

Computing Covariances for “Mutual Information” Coregistration

P. A. Bromiley*, M. Pokric, and N.A. Thacker

Imaging Science and Biomedical Engineering, Stopford Building,
University of Manchester, Oxford Road, Manchester, M13 9PT.

Abstract. Mutual information (MI) has become a popular similarity measure in multi-modality medical image registration since it was first applied to the problem in 1995. This paper describes a method for calculating the covariance matrix for MI coregistration. We derive an expression for the covariance matrix by identifying MI as a biased log-likelihood measure. The validity of this result is then demonstrated through comparison with the results of Monte-Carlo simulations of the coregistration of T1-weighted to T2-weighted synthetic MRI scans of the brain. We conclude with some observations on the theoretical basis of MI as a log-likelihood.

1 Introduction

The use of MI as a similarity measure for multi-modality coregistration was first proposed in 1995 [1], and it has since become the most popular information-theoretic approach to this problem. Research into coregistration has generally focused on the definition of similarity metrics or on the representation of the transformation model. There is however a growing recognition that characterisation of the accuracy of coregistration is essential if further quantitative processing of the images is to be performed using the resultant transformation model. For example, Crum et. al. [2] state that “...the veracity of studies that rely on non-rigid registration should be keenly questioned when the error distribution is unknown and the results are unsupported by other contextual information”. We present an analytical expression for the covariance matrix of the parameters of MI coregistration, based on the identification of the measure as a biased log-likelihood. This is only the first step towards a full characterisation of the error for the general coregistration problem: for example, it takes no account of the difference between image similarity and biological correspondence. It does however provide a lower bound on the error, which may be achievable for certain coregistration problems and definitions of correspondence.

Mutual information $\mathcal{I}(I; J)$ measures the Kullback-Leibler divergence [3] between the joint probability distribution $p(i, j)$ of two images or image volumes I and J and the product of their marginal distributions $p(i) \cdot p(j)$ [3],

$$\mathcal{I}(I; J) = \sum_{i,j} p(i, j) \log \frac{p(i, j)}{p(i) \cdot p(j)}$$

i.e. the divergence of the joint distribution from the case of complete independence of the images, where the sum is performed over a joint intensity histogram of the image pair. Therefore, maximisation of this measure with respect to a set of coregistration parameters will optimise the image alignment. Following [4], we can split the sum into

$$\mathcal{I}(I; J) = \sum_i p(i) \log \frac{1}{p(i)} + \sum_{i,j} p(i, j) \log \frac{p(i, j)}{p(j)}.$$

Recognising that the first term on the R.H.S. is the entropy $H(I)$ of image I [3] and that in the limit of large samples $p(i, j) = N_{ij}/N$, where N_{ij} is the number of entries in histogram bin (i, j) and N is the total number of entries in the histogram, we obtain

$$N[\mathcal{I}(I; J) - H(I)] = \sum_v \log \frac{p(i, j)}{p(j)} = \log P(I|J),$$

where v represents a sum over voxels rather than histogram bins, and P represents a probability summed over an entire pair of images or volumes. At this point we can make the arbitrary definition that I is the target (fixed) image and J the source image i.e. the image altered by the transformation model. If we ensure that the data sampled from the target image does not change by keeping the overlap with the source image constant¹, $H(I)$ will be a constant, indicating that MI is then a monotonic function of the log-probability of image I given image J ,

$$\log P(I|J) = N(\mathcal{I}(I; J)) + const. \quad (1)$$

* E-mail: paul.bromiley@talk21.com

¹Excluding an appropriately sized border around the target image will ensure that all of the remaining data overlaps the source image throughout the optimisation.

The covariances for a maximum-likelihood (ML) method are bounded by the minimum variance bound (MVB) [5]

$$C_{\theta_r, s}^{-1} = - \left. \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right|_{\theta_0}$$

where the θ represent parameters of some model, L represents the likelihood function, and θ_0 represents the parameters at the maximum of L . This bound becomes exact if the log-likelihood is quadratic i.e. the likelihood function is Gaussian. Proceeding with the Gaussian assumption, we can write

$$L = \prod_d A_d e^{-\frac{(I_d - I_M)^2}{2\sigma_d^2}} \Rightarrow \log L = \sum_d -\frac{(I_d - I_M)^2}{2\sigma_d^2} + \log A_d \Rightarrow \left. \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right|_{\theta_0} = \sum_d -\frac{1}{\sigma_d^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s} \Bigg|_{\theta_0}$$

where A_d is the normalisation of the Gaussian, I_d are the data and I_M the corresponding model predictions, and σ_d is the standard deviation of the data. Note that any constant normalisation of the Gaussian (A_d) disappears upon differentiation. In simple ML techniques e.g. linear least-squares fitting, the normalisation of L will indeed be constant. However, the MI measure is a ‘‘bootstrapped’’ likelihood, constructed from the joint histogram rather than from some explicit model and so the usual normalisation (to the area under the distribution) may no longer be constant: for example, simply altering the histogram bin size will alter the normalisation. Fortunately, a solution is available in the form of the χ^2 metric. If we normalise to the peak of the distribution, A_d becomes 1 and disappears upon taking logs. Maximisation of the log-likelihood is then directly equivalent² to minimisation of the χ^2

$$\log L = \sum_d -\frac{(I_d - I_M)^2}{2\sigma_d^2} = -\frac{\chi^2}{2}.$$

Whilst this is explicitly true for a Gaussian L , we would suggest that this statistic has higher utility regardless of the form of the underlying distribution as it provides appropriate normalisation. Furthermore, the χ^2 can be written in terms of a sum over individual data terms: the so-called χ of the χ^2

$$\chi^2 = \sum_d \chi_d^2 = \sum_i -2 \log(L_d) \Rightarrow \chi_d = \sqrt{-2 \log L_d}.$$

The expression for the MVB can also be rewritten in this form, through comparison with the previous result for a Gaussian likelihood function

$$\chi_d = \frac{(I_d - I_M)}{\sigma_d} \Rightarrow \sum_d \frac{\partial \chi_d}{\partial \theta_r} \frac{\partial \chi_d}{\partial \theta_s} = \sum_d \frac{1}{\sigma_d^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s} \Rightarrow C_{\theta}^{-1} = \sum_d (\nabla_{\theta} \chi_d)^T \otimes (\nabla_{\theta} \chi_d) \Bigg|_{\theta_0}. \quad (2)$$

The Gaussian assumption need only hold over a sufficient range around the maximum to allow the calculation of the derivatives, and since in coregistration the likelihood distribution is composed of tens of thousands of individual data terms (voxels) we can expect, due to the Central Limit Theorem, that this assumption will hold.

The equivalent χ^2 term in the MI measure can be identified using the log-likelihood from Eq. 1

$$\log P(I|J) = \sum_v \log \frac{p(i, j)}{p(j)} = \sum_v \log \frac{p(i, j)p(i_{max}, j)}{p(j)p(i_{max}, j)} = \sum_v \log \frac{p(i, j)}{p(i_{max}, j)} + \sum_v \log \frac{p(i_{max}, j)}{p(j)}.$$

The first term on the RHS is the χ^2 metric, normalised to the distribution peak $p(i_{max}, j)$ as required. The second is a bias term dependent on the non-uniform normalisation of the likelihood distribution. This expression elucidates the behaviour of the MI measure: it is a ML measure biased with a term that maximises the ‘‘peakiness’’ of the distributions in the joint histogram, in order to maximise the correlation between equivalent structures in the images. If we assume that the bias term varies slowly compared to the χ^2 term, which is reasonable since it depends on the marginal distribution, then Eq. 2 can be used. Applying the chain rule to expand the derivative of χ_v (the χ for each data term i.e. voxel) w.r.t. the model parameters in terms of the likelihood for each voxel L_v and the voxel values themselves, J_v , gives

$$C_{\theta}^{-1} = \sum_v \left(\frac{\partial \chi_v}{\partial L_v} \right)^2 \left(\frac{\partial L_v}{\partial J_v} \right)^2 (\nabla_{\theta} J_v)^T \otimes (\nabla_{\theta} J_v) \Bigg|_{\theta_0}$$

and so

$$C_{\theta}^{-1} = - \sum_v \frac{\left(\frac{\partial p(i, j)}{\partial J_v} - \frac{p(i, j)}{p(i_{max}, j)} \frac{\partial p(i_{max}, j)}{\partial J_v} \right)^2}{2p(i, j)^2 \log \frac{p(i, j)}{p(i_{max}, j)}} (\nabla_{\theta} J_v)^T \otimes (\nabla_{\theta} J_v) \Bigg|_{\theta_0}. \quad (3)$$

²The χ^2 used here refers to the general definition, as used in statistical tests for assessing the adequacy of fitting results for a number of degrees of freedom, not the specific computational forms used for comparing histograms or tables.

2 Method

The covariance estimation technique was tested on the rigid coregistration of T2 to T1 weighted simulated MR image volumes of a normal brain, obtained from Brainweb [6]. Each volume consisted of 55 slices of 217 by 195 voxels, with Gaussian noise added at 1% of the dynamic range. MI coregistration was implemented within the TINA software package (www.tina-vision.net), using simplex minimisation, allowing the coregistration to optimise the rotation (as Euler angles), translation and scaling of the images. The source images were rotated by 5° prior to coregistration in order to suppress interpolation artefacts³, following the suggestion by [7]. However, the coregistrations were started from the correct alignment.

Monte-Carlo simulations were run by adding Gaussian noise to the source image at levels of 0.25 to 2.5 times the original image noise, in ten steps of 0.25σ : covariances were calculated from 1000 coregistrations performed at each noise level. Then Eq. 3 was applied at each noise level, taking the median of 100 estimates of the covariances, over a range around the maximum that represented a change of around 0.5% in the χ^2 , in order to stabilise the calculation against the effects of interpolation artefacts, local minima etc. Finally, the estimated and practical (Monte-Carlo) covariances were compared. The covariance matrices were prepared from a set of $1 \times n$ vectors of parameters, and so had only n degrees of freedom despite containing n^2 parameters. It is therefore sufficient to compare only the n diagonal parameters, the variances, or their square-roots, the standard deviations.

3 Results

Fig. 1. shows the dependence of the standard deviations of the transformation model parameters on added noise. The Monte-Carlo results scale linearly with the addition of noise as expected, and linear least-squares fits to the data are shown. Some outliers are present at high noise levels due to bimodality in the Monte-Carlo results: the added noise sufficiently destabilised the coregistration that a local maximum close to the global maximum began to contribute. Therefore, these points were omitted from the fitting process. The estimates from the analytical expression are also shown together with linear fits. The estimated covariances of the translation parameters are identical to the Monte-Carlo results to within the noise on the data. The estimates for the rotational parameters show some divergence, and are also notably noisier, due to the non-linear nature of rotational transformations. The estimates for the scaling parameters show the greatest divergence at the higher noise levels. This is due to an effective underestimate of the covariance through the Monte-Carlo experiments. The scaling parameters are more susceptible to interpolation artefacts than the other parameters, leading to oscillations in the similarity metric around the global maximum. The Monte-Carlo results tend to fall into the local optima generated by these oscillations, leading to underestimation of the covariances, whereas the estimated covariance was stabilised against this effect by taking the median value over 100 points around the global maximum. The sources of the differences between the Monte-Carlo and estimated covariances were identified by plotting the likelihood function around its maximum. Overall, all of the estimated covariances either match the Monte-Carlo results closely, or converge at low noise levels, and are always within a factor of two.

4 Conclusion

This paper has provided a derivation of an analytical expression for the covariances of the parameters of MI coregistration. The result had been confirmed through comparison with Monte-Carlo simulations. The estimated variances match the Monte-Carlo results, both confirming the validity of the covariance estimation technique and justifying the assumption that the MI bias term is negligible in this case.

The covariance estimate presented here has a number of practical uses. Equivalence between estimated and Monte-Carlo variances can be used to demonstrate numerical stability in the implementation of a coregistration algorithm. Error propagation can be used to calculate spatial errors on voxel locations from the covariance matrix of the transformation model. Finally, the technique is equally valid for non-rigid coregistration, where the coregistration errors will vary spatially and so must be quantified if any further statistical analysis of the data is to be performed.

The derivation also illustrates some features of MI in general. Most important is the relationship between MI and log-likelihood. The consistency between the estimated covariances and the practical coregistration performance confirms that this interpretation is valid. We maintain that this is the true theoretical basis of the method, rather

³Such artefacts occur due to sudden changes in the number of voxels from the two image volumes that overlap exactly, and so are added to the joint histogram without interpolation. Interpolation smooths the data, so this results in a sudden change in the MI measure.

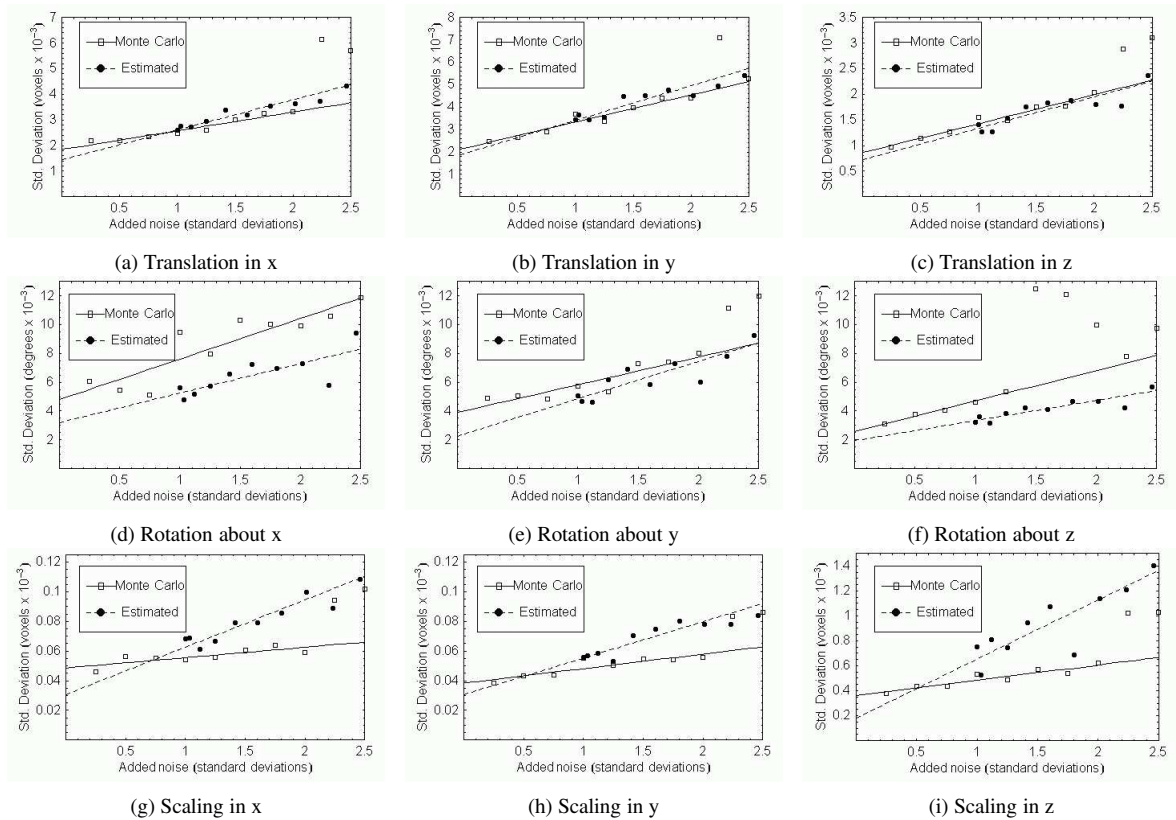


Figure 1. The standard deviations of the coregistration parameters for the synthetic data. The lines show linear least-squares fits to the data, omitting the outliers from the Monte-Carlo experiments due to evidence of bimodality around the maximum (see main text).

than its relationship to concepts of entropy. The likelihood interpretation may also provide new perspectives on mutual information and associated similarity measures, suggesting alternatives based on quantitative statistics. For instance, normalised MI measures [7] are currently used for coregistration problems with varying sample sizes. The approach adopted here suggests using a χ^2 metric i.e. an appropriately normalised log-likelihood, accommodating the variation in sample size as a change in the number of degrees of freedom. Ultimately, this could lead to a coregistration algorithm implemented in expectation-maximisation form.

Acknowledgements

The authors would like to acknowledge the support of the EPSRC and the MRC (IRC: From Medical Images and Signals to Clinical Information), and of the European Commission (An Integrated Environment for Rehearsal and Planning of Surgical Interventions). All software is freely available from our web site www.tina-vision.net.

References

1. P. Viola & W. M. Wells. "Alignment by maximisation of mutual information." *International Journal of Computer Vision* **24(2)**, pp. 137–154, 1997.
2. W. R. Crum, L. D. Griffin, D. V. G. Hill et al. "Zen and the art of medical image registration: correspondence, homology, and quality." *Neuroimage* **20**, pp. 1425–1437, 2003.
3. T. M. Cover & J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.
4. A. Roche, G. Malandain, N. Ayache et al. "Towards a better comprehension of similarity measures used in medical image registration." In *Proceedings MICCAI'99*, pp. 555–566. 1999.
5. R. J. Barlow. *Statistics: A Guide to the use of Statistical Methods in the Physical Sciences*. John Wiley and Sons Ltd., UK, 1989.
6. C. A. Cocosco, V. Kollokian, R. K.-S. Kwan et al. "Brainweb: Online interface to a 3D MRI simulated brain database." *Neuroimage: Proceedings of the 3rd International Conference on Function Mapping of the Human Brain, Copenhagen, May 1997* **5(4)**, pp. S425, 1997.
7. J. P. W. Pluim, J. B. Antoine Maintz & M. A. Viergever. "Interpolation artefacts in mutual information-based image registration." *Computer Vision and Image Understanding* **77**, pp. 211–232, 2000.