Chapter 7

Discussion

Contents

7.1 Original Contributions

This thesis has described a number of computational tools that can be used for both functional image analysis, and computational neuro-anatomy. These include various registration approaches, and image segmentation. Many of the methods have been incorporated into the SPM99 package, are already used widely by a large number of researchers in the field of functional imaging, and are also increasingly used in studies of morphometry.

A number of original contributions have been made within this thesis. Some of these may be completely original (or independently re-invented), but others simply involve combining different parts of pre-existing methods in a new way. For example, most of the techniques developed in Chapters 2 and 3 are significant improvements to methods previously presented by Friston *et. al.* (1995c). Within Chapter 2, the original contributions were predominantly in Section 2.6, which described a new approach to between modality registration. Although the approach is not as flexible as mutual information registration methods, it generally performs well when the various assumptions are satisfied.

The main novel component of Chapter 3 is the fast algorithm for computing the Hessian matrix (ignoring second derivatives) for the Gauss-Newton optimisation scheme. A simplified version of this same algorithm is also used for estimating intensity modulation fields in Chapter 5. In addition, Friston's original registration scheme (Friston *et al.*, 1995c) was placed within an Empirical Bayesian framework.

Chapter 4 introduced two main original ideas. The first is the use of prior probability distributions that are the same for both the forward and inverse deformations. The more pure version assumes log-normal distributions for length, area and volume changes. Because the computations required for this version would be prohibitively slow, a much faster approximation was also devised. The second main idea involves methods for making deformations internally consistent by simultaneously matching several images at the same time. A method for inverting deformation fields was also included in order to facilitate combining deformations with the inverses of other deformations.

A number of novel ideas can be found in Chapter 5. These include the use of prior probability maps to assist classification, and some aspects of the intensity non-uniformity correction. These include the use of a third-order regularisation scheme, and a rapid algorithm similar to that of Chapter 3.

The first original contribution in Chapter 6 are the evaluations of the applicability of the statistical parametric modelling component of SPM99 to voxel-based morphometry. The deformation-based morphometry section describes the use of standard multivariate statistical methods for comparing deformation fields among groups. Finally, a method of identifying regional shape differences is presented, which is based on applying multi-variate statistical tests to strain tensors.

A few recurring themes are encountered within this thesis. The first of these is about achieving consistency, not only in terms of between different data-sets, but also between different processing steps. A certain amount of modularity may need to be sacrificed in order to achieve optimal consistency between procedures. A second theme is the problem of assigning optimal hyper-parameters (weights) to different components of the various procedures. Although none of the chapters so far have presented any good methods for doing this, a section is included here that may provide a useful framework in which future methods may achieve this end.

7.2 Modularity

Currently, different image processing components are often thought of as distinct modules, whereby one set of processing is completed before another one begins. However, this may not be the best approach. A few examples will now be given that suggest that the whole should be considered greater than the sum of the parts.

Chapter 2, includes a section on between modality image registration. Part of the method includes simultaneously spatially normalising a pair of source images to corresponding templates, while maintaining a rigid body relationship between the source images. This component could be extended to include nonlinear warping in addition to the current zooms and shears. In theory, this should provide a framework for improved rigid registration, as well as better estimates of the deformations that spatially normalise the images. One of the themes of the thesis is about improving internal consistency, of which this is another example, as all the information would be included within the same internally consistent registration model. Another matching criterion such as mutual information could also be recruited at the same time, which should effectively bring additional information to the problem, further enhancing the final result. The more useful information that is included, then the better the parameter estimates should be – providing that the optimal weighting for the different components can be found.

Section 2.6 also suggests another move away from a modular image processing stream. Simultaneous rigid registration and spatial normalisation has just been mentioned, but the section also includes a segmentation component. Segmentation can be performed more efficiently when the prior probability of each voxel belonging to particular tissue classes is known, and in order to achieve this, the images need to be in register with prior probability images. Although the chapter only describes using affine registration to effect this mapping, more accurate results could be expected after nonlinear matching. Conversely, segmented images can be useful for image registration. For example, a number of nonlinear image registration methods are based on matching brain surfaces together, either in 3D (Thompson & Toga, 1996), or in 2D on a flattened cortical surface (see Drury *et. al.* (1999) or Thompson and Toga (1999)). It seems that segmentation and spatial normalisation could be performed most accurately when both done simultaneously.

Many issues relating to morphometry are also applicable to image warping. Image registration methods generally involve some model of how brains should deform. These can generally be thought of as some form of multi-dimensional probability distribution. Similarly, morphometric methods also require models of brain shape variability. It is envisaged that future work on morphometry should develop in concurrence with the methods used for estimating deformations. The parameter distributions imposed upon the deformations by the registration method could be used in the morphometry studies. Similarly, knowledge of the variability of brain shapes obtained from morphometry could be used as *a priori* information for Bayesian image registration methods. Both fields would clearly benefit by having a compact and concise representation of the anatomical variability of brains.

The link between warping methods and morphometric methods may be even tighter. Not only do they both need similar models of shape variability, but they also need the hyper-parameters that describe the amount of variability. These can not be estimated properly by warping a number of images to the same template, followed by simply computing the variance of some features of the resulting deformation fields. This is because the amount of regularisation used by the warping model would greatly influence the estimated result. In order to obtain unbiased estimates of shape variability, this estimation would need to be a component of the warping model. The next section elaborates on this theme.

7.3 Hyper-parameter estimation

An ongoing argument in the field of image registration is something to the effect of "if you use too many parameters in a registration model, then the model will be overfitted", versus "if not enough parameters are used, then the images can not be registered properly". Increasing the relative weighting of the prior potential component effectively reduces the search space of the optimisation problem, so this argument can be rephrased in terms of the relative weights used for the likelihood and prior potential components. This section elaborates on a possible approach that may help to resolve these issues.

This thesis contains many examples of problems where a model is used, but optimal weights for determining the best parameter estimates are not known. These problems occur in situations where the optimised functions consist of several terms, where each term is weighted differently. For example, in Section 2.6, more than one pair of images is registered simultaneously. Optimal results are obtained only when the contribution from each image pair to the cost function, is properly balanced. Another area is in image warping, in which it is necessary to know the optimal balance between prior and likelihood potential (see Sections 3.2.1 and 4.2.2). Also, some method of estimating the *a priori* variability of the non-uniformity is needed to improve the segmentation (see Section 5.2.3). In order to obtain the most accurate results, optimal weighting should be assigned to the different components of the problems.

One method of choosing weights empirically is by using L-curves, whereby a model is fitted using a range of different weights. After all the fits are complete, the logs of one cost function term are plotted against the logs of the other, over the range of different weights. The weighting is deemed optimal for values that fall close to the inflection of the plot. This method is time consuming, and is not practical for complex models where there may be many different terms. Fortunately, there may be more practical solutions to the problem of hyper-parameter estimation, which would allow the unknown weights, in addition to the parameters of interest, to be estimated.

So far, any hyper-parameter estimation in this thesis has been performed in a pragmatic, but rather ad hoc manner. In particular, the estimation of the degrees of freedom in the models has involved slightly questionable methods. There now follows a more rigorous framework for hyper-parameter estimation that could be used for future work. As an illustration, consider the iterative parameter updating scheme from Chapter 3. After rearranging and a slight modification to the notation[1], Eqn. 3.6 can be written as:

$$\mathbf{q}^{(n+1)} = \mathbf{q}^{(n)} - \left(\lambda_1 \mathbf{A_1}^T \mathbf{A_1} + \lambda_2 \mathbf{C_0}^{-1}\right)^{-1} \left(\lambda_1 \mathbf{A_1}^T \mathbf{b_1} + \lambda_2 \mathbf{C_0}^{-1} \left(\mathbf{q}^{(n)} - \mathbf{q_0}\right)\right) \tag{7.1}$$

Obtaining the best balance between minimising the residual squared difference between the images, and maximising some form of deformation smoothness, can be considered as finding some optimum estimates for λ_1 and λ_2. It is possible to do this from the data using an empirical Bayes approach, providing that a few assumptions about the forms of the errors are made. In this example, both the parameters themselves and the residual difference between the image pair are assumed to be independent and identically distributed. An algorithm similar to expectation maximisation (EM) can be used for estimating the hyper-parameters.

EM is an iterative approach that alternates between two steps: the expectation (E) step and the maximisation (M) step (see Chapter 5). The algorithm is initialised by assigning some (non-zero) starting estimates to the hyper-parameters (λ_1 and λ_2). The first E step involves computing the expectation of a set of unobserved parameters while holding the hyper-parameters constant. Within a Bayesian registration scheme, this would involve computing a minimum variance estimate of the parameters ($\mathbf{q}$). This is where this approach deviates from a pure EM algorithm, as a MAP estimate of $\mathbf{q}$ is used instead, as it is much faster to compute. This can be achieved by repeated application of Eqn. 7.1 until convergence is attained.

The M step involves re-estimating the hyper-parameters such that the likelihood of observing the data is maximised, while holding the parameters fixed. The hyper-parameters are computed from the reciprocals of variance estimates derived from the observations and parameters. This step involves estimating the residual variance attributed to the different parts of the model, divided by the appropriate degrees of freedom, computed by:

$$\begin{aligned} p_1 &= m_1 - tr\left(\left(\lambda_1 \mathbf{A_1}^T \mathbf{A_1} + \lambda_2 \mathbf{C_0}^{-1}\right)^{-1} \lambda_1 \mathbf{A_1}^T \mathbf{A_1}\right) \\ p_2 &= m_2 - tr\left(\left(\lambda_1 \mathbf{A_1}^T \mathbf{A_1} + \lambda_2 \mathbf{C_0}^{-1}\right)^{-1} \lambda_2 \mathbf{C_0}^{-1}\right) \end{aligned} \tag{7.2}$$

where m_1 is the number of rows an $\mathbf{A_1}$, and m_2 is the number of columns (also the same as the

[1] $\mathbf{C_0}$, $\mathbf{A}$, $\mathbf{b}$ and σ^2 become $\lambda_2\mathbf{C_0}$, $\mathbf{A_1}$, $\mathbf{b_1}$ and $1/\lambda_1$ respectively.

number of rows or columns of $\mathbf{C_0}^{-1}$. The new estimates of λ_1 and λ_2 are then:

$$\lambda_1 = \frac{p_1}{(\mathbf{b_1} - \mathbf{A_1 q})^T (\mathbf{b_1} - \mathbf{A_1 q})}$$
$$\lambda_2 = \frac{p_2}{(\mathbf{q} - \mathbf{q_0})^T \mathbf{C_0}^{-1} (\mathbf{q} - \mathbf{q_0})} \tag{7.3}$$

This alternating EM procedure is continued until a stopping criterion is reached.

More generally, Eqn. 7.1 can be expressed as:

$$\mathbf{q}^{(n+1)} = \mathbf{q}^{(n)} - \left(\mathbf{A}^T \mathbf{C}^{-1} \mathbf{A}\right) \mathbf{A}^T \mathbf{C}^{-1} \mathbf{b} \tag{7.4}$$

where $\mathbf{A} \equiv \begin{bmatrix} \mathbf{A} \\ \mathbf{I} \end{bmatrix}$, $\mathbf{b} \equiv \begin{bmatrix} \mathbf{b_1} \\ \mathbf{q}^{(n)} - \mathbf{q_0} \end{bmatrix}$ and $\mathbf{C} \equiv \begin{bmatrix} \lambda_1^{-1}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \lambda_2^{-1}\mathbf{C_0} \end{bmatrix}$.

This EM approach maximises the likelihood of the data given the hyper-parameters:

$$p(\mathbf{b}|\lambda) = \sqrt{\frac{|\mathbf{A}^T\mathbf{A}|}{(2\pi)^{m-n}|\mathbf{C}||\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}|}} e^{-\frac{1}{2}(\mathbf{b}-\mathbf{Aq})^T\mathbf{C}^{-1}(\mathbf{b}-\mathbf{Aq})} \tag{7.5}$$

This is equivalent to minimising the *restricted maximum likelihood* (REML) objective function (Patterson & Thompson, 1971; Harville, 1974):

$$-\log(p(\mathbf{b}|\lambda)) = \frac{1}{2}\log|\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}| + \frac{1}{2}\log|\mathbf{C}| + \frac{1}{2}(\mathbf{b}-\mathbf{Aq})\,\mathbf{C}^{-1}(\mathbf{b}-\mathbf{Aq}) + const \tag{7.6}$$

There are a number of more general EM approaches for solving these problems (Harville, 1977), but their complete description is beyond the scope of this thesis. The idea would be to estimate coefficients for a linear combination of basis functions that parameterise $\mathbf{C}$ such that the REML objective function is optimised[2]. For image registration methods with nonlinear priors, such as those described in Chapter 4, the covariance matrices would be replaced by inverse Hessian matrices (second partial derivatives of the prior potential with respect to changes in the parameters - see Section 2.4). An extension of the scheme described in Section 4.3.3 could allow spatially varying estimates of λ_1 and λ_2 to be obtained. A more useful model of spatial variability would allow anisotropic variability, with different amounts of distortion in different directions, and possibly also model some covariance among distortions in different directions, and also among distortions of neighbouring tetrahedra. This could be based on a multi-normal model of Hencky tensor elements (see Section 6.4). With more complex models of structural variability, and lots of example datasets, it should be possible for algorithms to empirically learn to represent neuroanatomical variability. Given improved variability models, the estimates of the deformations should become more accurate, which, in turn should lead to better representations of neuroanatomical variability.

---

[2] In the current example, the coefficients are λ_1^{-1} and λ_2^{-1}, with basis functions $\begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ 0 & \mathbf{C_0} \end{bmatrix}$.