misalignment of tissue during the registration process introduces spatial correlation in the residual image. This is due to regions of tissue types often being spatially contiguous. An example difference image after affine registration is given in Fig. 3, which illustrates the spatial correlation of the model fit due to misalignment. This spatial covariance in e needs to be compensated for, to avoid over-emphasising the noise precision and therefore the relative importance of the likelihood to the spatial prior.

The most direct method to compensate for the spatial covariance of the residual noise (e) is to model spatially smooth noise using a Gaussian process. This would allow an adaptive determination of the noise covariance. Unfortunately, this approach is computationally very demanding. An alternative solution has been proposed with minimal computational overhead. By calculating the number of RESELS (RESolution EllementS), the number of independent signals in the data can be approximated. If this residual noise is assumed as having been smoothed using a Gaussian kernel, the degrees of freedom of the unsmoothed image can be approximated (Worsley et al., 1995). All terms that sum over voxels are weighted by the ratio of the degrees of freedom in the image to the number of voxels in the overlap Nv. This is equivalent to decimating the data, a process which reduces the number of data samples to remove redundancy. However, as decimating the data requires removing voxels which may still contain valuable information, a weighting term is used instead, providing a virtual decimation (VD) (Groves et al., 2011). The VD weighting factor, α, can be calculated using Eq. (7) as given in (Worsley et al., 1995):

$$\alpha = \frac{\text{RESELS}}{N_v} \left(4\log(2)/\pi \right)^{D/2} = \left( \frac{0.9394}{\text{FWHM}_x} \right) \left( \frac{0.9394}{\text{FWHM}_y} \right) \left( \frac{0.9394}{\text{FWHM}_z} \right)$$

where FWHM(x,y,z) is the full width at half maximum (FWHM) of the equivalent Gaussian smoothing kernel in each direction. The smoothing kernel’s FWHM is estimated by assessing the correlation between adjacent voxels in each direction and

$$(\text{FWHM}_l)^2 = -2\log(2)/\log(\text{corr}_l)$$

where corr_l is the correlation between adjacent voxels in the direction l, l={x,y,z}. This adjustment weights all terms that sum over voxels such that we only consider the number of “independent” noise observations. Although this approach is non-Bayesian, it is still determined from the data, and therefore fits the adaptive framework that we wish to consider.

**Regularisation model**

When encoding regularisation as a prior on the deformation parameters, w, it can be encoded using a spatial kernel which is specified in the form of a precision (inverse covariance) matrix. The precision matrix representation is preferred to a covariance matrix as it can describe the regularisation models of interest using a sparse form, unlike the covariance matrix, which in high-resolution registration would be computationally intractable to store. A can encode a variety of priors, depending on the choice of regularisation model. In this work we are concerned with using the thin-plate spline bending energy. For each transformation vector direction G, which provides a directional component of the mapping between x and t(x,w), the bending energy equation is given as:

$$\int_0^1 |Df|^2 \left( \left( \frac{\partial^2 G}{\partial x^2} \right)^2 + \left( \frac{\partial^2 G}{\partial y^2} \right)^2 + \left( \frac{\partial^2 G}{\partial z^2} \right)^2 + 2 \left( \left( \frac{\partial^2 G}{\partial x \partial z} \right)^2 + \left( \frac{\partial^2 G}{\partial y \partial z} \right)^2 \right) \right) \, dx \, dy \, dz$$

where x,y,z refer to locations within the co-ordinate system box, bounded by the origin and X,Y,Z.

This spatial kernel is derived by differentiating G, with respect to the set of transformation parameters which move in a single axis. A pictorial representation of this kernel, as well as membrane, and elastic energy is given in (Ashburner, 2007). As the transformation is wholly defined by the transformation parameters w, the regularisation of these parameters provides regularisation of the transformation. These regularisation models provide a prior distribution on the transformation parameters which either penalise the second derivative of the deviation from the identity transformation, or the bending energy of the transformation parameters. This regularisation model is widely utilised in parameterised registration approaches, such as (Andersson et al., 2007; Ashburner, 2007; Rueckert et al., 1999).

**Optimisation**

We have described our proposed probabilistic generative model for non-rigid registration between two MRI images of the human brain. This generative model, described in Fig. 1, contains a set of unknown parameters which are modelled using probability distributions. These parameters describe the generation of the target image data y using the source image data x. The distributions on the unknown parameters provide a probabilistic description of the image deformation w and describe the error in model fit (d).

As prior distributions are specified on the model parameters, the inference of the model parameters must account for this. Bayesian
statistical inference provides the only coherent framework for the adjustment of belief (in the form of probability density functions (PDFs)) in the presence of new information (Cox, 1946). Therefore, the Bayesian framework is the most appropriate method to infer on this model.

Bayesian inference
Bayesian inference provides a principled method to infer on the registration model parameters, given their prior distributions. Prior knowledge is incorporated into inferring the posterior distribution using Bayes’ rule which states:

\[ P(\Theta | D) = \frac{P(D | \Theta) P(\Theta)}{P(D)} \]

(10)

where \( \Theta \) is the set of model parameters and \( D \) is the observable data. The posterior probability of the model parameters \( P(\Theta | D) \) is given in terms of the likelihood of the data given the model parameters \( P(D | \Theta) \), multiplied by the prior probability of the parameters \( P(\Theta) \) normalised by the model evidence \( P(D) \). The evidence is calculated by integrating the likelihood over the set of model parameters, which is often intractable to compute, and for most optimisation methods can be ignored. This gives rise to the following relationship:

\[ P(\Theta | D) \propto P(D | \Theta) P(\Theta) \]

(11)

Full Bayesian model inference can be provided using tools such as Markov Chain Monte Carlo (MCMC), but this would be particularly computationally demanding in this setting due to the large number of parameters. Therefore we resort to an approximate solution, using the mean-field variational Bayesian (VB) methodology for model inference.

Variational Bayes
This work follows the mean-field VB approach for inference of graphical models (Jaakkola, 2001; Jordan et al., 1999), which builds on previous work using the mean-field approximation (Peterson and Anderson, 1987). VB is related to the popular expectation–maximisation (EM) approach (Dempster et al., 1977) and is sometimes referred to as variational EM in the literature. VB allows tractable Bayesian inference of an approximate posterior probability distribution of the model parameters by approximating the posterior parameter distribution \( P(\Theta | D) \) as a simpler, parametric probability distribution function \( q(\Theta) \):

\[ q(\Theta) \approx P(\Theta | D) \]

(12)

The hyper-parameters of \( q(\Theta) \) can then be optimised to fit the distribution.

Mean-field VB uses the mean-field approximation over groups of approximate posterior distributions, by assuming independence between groups. In our case, transformation, noise and regularisation parameter distributions are factorised in the approximate posterior:

\[ P(\mathbf{w}, \phi, \lambda | y) \approx q(\mathbf{w}, \phi, \lambda) = q(\mathbf{w})q(\phi)q(\lambda) \]

(13)

This assumption of independence does not prohibit correlation in the mean of the posteriors, and allows for a tractable solution. As VB is not tractable for arbitrary non-linear forward models, a first order Taylor series approximation is used to provide a linear approximation of the transformation function.

Objective function
The objective function in VB is the negative variational free energy \( \mathcal{F} \), which provides a summary of the current model fit, whilst penalising the deviation of the parameters from their prior distribution. \( \mathcal{F} \) provides a lower bound on the log evidence for the model, which becomes closer to the true evidence as the model fit improves. In the case of our registration model, using a mean-field approximation on our inference, \( \mathcal{F} \) can be written as the sum of four terms:

\[ \mathcal{F} = L_{\text{reg}} - D_{\text{KL}}(q(\mathbf{w}) || p(\mathbf{w})) - D_{\text{KL}}(q(\lambda) || p(\lambda)) - D_{\text{KL}}(q(\phi) || p(\phi)) \]

(14)

where \( L_{\text{reg}} \) is the expectation of the log-likelihood given in Eq. (3), with respect to both the Gamma distribution on \( \phi \) and the Gaussian noise model. \( D_{\text{KL}}(q || p) \) is the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) between the approximate posterior \( q \) and the prior \( p \) for each of the model parameters. The KL terms penalise over-confidence in the parameter estimates and deviation from the prior distribution. Maximisation of \( \mathcal{F} \) is equivalent to minimising the Kullback–Leibler divergence between the approximate posterior distribution \( q(\mathbf{w}, \lambda, \phi) \) and the true posterior \( P(\{\mathbf{w}, \lambda, \phi\} | y) \) which cannot be directly calculated.

Variational calculus is used to derive an analytic iterative update for each hyper-parameter of the approximate posterior distributions to maximise the objective function \( \mathcal{F} \). The definition of \( \mathcal{F} \) for this registration model is provided in Appendix A. For computational efficiency in practice we only calculate a small portion of \( \mathcal{F} \) to measure convergence, which we call \( c \).

Model inference
Using the VB framework, analytic updates can be derived for the approximate posterior distributions of the transformation, regularisation and noise parameters, which seek to maximise \( \mathcal{F} \). We use free-form variational Bayes, in which the form of the factorized posterior distribution is not fixed arbitrarily but is instead determined algebraically by combining the likelihood expression with the prior distribution. The derivation of the updates presented below is provided in Appendix B.

Inference on transformation parameters
The approximate posterior distribution form of the transformation parameters, \( \mathbf{w} \), follows a multivariate normal distribution:

\[ q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mu, \mathbf{T}^{-1}) \]

(15)

The VB update rules allow for the derivation of the updates for the mean and covariance of the transformation parameters, which fully define the approximate posterior distribution of \( \mathbf{w} \):

\[ \mathbf{T} = \begin{bmatrix} \alpha \phi^T J + \bar{\lambda} \end{bmatrix} \]

(16)

\[ \mathbf{T}_\mu_{\text{new}} = \begin{bmatrix} \alpha \phi^T (J \mu_{\text{old}} + k) \end{bmatrix} \]

(17)

where \( J \) is the \( N_x \times N_x \) matrix of first order partial derivatives of the transformation parameters with respect to the transformed image \( \mathbf{t} (\mathbf{x}, \mu_{\text{old}}) \), centred about the previous estimate of the mean held in the \( N_x \times 1 \) vector \( \mu_{\text{old}} \). \( k \) is a length \( N_x \times 1 \) vector representing the residual image \( y - \mathbf{t}(\mathbf{x}, \mathbf{w}) \). \( \mu_{\text{new}} \) is a \( N_x \times 1 \) vector describing the current estimated transformation parameters, and is dependent on the old estimated values. The approximate posterior precision matrix of the set of transformation parameters is given by the \( N_x \times N_x \) matrix \( \mathbf{T} \). \( \bar{\lambda} \) is the expectation of the posterior spatial precision distribution and \( \phi \) is the expectation of the estimated noise precision. As both \( \mu_{\text{new}} \) and \( J \) depend on the current parameters, \( \mu_{\text{old}} \), these updates need to be iteratively applied until convergence of \( \mu \) is achieved.

These updates are equivalent to those in non-linear least-squares (NLLS), and for fixed values of \( \phi \) and \( \lambda \) this method is
equivalent to standard NLLS approaches to registration such as FNIRT (Andersson et al., 2007).

Inference on regularisation parameters

The Variational Bayesian methodology can be used to provide updates on the posterior distribution of the regularisation control parameter $\lambda$. The approximate posterior distribution of $\lambda$ is Gamma distributed, $q(\lambda) = \mathcal{Ga}(\lambda; s, c)$. The hyper-parameter updates are as follows:

$$c = c_0 + \frac{N_f}{2}$$

$$\frac{1}{s} = \frac{1}{s_0} + \frac{1}{2} (\text{Tr}(\mathbf{T}^{-1}\mathbf{A}) + \mu^T \mathbf{A} \mu)$$

where $\text{Tr}$ refers to the matrix trace operation. The expectation of the approximate posterior distribution over $\lambda$, is given as $\lambda = sc$.

Inference on noise parameters

The approximate posterior noise parameter distribution $\phi$ is Gamma distributed, $q(\phi) = \mathcal{Ga}(\phi; a, b)$, with hyper-parameter updates given by:

$$b = b_0 + \frac{N_f c a}{2}$$

$$\frac{1}{a} = \frac{1}{a_0} + \frac{1}{2} \phi (\mathbf{k}^T \mathbf{k} + \text{Tr}(\mathbf{T}^{-1} \mathbf{J} \mathbf{J}^T))$$

As described in the Noise model section, $N_f$ is scaled by the virtual decimation factor, $a$, such that it measures the number of independent noise voxels. This is required to prevent over-emphasising $\phi$. The expectation of the approximate posterior distribution over $\phi$, is given as $\phi = ab$.

Implementation

Approximations

From a practical perspective, the update terms required for both the noise (Eq. (21)) and the smoothness (Eq. (19)) precision contain terms which are computationally unfeasible to calculate. Specifically, the difficult term to compute is the inverse of the posterior precision matrix of the transformation parameters, $\mathbf{T}^{-1}$. As $\mathbf{T}$ is a large sparse matrix, calculating the inverse is computationally very intensive, and would require a very large amount of memory. Therefore, only for the updates in Eqs. (21) and (19), an approximation to the posterior covariance matrix is made, which assumes that the control point at each location is independent of its neighbours and only has cross-directional covariance. This allows a sparse inverse approximation which can be rapidly calculated and provides a sufficiently accurate estimation of $\phi$ and $\lambda$, with a slight tendency to over-weight $\lambda$ and under-estimate $\phi$. The accuracy of this approximation is described in Appendix C. For calculating the update on $\mu$ in Eq. (17), $\mathbf{T}^{-1}$ is not necessary to be explicitly calculated, as approximate methods can be used on the full precision matrix $\mathbf{T}$ to find $\mu_{\text{new}}$; in this implementation a conjugate gradient method was used.

Although our updates seek to maximise the negative variational free energy, we choose not to calculate $\mathbf{F}$ in full for measuring the convergence of each update. This is because of the large computational expense of calculating the log determinant of very large matrices. Instead we measure $\mathcal{C} = \alpha \phi |k|^2 + \lambda \mu^T \mu$ for convergence after each update for $\mu$.

Implementation

The 3D implementation of this registration model was incorporated into the FMRIB Non-linear Image Registration Tool (FNIRT) (Andersson et al., 2007). FNIRT is a non-linear least-squares (NLLS) implementation of a FFD B-spline model as popularised by (Rueckert et al., 1999). FNIRT uses MAP inference with Gauss-Newton optimisation, using Levenberg–Marquardt to deal with regions of non-linearity. It is regularised using a fixed-parameter bending energy prior, with a simple noise model which depends on the SSD of the entire image residual, including background regions. FNIRT was chosen as a basis for implementation of the described model due to its similar formulation to our approach. Both are generative models, solved in a NLLS framework. FNIRT also has a highly efficient mechanism for calculating the Hessian matrix, which would otherwise be computationally expensive. However, it must be stressed that this work describes a generic framework and is not restricted to application in FNIRT. It could be implemented in any other generative model regularised using an elastic prior on the transformation parameters, including the diffeomorphic by design approaches of a stationary velocity field (Ashburner, 2007) or using geodesic shooting (Ashburner and Friston, 2011), which may have some advantages in mapping larger deformations. An algorithmic summary of the proposed method is presented in Algorithm 1.

Algorithm 1. Pseudo-code description of the VB registration algorithm:

```plaintext
\begin{algorithm}
  \caption{VB registration algorithm}
  \label{alg:vb}
  \begin{algorithmic}
    \Statex \textbf{function} VBRegistration($\mathbf{x}$)
    \Statex \textbf{input} \textbf{target} $\mathbf{y}$, \textbf{initial} transformation $\mathbf{t}$, \textbf{accuracy} $\epsilon$
    \Statex \textbf{output} transformed image $\mathbf{y}'$
    \Statex $\mathbf{u}_0 := \mathbf{t}$
    \Statex $\mathbf{u}_i := \mathbf{u}_{i-1}$
    \Statex \Repeat
    \Statex \textbf{foreach} voxel $v$
    \Statex \State \textbf{calculate} $\mathcal{C}(v, \mathbf{u})$
    \Statex \State \textbf{calculate} $\mathcal{L}(v, \mathbf{u})$
    \Statex \State $\mathbf{u}_{i+1} := \text{calculate}(\mathbf{u}_i, \mathbf{u})$
    \Statex \Until \textbf{converged}
    \Return $\mathbf{u}_{\text{new}}$
  \end{algorithmic}
\end{algorithm}
```

Results

Experiments

To evaluate our proposed method we first examine the variability in inferred values of $\lambda$ across a range of signal-to-noise ratio (SNR), resolution levels, and between individuals. We compare our proposed method to the original FNIRT implementation using structural overlap measurements, transformation complexity, and the level of image folding of the transformation. We also compare the individual inference of $\lambda$ in each registration, as opposed to fixing it based on the average of a set of registrations. Finally, an illustration of the probabilistic confidence output of the proposed framework is provided.

Materials

All of the experiments conducted used data available in the Non-rigid Image Registration Evaluation Project (NIREP) (Christensen et
NIREP contains a set of 16 3D T1 weighted MR images of the human brain, taken from 16 healthy subjects, 8 male (mean age 32.5 years, std. dev. 8.4) and 8 female (mean age 29.8 years, std. dev. 5.8). These images have a high SNR and have 32 accurately delineated cortical labels. These images have been pre-processed by correcting for bias fields and extracting the brains. Prior to our experiments we affinely align each image to the MNI152 atlas using FLIRT (Jenkinson and Smith, 2001). Each of our individual non-rigid registrations is then initialised with a rigid alignment between the atlas aligned images.

Multi-resolution registration scheme

Multi-resolution registration schemes provide an efficient approach to deal with the different scales of deformations required to warp between two subjects. These schemes provide a coarse to fine alignment, with image sub-sampling and wider control point spacing at the coarser resolution levels. In this implementation, the transformation parameters are refined at each multi-resolution level, such that a single set of transformations parameters are estimated for the entire deformation. The images are smoothed prior to sub-sampling, which is both to improve the SNR of the image data and to provide a scale-space representation of the image data which only preserves image features of a specific scale (Witkin, 1987). As different scales of image features are visible at different multi-resolution levels, and each level will have varying residual magnitudes and degrees of freedom which greatly affect the model fit, they will commonly require different levels of regularisation to provide an optimal balance against the noise precision.

In our comparison with FNIRT we utilise their standard multi-resolution scheme which is recommended for general use in inter-subject structural brain registration. As FNIRT only provides a relatively coarse registration with a control point spacing of 10 mm by default, we extrapolate these default regularisation parameters to provide a more accurate 5 mm control point spacing. We choose both a highly regularised, and a lower regularised 5 mm configuration for comparison. The multi-resolution scheme has 4 levels of sub-sampling resolutions and control point spacings. At each of these levels, there are two separate sub-levels with differing amounts of regularisation and image pre-smoothing. As discussed in Variability in regularisation across signal-to-noise ratios, we also propose a different pre-smoothing scheme. The details of the extrapolated regularisation multi-resolution scheme are given in Table 1.

Variability in inferred regularisation

Variability in regularisation across signal-to-noise ratios

Theoretically, more regularisation should be used in situations where the image information is of lower quality, in order to avoid the registration becoming under-constrained. Therefore, we choose to measure the variability in the inferred level of regularisation over a range of SNRs. This is achieved by adding Gaussian noise to the source image, which had an original SNR of 45, to produce a set of images with a range of SNR between 10 (20 dB) and 45 (33 dB). The SNR of the target was estimated at 52 (34 dB). Each noisy image is registered to the unmodified target image with the proposed framework. As the anatomy of both images remains the same, the difference in inferred regularisation between the registrations will only depend upon the SNR of the source image. We chose to add noise to the source image as this mimics the common spatial normalisation procedure, where collected data is registered to some common template. This template is often an average of many subjects, and hence has little noise. Adding noise to the target image instead may lead to a slightly different behaviour, as gradients are calculated based on the source image, and noise in the source image is smoothed by the interpolation used to warp the source image.

20 noisy images were sampled for each SNR level to provide error bounds on our estimates. The details of the pre-smoothing scheme are likely to affect the resulting registration. Image smoothing increases the SNR of the data, but preserves fewer small features. Therefore, we may expect that pre-smoothing using a larger kernel will result in smoother deformation fields, even if a constant level of regularisation were used. Furthermore, there will exist a dependence of the necessary regularisation on the level of pre-smoothing. Finding an optimal pre-smoothing is not the objective of this paper, and it would be likely to vary between applications. We choose to test three schemes; the original FNIRT scheme, a scheme employing no pre-smoothing and a moderate sized blurring scheme. The moderate pre-smoothing scheme was proposed as an intermediate between the higher pre-smoothing used by FNIRT and none. Details of the pre-smoothing schemes are given in Table 1.

The results of this experiment are shown in Fig. 4. Where no pre-smoothing is used, we see an exponential trend to infer higher spatial precision at lower SNR, heavily constraining the registration in these cases. Using the FNIRT pre-smoothing scheme shows a more linear variation in λ with SNR, and we also see that λ remains substantially higher at high SNR compared to the other two schemes. This is because the substantial smoothing of the images removes some of the

<table>
<thead>
<tr>
<th>Multi-resolution level</th>
<th>Image resolution (mm)</th>
<th>Control point spacing (mm)</th>
<th>FNIRT pre-smoothing FWHM (mm)</th>
<th>Moderate pre-smoothing FWHM (mm)</th>
<th>Fixed FNIRT λ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>8</td>
<td>40</td>
<td>8</td>
<td>8</td>
<td>300</td>
</tr>
<tr>
<td>1b</td>
<td>8</td>
<td>40</td>
<td>6</td>
<td>4</td>
<td>150</td>
</tr>
<tr>
<td>2a</td>
<td>4</td>
<td>20</td>
<td>5</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>2b</td>
<td>4</td>
<td>20</td>
<td>4.5</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>3a</td>
<td>2</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>3b</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>4a</td>
<td>1</td>
<td>5</td>
<td>1.5</td>
<td>1</td>
<td>40/20</td>
</tr>
<tr>
<td>4b</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0.5</td>
<td>30/10</td>
</tr>
</tbody>
</table>

Fig. 4. Plot showing the inferred spatial precision (λ), at the finest multi-resolution level when registering a single pair of images taken from the NIREP database. Independent additive Gaussian noise was added to the source image at a range of signal-to-noise ratios between 10 (20 dB) and 45 (33 dB). The error bars shown in the plot describe the standard deviation of these inferred values across twenty separate registrations with different instances of random noise added to the source images. These images were registered under three different image pre-smoothing schemes, the standard FNIRT scheme, a moderate pre-smoothing scheme, both of which are described in Table 1, and a scheme with no image pre-smoothing. At low SNR and without pre-smoothing the data, we see very high regularisation which decreases exponentially with improved SNR which overly constrains the registration where noise is present. Using the FNIRT pre-smoothing scheme yields a more linear trend in regularisation with SNR, but only produces overly smooth deformations, even at high SNR. The moderate pre-smoothing scheme provides a balance between the two schemes, with lower regularisation at high SNR and a linear relationship between SNR and λ.
image features required to drive a high-resolution registration. The moderate pre-smoothing scheme provides a trade-off between the two schemes, with a linear decrease in $\lambda$ with SNR, and a reasonably low $\lambda$ at high SNR, permitting accurate registration. We therefore choose to use the moderate pre-smoothing scheme with the proposed method in our experiments.

Variability in regularisation across subjects and multi-resolution levels

Each level of the multi-resolution scheme will require a different level of regularisation depending on the control point spacing, the image pre-smoothing and the image sub-sampling, as these factors affect either the transformation model, or the data to be registered. Additionally, each pair of individual brain images will require a different regularisation level to provide the mapping between them, which will depend on both the SNR of the source and target images, and the anatomical differences.

In this experiment we register all 16 subjects in the NIREP database to one another, resulting in a total of 240 registrations. Each of the images in NIREP was acquired using the same protocol, hence their SNR should be similar. Therefore we can assume that the majority of variability in inferred regularisation is due to anatomical differences. The resulting inferred $\lambda$ values, in terms of their FNIRT equivalent values, are shown in Fig. 5. The FNIRT equivalent values are calculated by calculating the FNIRT $\lambda$ value which would be given for the ratio of the inferred spatial precision over noise precision.

Fig. 5 illustrates the variability in $\lambda$ across subjects at each multi-resolution level and allows comparison with the original FNIRT configuration. As can be seen, $\lambda$ has a wide variability between subjects at each stage of the registration. This variability in inferred regularisation across pairwise examples is intuitive, as the complexity of any deformation necessary to map between two individual brains is likely to vary widely depending on the subjects presented. This is further demonstrated in the next section, which shows a wide variability in transformation complexity across registrations. There appears to be no single $\lambda$ which is inferred and would provide an optimal level of regularisation. The level of $\lambda$ will depend not only SNR, but also on the structural differences between individual subjects and therefore should be individually selected.

Comparison against FNIRT

It is appropriate to compare the performance of the proposed framework against the closely related method FNIRT. Each of the 16 NIREP subjects is again registered to every other subject giving a total of 240 registrations. The proposed method was compared against FNIRT with a high and low level of regularisation at the highest resolution level. We compare these methods in terms of transformation complexity, image folding and overlap of segmented regions.

Fig. 6(a) shows the level of complexity across all 240 registration for each method. As can be seen, for fixed levels of regularisation the bending energy forms a tight distribution. When the level of regularisation is inferred, a much larger spread of transformation complexity is obtained...
with a similar mean complexity compared to the higher regularised FNIRT scheme. Where FNIRT uses little regularisation at the finest level, a substantially higher transformation complexity is obtained than with other settings.

The percentage of folded voxels in the transformation can be considered a surrogate measure of model over-fitting, indicating how under-constrained (under-regularised) the registration is, as shown in Fig. 6(b). Here, FNIRT with low regularisation leads to a large amount of image folding. FNIRT with higher fixed regularisation still has a mode of 40% more folded voxels than the proposed framework. Therefore, we may conclude that fixed regularisation methods over-fit the data in some cases, by providing a more complicated transformation without sufficient information to justify this complexity.

Fig. 7 shows the average overlap of the 32 segmented cortical regions over the 240 registrations for each method measured using the Dice metric (Dice, 1945), which is defined as:

$$D = \frac{2|A \cap B|}{|A| + |B|}$$

(22)

where $A$ represents the propagated segmentation labels, and $B$ the target label segmentation. As can be seen, after the affine registration using FLIRT, the average overlap of these regions is very low. FNIRT using a low fixed level of regularisation provides the highest level of overlap, at the expense of much increased complexity and image folding as described previously. With a higher level of regularisation, FNIRT suffers from more negative Jacobians and achieves a slightly lower average overlap than our inferred method. Our method shows a large improvement in overlap from initial alignment using FLIRT and provides a more flexible approach to registration which results in a better constrained transformation.

**Comparison with average of inferred regularisation values**

In order to illustrate whether regularisation needs to be inferred on an individual basis, or to establish whether this framework should be used to infer a good set of fixed parameter values, we re-run all 240 registrations using our framework, but without inferring the value of $\lambda$ and instead using the average inferred $\lambda$ across the 240 registrations. This is the right-most plot in Figs. 6(a), (b), and 7. As could be expected, some of the variability in the transformation complexity is lost as a result of each registration using the same amount of regularisation. From Fig. 7 we can see that the overlap is very slightly improved with this method over individual adaptivity, but this comes at the expense of a 25% increase in the mode of the number of negative Jacobians. In Fig. 8, we show an example registration where $\lambda$ is inferred, or is fixed at the mean of the inferred $\lambda$ values. In this example we show the propagation of a cortical label (right post-central gyrus) which is better aligned where a higher level of regularisation is inferred for all of the multi-resolution levels except for the finest. In this case inferring $\lambda$ results in a 25% decrease in folding, and leads to a small increase in average overlap. Based on these results, we can summarise that even for a similar group of healthy individuals, without individually inferred regularisation, the registration will be improperly constrained in more cases, resulting in less smooth mappings which have little to no improvement in overlap. An alternative approach would have been to use an informative prior distribution for $\lambda$, where $S_0$ and $C_0$ were derived from the individual runs.

![Fig. 6.](image_url) Violin plots illustrating the probability density functions of the bending energy, and percentage of voxels with negative Jacobians across the different non-rigid methods. Violin plots are interpreted as a superposition of a kernel density estimation plot (in blue), on top of a boxplot (black box, and line). The data in these plots comes from the 240 pairwise registrations of the 16 individuals in the NIREP database. The plot on the left shows the level of bending energy, which measures transformation complexity, across the different non-rigid methods. The fixed regularisation FNIRT approaches have a very tight distribution, implying a similar level of complexity used to map between all pairs of individuals. Conversely, where $\lambda$ is inferred from the data, the level of bending energy takes a much wider range of values, as the transformation complexity is data dependent. With the VB approach with $\lambda$ fixed to the mean of the inferred distribution, we still see a fairly wide distribution due to the inferred noise level. Plot (b) shows the percentage of voxels with negative Jacobians (image folding) across the different non-rigid methods. The level of folding is substantially higher in FNIRT where $\lambda$ is low at the finest resolution levels, compared to all other methods. Where $\lambda$ is inferred from the data we see a lower mode and median than where a fixed level of regularisation is used.

![Fig. 7.](image_url) A boxplot showing the average overlap scores (Dice) over the 32 segmented cortical regions between the 4 different non-rigid methods and FLIRT from the 240 pairwise registrations of the NIREP database. As can be seen, all the FFD methods outperform the affine registration of FLIRT. FNIRT with low $\lambda$ at the finest levels achieved the highest overlap, at the cost of a lot of transformation complexity and image folding (Fig. 6(a) and (b)). The proposed framework achieves a very similar result in terms of overlap to that of FNIRT with high regularisation, and the VB framework with fixed regularisation, but with a reduction in the level of image folding.
We have demonstrated that our proposed approach provides an appropriate framework for non-rigid registration which adapts to the SNR of the data and the anatomical variability between subjects.

A benefit of such an adaptive approach to regularisation in non-rigid registration is that the regularisation parameters are not required to be derived using a time consuming trial and improvement approach, which will likely need to be re-tuned for different datasets. Regularisation parameter values could be estimated by assessing registration performance on a training set of labelled data according to some quantitative metric. However, this relies on having sufficient labelled training data which has similar properties to any data which will be registered. Allowing the inference of regularisation on an individual basis overcomes the “one size fits all” approach to registration complexity. This is likely to be flawed due to cross-population anatomical variability, which implies that a fixed regularisation solution will be under- or over-constrained in some situations.

In this work we have inferred a wide range of values for \( \lambda \) across different registrations at each multi-resolution level. These values produce equivalent average structural overlap measurements to hand-tuned regularisation methods, but in general result in a better constrained mapping, as indicated by the lower level of image folding. The differences in bending energy across the set of 240 registrations when inferring \( \lambda \) show that varying complexity is required for registering different individuals from a population, even where all the subjects are healthy controls from a reasonably small age range.

There are two key advantages in our model formulation using a principled probabilistic framework; firstly, as our framework is fully probabilistic we have an estimate of the uncertainty of the registration model parameters. Of these parameters, most interestingly we have obtained an estimate of the uncertainty of the transformation parameters in the form of a covariance matrix. This provides both a measure of the spatial uncertainty of each voxel in the transformed source image, and of any registration derived features, for example Jacobian features. An example map illustrating the estimated spatial variance of each direction is given in Fig. 9. This map shows the variance of each transformation parameter, interpolated across the image domain. The variance was accurately estimated by inverting a 5\( ^2 \) set of transformation parameters, in all 3 directions, for each control point. Previous attempts to produce probabilistic output from registration have utilised stochastic variation of control points (Hub et al., 2009), whereas our method explicitly models the variance of each transformation parameter. Other attempts have analysed the Hessian of the cost function under the assumption of normality (Kybic, 2010), however this does not use a
full probabilistic model, whereas our method models the image noise and regularisation, both of which will greatly affect the variance of the method. Risholm et al. (2010a) discuss some possible uses of the probabilistic output from their Bayesian elastic registration tool. However, their estimates of uncertainty are limited due to the ad-hoc regularisation that they use for this particular task. Additionally, their approach is computationally expensive for estimating a high-resolution mapping, as they use MCMC for inference. In contrast, the method we propose provides a tractable and principled approach to estimating both the regularisation weighting, and approximating the level of spatial uncertainty in high-resolution registration. We envisage that this spatial uncertainty information could be utilised in further statistical processing for a wide variety of tasks. In (Simpson et al., 2011) we showed that the mapping uncertainty could be used to estimate an appropriate smoothing kernel to compensate for misregistration. This was shown to increase the ability to classify spatially normalised subjects with Alzheimer’s disease from healthy elderly controls. Further work could either take the form of directly modeling spatial uncertainty at a group statistics level, e.g. in MRI data analysis as demonstrated in (Keller et al., 2008), or of incorporating this uncertainty into registration derived features which we might use for further analysis, such as Jacobian values for tensor based morphometry, which will also be probabilistic.

The second advantage of our formulation is that we have developed a generic Bayesian framework for non-rigid registration which would allow us to incorporate more complex adaptive spatial priors, or noise models, providing a more flexible approach to registration. For example, a more complex spatial prior could be inferred in the same manner, this could allow the inference of multiple $\lambda$ values either varying over space, or for a linear combination of precision components, for example the Lamé parameters for linear elastic regularisation.

This framework could also be implemented with a variety of transformation models, including those which are guaranteed to produce diffeomorphic mappings, for example stationary velocity fields (Ashburner, 2007) or variable velocity fields via geodesic shooting (Ashburner and Friston, 2011). This could lead to some improvements in modelling larger deformations. The VB framework also permits model comparison using the variational free energy $\mathcal{F}$ which could be used, for example, to compare different control point spacings or spatial priors (as long as they are non-singular). However, for the framework presented here there are two restrictive factors: firstly the VD factor $\alpha$, used to compensate for the spatial covariance of the residual image defined in Noise model, would need to be fixed between methods, as this would otherwise alter the data. Additionally, the Taylor series expansion will give different uncertainties at different local minima. Therefore, we decided to not consider model comparison in this paper.

One of the questions which remains unanswered in this paper is how to pre-smooth the images. For this work, we chose a moderate level of image pre-smoothing image to provide accurate registration across a range of SNRs. However, the optimal image smoothing is

![Fig. 9. An example map of the spatial uncertainty of our transformation. The top row shows the transformed source image, the next three rows contain maps showing the standard deviation, for each transformation direction, of the final mapping between images. Notice the lower uncertainty surrounding edges, and higher uncertainty in homogeneous regions.](image-url)
likely to depend on the specific application in question and will be a subject of further investigation.

Conclusions

In this work we have demonstrated that it is possible and computationally tractable to infer regularisation parameters in non-rigid registration as part of a probabilistic framework. This was developed as a generic registration framework, capable of using a range of transformation and regularisation models. We demonstrated its use with a high-resolution free-form deformation transformation model applied to the problem of inter-subject brain registration. We have shown that our method appropriately regularises the registration dependent on the data quality. It also provides adaptivity to individual variability, by penalising against overly complex registrations where there is little support for this from the data. We compared this method to FNIRT, from which this implementation was derived, and found that little support for this from the data. We compared this method to FNIRT, from which this implementation was derived, and found that little support for this from the data. We compared this method to FNIRT, from which this implementation was derived, and found that little support for this from the data.

Acknowledgment

IJAS would like to acknowledge funding from the EPSRC through the Life Sciences Interface Doctoral Training Centre, Oxford, UK.

Appendix A. Variational free energy

Convergence could in principle be monitored by the negative variational free energy $F$. This would be an appropriate choice as it is the measure we are seeking to maximise. As we are using the mean-field approximation for the proposed model, $F$ can be split into 4 separate terms:

$$ F = L_{av} - D_{KL} (q(\lambda)||p(\lambda)) - D_{KL} (q(\phi)||p(\phi)) - D_{KL} (q(w)||p(w)) $$ (A.1)

where $L_{av}$ is the average likelihood of the model parameters, and $D_{KL}$ is the Kullback–Leibler divergence between distributions. The standard Kullback–Leibler divergences for normal and Gamma distributions were given in Roberts and Penny (2002).

The average log-likelihood $L_{av}$ of the model parameters, is the likelihoo given in Eq. (3), where we take the mode of our parameter distributions:

$$ L_{av} = \frac{N_c}{2} \log(\alpha) + \psi(b) - \frac{\phi}{2} \alpha \left( k^T k + \operatorname{Tr}(T^{-1}J J^T) \right) - \frac{N_c}{2} \log(2\pi) $$ (A.2)

where $\psi$ is the di-gamma function.

The KL divergence between the approximate posterior and prior on $\phi$ is given by:

$$ D_{KL} (q(\phi)||p(\phi)) = (b-1) \log(b) - b \log(\Gamma(b)) + \log(\Gamma(b_0)) + b_0 \log a_0 - (b_0-1) \log(b_0) + \frac{ab}{b_0} $$ (A.3)

The KL divergence between the approximate posterior and prior on $\lambda$ is given by:

$$ D_{KL} (q(\lambda)||p(\lambda)) = (c-1) \log(c) - c \log(\Gamma(c)) + \log(\Gamma(c_0)) + c_0 \log(c_0) - (c_0-1) \log(c_0) + \frac{sc}{s_0} $$ (A.4)

The KL divergence between the approximate posterior and prior on $w$ is given by:

$$ D_{KL} (q(w)||p(w)) = -\frac{1}{2} \log \left( \frac{\lambda A^{-1}}{w - A} \right) + \frac{1}{2} \lambda \operatorname{Tr}(T^{-1}A) + \frac{1}{2} \lambda \omega_{Aw} - \frac{N_c}{2} $$ (A.5)

In practise, we do not calculate $F$ in full due to the computational complexity of calculating the log determinant of $T$, but calculate a small portion of it $C = \alpha \phi k^T k + \lambda \mu^T \mu$ to monitor for convergence after updating $\mu$.

Appendix B. Derivation of approximate posterior updates

In order to infer the posterior parameter distributions using VB, we first need to decide on the approximate form of the posterior distribution. By using the mean-field approximation, the transformation, noise and smoothness posterior distributions are assumed to be independent.

$$ P(w, \lambda, \phi, \gamma) \approx q(w, \lambda, \phi) = q(w)q(\lambda)q(\phi) $$ (B.1)

As we are using a local approximation of a linear relationship between the transformation parameters $w$, and the Gaussian likelihood of the image residuals, our approximate posterior distribution on $w$ will be normally distributed, hence $q(w) = \mathcal{N}(w; \mu, T^{-1})$. The approximate log posterior for $w$ is given as:

$$ \log q(w) = -\frac{1}{2} w^T T w + \frac{1}{2} w^T T \mu + \frac{1}{2} \mu^T T w + \log \text{const}(w) $$ (B.2)

where $\log \text{const}(w)$ contains all terms which are constant with respect to $w$. The posterior distribution on $\lambda$ as a conjugate to the prior, is Gamma distributed, and parameterised using shape and scale parameters, $c$ and $s$ respectively. The approximate log posterior is given by:

$$ \log q(\lambda) = -\lambda s + (c-1) \log \lambda + \text{const} \{ \lambda \} $$ (B.3)

Similarly, the posterior distribution on $\phi$ is Gamma distributed, with shape and scale parameters $b$ and $a$. The log posterior is given as:

$$ \log q(\phi) = -\frac{\phi}{a} + (b-1) \log \phi + \text{const} \{ \phi \} $$ (B.4)

The full log posterior of the transformation, noise and smoothness parameters is given as:

$$ L = \log P(y, x, w, \phi, \gamma) + \log P(w) + \log P(\lambda) + \log P(\phi) + \log \text{const} \{ w, \lambda, \phi \} $$ (B.5)

$$ = \alpha \frac{N_c}{2} \log \phi - \frac{\phi}{2} \left( y - T w \right)^T \left( y - T w \right) - \frac{\lambda}{s_0} (c_0-1) \log \lambda - \frac{1}{2} \frac{w^T A w}{\lambda} + \frac{N_c}{2} \log \lambda - \frac{\lambda}{s_0} (c_0-1) \log \lambda + \log \text{const}(w, \lambda, \phi) $$

The log posterior is used to analytically derive our updates for each parameter group, by applying the calculus of variations. When
using mean-field variational Bayesian inference with conjugate, exponential distributions we find analytic updates for the hyper-parameters of each approximate posterior distribution by integrating the log posterior \( \mathcal{L} \) over the other approximate posterior distributions. Updates are found by comparing the coefficients of the approximate log posterior hyper-parameters with those in our marginalised full log posterior.

**Update on \( \mathbf{w} \)**

In order to derive the update for \( \mathbf{w} \), we need to integrate out the approximate posterior distributions of \( \phi \) and \( \lambda \) from our log posterior \( \mathcal{L} \):

\[
\log q(\mathbf{w}) = \int \mathcal{L}(\lambda) q(\phi) d\lambda d\phi + \text{const}(\mathbf{w}) \tag{B.6}
\]

The right hand side of the above equation can be written as:

\[
\int \left( \frac{\partial \mathbf{y}}{\partial \mathbf{w}} - \frac{\lambda_0}{2} \right) (\mathbf{y} - \mathbf{t}(\mathbf{w})) + \frac{\lambda_0}{2} \mathbf{w}^T \mathbf{A} \mathbf{w} + \text{const}(\mathbf{w}) \tag{B.7}
\]

We need to make a linear approximation to the transformation \( \mathbf{t}(\mathbf{w}) \) in order to make this tractable to VB. This can be achieved by taking a first order Taylor series expansion:

\[
\mathbf{t}(\mathbf{w}) \approx \mathbf{t}(\mathbf{w}_0) + \mathbf{J}(\mathbf{w} - \mathbf{w}_0) \tag{B.8}
\]

where \( \mathbf{J} \) is the Jacobian matrix of partial derivatives of each voxel in the image, with respect to each transformation parameter. Now we can rewrite \( \mathbf{y} - \mathbf{t}(\mathbf{w}) \):

\[
\mathbf{y} - \mathbf{t}(\mathbf{w}) = \mathbf{y} - \mathbf{t}(\mathbf{w}_0) - \mathbf{J}(\mathbf{w} - \mathbf{w}_0) \tag{B.9}
\]

where \( \mathbf{k} \) is the vectorisation of \( \mathbf{y} - \mathbf{t}(\mathbf{w}_0) \). We can now rewrite the right hand side of Eq. (B.7), using our Taylor series expansion as:

\[
-\frac{1}{2} \left( \left( \mathbf{k} - \mathbf{J}(\mathbf{w} - \mathbf{w}_0) \right)^T \frac{\partial \phi}{\partial \mathbf{w}} + \mathbf{w}^T \mathbf{A} \mathbf{w} \right) + \text{const}(\mathbf{w}) \tag{B.10}
\]

By comparing the coefficients of \( \mathbf{T} \) and \( \mathbf{\mu} \) between Eq. (B.11) and the posterior distribution for \( \mathbf{w} \) given in Eq. (B.2), the hyper-parameter updates are given as:

\[
\mathbf{T} = \mathbf{J}^T \frac{\partial \phi}{\partial \mathbf{w}} + \Lambda \mathbf{\lambda} \tag{B.12}
\]

\[
\mathbf{T} \mathbf{\mu}_{\text{new}} = \mathbf{J}^T \mathbf{\mu}_{\text{old}} + \mathbf{\lambda} \frac{\partial \phi}{\partial \mathbf{w}} \tag{B.13}
\]

where \( \mathbf{\mu}_{\text{new}} \) and \( \mathbf{\mu}_{\text{old}} \) are the new and old estimates for \( \mathbf{\mu} \) respectively. These updates give us a non-linear least-squares solution for estimating the transformation parameters. If \( a,b,c \) are constant, this is equivalent to current registration approaches such as FNIRT.

**Updates for \( \lambda \)**

To update \( \lambda \), we integrate out the posterior on \( \mathbf{w} \) and \( \phi \) from our full log posterior:

\[
\log q(\lambda) = \int \mathcal{L}(\mathbf{w}, \mathbf{\phi}) q(\mathbf{w}, \phi) d\lambda d\phi + \text{const}(\lambda) \tag{B.14}
\]

The right hand side of the above equation can be written as:

\[
\int \left[ -\frac{\lambda}{2} + \frac{N_c}{2} (\mathbf{w}^T \mathbf{A} \mathbf{w} + \text{const}(\lambda)) \right] \mathcal{N}(\mathbf{w}; \mathbf{\mu}, \mathbf{T}^{-1}) d\mathbf{w} + \text{const}(\lambda)
\]

where \( \mathcal{N}(\mathbf{w}; \mathbf{\mu}, \mathbf{T}^{-1}) \) refers to the matrix trace operation.

By comparing coefficients of \( s \) and \( c \) from Eq. (B.15) and the posterior distribution for \( q(\lambda) \) given in Eq. (B.3), we can see the hyper-parameter updates are given as:

\[
c = c_0 + N_c \frac{a}{2} \tag{B.16}
\]

\[
\frac{1}{s} = \frac{1}{s_0} + \frac{1}{2} \left( \text{Tr} \left( \mathbf{T}^{-1} \mathbf{A} \right) + \mathbf{\mu}^T \Lambda \mathbf{\mu} \right) \tag{B.17}
\]

**Updates for \( \phi \)**

To update \( \phi \), we integrate out the posterior on \( \mathbf{w} \) and \( \lambda \) from our full log posterior:

\[
\log q(\phi) = \int \mathcal{L}(\mathbf{w}) q(\lambda) q(\mathbf{w}) d\lambda d\phi + \text{const}(\phi) \tag{B.18}
\]

The right hand side of the above equation can be written as:

\[
\int \left[ \alpha N_c \frac{\partial \phi}{\partial \mathbf{w}} - \phi \right] (\mathbf{y} - \mathbf{t}(\mathbf{w})) + \frac{\partial \phi}{\partial \mathbf{w}} - \frac{\partial \phi}{\partial \mathbf{w}} \left( \mathbf{k}^T \mathbf{\phi} + \text{Tr} \left( \mathbf{T}^{-1} \mathbf{J}^T \mathbf{J} \right) \right) + \text{const}(\phi)
\]

By comparing coefficients of \( a \) and \( b \) from Eq. (B.19) and the posterior distribution for \( q(\phi) \) given in Eq. (B.4), we can see the hyper-parameter updates are given as:

\[
b = b_0 + N_c \frac{a}{2} \tag{B.20}
\]

\[
\frac{1}{a} = \frac{1}{a_0} + \frac{\alpha}{2} \left( \mathbf{k}^T \mathbf{\phi} + \text{Tr} \left( \mathbf{T}^{-1} \mathbf{J}^T \mathbf{J} \right) \right) \tag{B.21}
\]

**Appendix C. Accuracy of approximate inference of \( \lambda \) and \( \phi \)**

As mentioned in Approximations, for the noise and spatial precision updates in Eqs. (21) and (19), an approximation of control point independence is used. This is because the inverse of the large sparse precision matrix \( \mathbf{T} \) is required, and is computationally impractical to calculate or store. Therefore, we choose to approximate this matrix as having independence between control points, only including the covariance between directions. The effect of this approximation is that when calculating the covariance matrix, the estimated variance will be lower in regions which contain a large amount of information, e.g. edges, and higher in homogeneous regions, which have a large amount of image covariance. This will induce a preference for slightly higher spatial precision, and lower noise precision. In order to test the accuracy of this approximation, a series of 200 2D registrations
was carried out, resolving smooth artificial deformations of varying magnitudes applied to random image slices from the NIREP database. The results presented in Fig. C.10 show that this approximation produces reasonably accurate inference of these two parameters, with a 5–14% bias towards a higher spatial precision and a 3–7% bias to lower noise precision.

### References


