Computing Covariances for Mutual Information Co-registration

N.A.Thacker, M.Pokric
Division of Imaging Science and Biomedical Engineering, University of Manchester
email: neil.thacker@man.ac.uk

Abstract

This paper identifies the important role that covariance estimation has to play in the construction of analysis systems. The problem of co-registration for inter-modality clinical volumes is often solved by maximising the so-called mutual information measure. This paper extends the existing theory in this area and suggests a viable way of constructing covariances for mutual information approaches by treating this algorithm as a bootstrapped likelihood based approach. We provide both theoretical and practical tests of the validity of this method. In doing so we identify important subtleties in the current use of these measures for coregistration. These issues suggest potential improvements in the way that such measures might be constructed and used.

Introduction

Co-registration is a cornerstone of many medical image analysis processes. It is often required as a precursor to the analysis of change or for the construction of multi-dimensional data [1]. When constructing systems from separate analysis modules it is crucially important to know the accuracy of the data passing between them and to make appropriate use of this knowledge in subsequent processing. The most common way to represent such data is the covariance matrix. Confirmation that the covariance matrix agrees with the theoretical prediction on sample data is also a very good way of confirming the validity of the assumptions made in the parameter estimation technique. All aspects of the quantitative statistical method must be understood in order to achieve agreement between the theroretical and empirical estimates of parameter accuracy. In achieving this the algorithm module is tested to a level that allows full exploitation in a larger system.

Defining $p(i,j)$ as the joint probability distribution for grey level values i and j at equivalent locations in two images I, J, this measure is defined as;

$$I(I,J) \ = \ \sum_i \sum_j p(i,j) log \frac{p(i,j)}{p(i)p(j)}$$

which has been shown [2] to be monotonically related to the negative log probability of the equivalence between image values;

$$-log(P) \ = \ - N \sum_i \sum_j p(i,j) log \frac{p(i,j)}{p(j)}$$

via

$$-log(P) \;=\; N(H(I) \;-\; I(I,J))$$

where N is the number of voxels and $H(I)$ is the entropy of image I and is fixed. Thus the maxima of a mutual information measure is also the minima of the log probability of the similarity between the two images. The measure is perhaps more easily recognised when written as a sum over voxels v_{ij} in the original data rather than over the histogram.

$$-log(P) \;=\; -\sum_v log\frac{p(i,j)}{p(j)}$$

We would like to be able to compute a covariance for the estimated parameters from such an optimisation, but in fact the log probability term identified here is not appropriate for this process because it uses probabilities of obtaining particular i j combinations within some interval, rather than a likelihood. As a consequence, redefinition of the grey level intervals produces a scaling of this measure. In order to get the log probability into an invariant form we need to rewrite it as follows;

$$-log(P) \;=\; -\sum_v log\frac{p(i,j)\;p(j_{max})}{p(j_{max})\;p(j)}$$

$$=\;-\;\sum_v log\frac{p(i,j)}{p(j_{max})}\;-\;\sum_v log\frac{p(j_{max})}{p(j)}$$

where $p(j_{max})$ is the maximum probability within the distribution. Written in this way the behaviour of the mutual entropy algorithm now becomes explicit. The first term in this equation now corresponds to the conventional χ^2 likelihood statistic which is minimised to achieve alignment. The second term explicitly optimises the "peakiness" of the estimated distribution in order to achieve the maximum correlation between equivalent structure. Unfortunately, this second term is not a true statistic, as it is dependent upon the specific quantisation of the data. However, it is legitimate to ignore bias due to this term at the optimal co-registration of two data sets, provided the likelihood term has sufficient information to generate an accurate estimate of the parameters It is the quadratic approximation of the variation of this first term about the estimated minima which defines the covariance matrix.

Method

We can now make an association between individual data terms and more conventional log-likelihood approaches. In particular we can express the negative log-likelihood in the form;

$$\chi^2 \;=\; \sum_v \chi_v^2 \;=\; \sum_v -log(f_v) \;=\; \sum_v (\sqrt{-log(f_v)})^2$$

where f_v is the underlying continuous likelihood distribution which gives rise to our quantised estimate $p(i,j)/p(j_{max})$ [1].

[1] this function needs to be defined as continous and differentiable in order to make any attempt at covariance estimation

We can now use conventional techniques from the numerical literature [3] as a basis for the estimation of an inverse covariance C_Θ^{-1}on a set of coregistration parameters Θ as follows

$$C_\Theta^{-1} \;=\; \sum_v (\nabla_\Theta \chi_v)^T \otimes (\nabla_\Theta \chi_v)$$

We can estimate this expression using the chain rule as

$$C_\Theta^{-1} \;=\; \sum_v (\partial\chi_v/\partial f_v)^2 (\partial f_v/\partial J_v)^2 (\nabla_\Theta J_v)^T \otimes (\nabla_\Theta J_v)$$

Which expresses the covariance in terms of image derivatives $\nabla_\Theta J_v$ derivatives of the likelihood estimation $\partial\chi_v/\partial f_v$ and the derivative of the likelihood function $\partial f_v/\partial J_v$. Notice that this has the expected properties for image alignment that the maximum contribution to the inverse covariance is made by data which are close to edge features.

From our expression for χ_v we get

$$\partial\chi_v/\partial f_v \;=\; \frac{1}{2f_v\sqrt{-log(f_v)}}$$

The inverse covariance

$$C_\Theta^{-1} \;=\; \sum_v \frac{-(\partial f_v/\partial J_v)^2}{4f_v^2 log(f_v)} (\nabla_\Theta J_v)^T \otimes (\nabla_\Theta J_v) \tag{1}$$

can be considered a general result for the calculation of covariances on parameters θ for any image based bootstrapped likelihood [4]. The equation illustrates that low probability data points will have the main influence over location and stability of the minima. Notice also the lack of scaling due to inherrent image noise, as this information is already encoded in the sampled likelihood distribution.

Results

Theoretical Testing

We can check that this result is sensible by appling it to a naive gaussian model where $f_v = exp(-(I_{jmax} - I_v)^2/(2\sigma_j))$. The covariance estimated using equation (1) is then given as

$$C_\Theta^{-1} \;=\; \sum_v \frac{(\nabla_\Theta J_v)^T \otimes (\nabla_\Theta J_v)}{2\sigma_j^2}$$

This is the same as would have been defined directly for the corresponding gaussian likelihood model, proving the theoretical validity of our derivation.

The Gaussian model results in a pure quadratic form for the log-likelihood function, this covariance estimate is therefore exact in this case. For non-Gaussian data the estimated likelihood functions f_v must be linear over a range determined by the stability of the estimated parameters θ. This can be expected to be true for smoothly varying likelihood functions and large quantities of voxel data. However, as the true probability distributions f_v are unknown, they are generally bootstrapped from the data itself. This problem will now be addressed in the next section.

Experimental Testing

Derivatives of the likelihood function can be estimated to second order using finite differences as can the $\nabla_\Theta J_v$ term. In order to confirm our method for the estimation of covariances we have performed a Monte-Carlo study. The alignemnt technique used selects three orthogonal panels from the first (reference) data set and aligns these within the second (reslice) data set using the mutual information measure. Two data sets were selected on CT and MR for overlapping regions of the head and brain. These were then co-registered to give an initial estimate of the alignment and the covariance estimated. The variation of the alignment as a function of perturbative noise was then explored by repeating the alignment multiple times while adding random gaussian noise to the resliced (MR) data at a level estimated to be in the original data. The distribution of parameter estimates was then compared to the covariance estimate by taking the error function of the scaled mahalanobis distance for each noisy alignment. This produces a series of probability curves (one for each scaling) figure 1. If the covariance estimate matches a particular scaled estimate of the parameter stability then the corresponding distribution will be uniform (flat). For this curve the estimated parameter deviations are tabulated in table 1.

Discussion and Conclusions

The key stage in this analysis is the realisation that the standard mutual information measure is not a true likelihood statistic. Though it is true that it can be written in the form of a log-probability this is not enough so that it can be used for covariance estimation. This is a subtle but important point. Attempting to compute a covariance from the log probability directly would introduce undesirable scaling between data-points. This can be proved quite easily using the Gaussian model with varying σ_j and the standard probability $p(j)$ scaling instead of $p(j_{max})$. The resulting covariance estimate then includes an additional $log(\Delta_i/\sqrt{2\pi}\sigma_j)$ where Δ_i is the histogram interval.

This analysis illustrates that although the log-likelihood function can be related to a measure similar in form to mutual information, as with many image processing algorithms which borrow equations from physics, this is **not** the theoretical foundation of the approach. While it is convenient to refer to the resulting algorithm as maximisation of "mutual information" the similarity of the underlying statistical theory and true mutual information is purely coincidental. Indeed the body of theory from which true mutual information measures arise would not be capable of defining a covariance. It is therefore important that algorithmic design choices are not made on the basis of this chance similarity in the misguided belief that this interpretation is in some way optimal. For example, the quantisation of data necessary for the constuction of a correlation histogram is not only theoretically unnecessary but algorithmically unsound (as it leads to local minima in the cost function). It is probably better to strive to work with approximations to continuous distributions wherever possible, as in the original work [5]. The second term remaining in the standard approach relating to the "peakiness" of the data distribution is a particular cause for concern and can only be reconciled with statistics by interpreting it as a badly normalised second likelihood or a prior probability distribution. Although this term has sensible behaviour it does not have a fully justifiable form. It is issues such as this which may have caused some researchers to have difficulty in implementing these algorithms despite the apparent simplicity of the approach.

This should perhaps lead us to consider alternative formulations which have more validity.

The covariance expression has been derived assuming uncorrelated data terms χ_v in the log-likelihood formulation. Spatial correlation in the data will inevitably reduce the effective number of degrees of freedom, thereby scaling the estimated covariances. Agreement between theory and experiment has been achieved here using only a fraction of the data available for coregistration (three orthogonal panels rather than the entire volume). In general comparison of the true localisation error versus that estimated using the above covariances is likely to behave in a similar manner to all data fitting processes. The estimated errors on the covariances are likely to be smaller than practically observed until the model complexity matches the data. Before this point the main contribution to the error on localisation will be due to an inability of the model to fit the data rather than the stability of estimated parameters. The work of West et al [6] would seem to suggest for example that medical data sets are only rigid to an accuracy of 0.1 voxels. Estimates of covariances for rigid coregistration which predict voxel alignments with accuracy much greater than this are therefore likely to be optimistic. Techniques which attempt deformable coregistration will of course have may more free parameters, the statistical error on these parameters would be expected to dominate accuracy.

References

1. M. Pokric, N.A. Thacker, M.L.J.Scott and A.Jackson, Multi-Dimesional Medical Image Segmentation with Partial Voluming, MIUA, Birmingham, 77-80, 2001.

2. A. Roche, G.Malandain, N.Ayache and S.Prima. Towards a Better Comprehension of Similarity Measures Used in Medical Image Registration. MICCA, 555-566, 1999.

3. W.H.Press B.P.Flannery S.A.Teukolsky W.T.Vetterling,Numerical Recipes in C, Cambridge University Press 1988.

4. A.Lacey, N.A.Thacker, P.Courtney and S.Pollard. TINA 2001: The Closed Loop 3D Model Matcher. BMVC 2001, Manchester, 203-212, 2001.

5. Paul Viola, Alignment by Maximisation of Mutual Information, M.I.T. PhD Thesis, 1995,

6. J. West, J.M. Fitzpatrick, et al, *Comparison and Evaluation of Retrospective Intermodality Brain Image Registration Techniques*, J. Comput. Assist. Tomography, 21, 1997, pp.554-566.