

Empirical Evaluation of Covariance Estimates 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 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 mutual information coregistration. We derive an expression for the matrix through identification of mutual information with a 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 and genuine MRI scans of the brain. We conclude with some observations on the theoretical basis of the mutual information measure as a log-likelihood.

1 Introduction

The use of mutual information (MI) as a similarity measure for multi-modality coregistration was first proposed in 1995 [1], and since then has 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 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. However, it provides a lower bound on the error, which may be attainable for certain coregistration problems and definitions of correspondence.

Mutual information $\mathcal{I}(I; J)$ measures the Kullback-Leibler divergence between the joint probability distribution $p(i, j)$ of two images 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 histogram. Therefore, maximisation of this measure with respect to a set of coregistration parameters will optimise the image alignment. Following [4], we can write

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

Recognising that the first term on the R.H.S. is the entropy $H(I)$ of image I [3] and that $p(i, j) = N_{ij}/N$, 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

$$\log P(I|J) = N[\mathcal{I}(I; J) - H(I)] = \sum_v \log \frac{p(i, j)}{p(j)} \quad (1)$$

where v represents a sum over voxels rather than histogram bins. 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 overlapping regions of the images always include the whole of the target image, for example by excluding an appropriately sized border around the reference image, $H(I)$ will be a constant, giving

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

Therefore the MI is a monotonic function of the log-probability of image I given image J .

The covariances for a maximum likelihood technique are given by the minimum variance bound [5]

$$C_\theta^{-1} = - \left. \frac{\partial^2 \log L}{\partial \theta_m \partial \theta_n} \right|_{\theta_0}$$

where θ represent parameters of some model, θ_0 represents the parameters for optimal alignment, and L represents the likelihood function. This bound becomes exact if the log-likelihood is quadratic i.e. the likelihood function in Gaussian. Assuming a Gaussian likelihood function

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

where A_d is the normalisation of the Gaussian, I_d are the data and I_M the corresponding model predictions, and σ_d are the standard deviations of the data. Note that any constant normalisation of the Gaussian (A_d) disappears upon differentiation. In simple maximum likelihood techniques e.g. linear least-squares

fitting, the normalisation of L will indeed be constant. However, MI is constructed from a so-called “bootstrapped” likelihood, constructed from the joint histogram rather than an explicit model. In that case, 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, then A_d becomes 1 and disappears upon taking logs. The maximisation of the log-likelihood is then directly equivalent 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.

The χ^2 can be written in terms of a sum over individual data terms, the so-called χ of the χ^2

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

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

$$\chi_d = \frac{(I_i - I_M)}{\sigma_d} \Rightarrow \sum_d \frac{\partial \chi_d}{\partial \theta_r} \frac{\partial \chi_d}{\partial \theta_s} = \sum_d \frac{1}{2\sigma_d} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s}$$

Comparing this to the previous expression for the covariances of a Gaussian likelihood, we can write,

$$\Rightarrow C_\theta^{-1} = \sum_d 2(\nabla_\theta \chi_d)^T \otimes (\nabla_\theta \chi_d) \Big|_{\theta=\theta_{max}} \quad (3)$$

The Gaussian assumption need only be true over a sufficient range around the minimum that the derivatives can be calculated, and since in rigid coregistration we are dealing with a likelihood composed from ≈ 100000 voxels we would expect, via. the Central Limit Theorem, that this would be a good approximation.

In order to identify the equivalent χ^2 term in the MI measure, we can split Eq. 1 into two terms

$$-\log P(I|J) = -\sum_v \log \frac{p(i, j)}{p(i_{max}, j)} - \sum_v \log \frac{p(i_{max}, j)}{p(j)} \quad (4)$$

The first term on the RHS is the χ^2 metric, normalised to the distribution peak 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 maximum likelihood 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. 3 can be used: expanding using the chain rule, substituting for the differential of χ_v from Eq. 2, using Eq. 2 and Eq. 4 to substitute for L_v , and remembering that the model terms I_M in this case are represented by the source image voxels J_v gives

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

$$C_\theta^{-1} = - \sum_v \frac{\left(\frac{\partial p(i,j)}{\partial J_v} - \frac{p(i,j)}{p(i_{max},j)} \frac{\partial p(i_{max},j)}{\partial J_v} \right)^2}{2p(i,j)^2 \log \frac{p(i,j)}{p(i_{max},j)}} (\nabla_\theta J_v)^T \otimes (\nabla_\theta J_v) \Big|_{\theta=\theta_{max}} \quad (5)$$

2 Method

The covariance estimation technique was first tested on the rigid coregistration of T2 to T1 weighted simulated MR of a normal brain, obtained from Brainweb [6]. Each volume consisted of 55 slices of 217 by 195 voxels, with Gaussian random noise added at 1% of the dynamic range of the images. The technique was repeated on the coregistration of genuine T2 to T1 weighted MR from a normal volunteer. These image volumes consisted of 29 3mm thick slices of 256 by 256 (0.89mm by 0.89mm) voxels. The noise on the images, measured using the width of zero crossings in horizontal and vertical gradient histograms, was again approximately 1% of the dynamic range of the images. MI coregistration was implemented within the TINA machine vision software package (www.tina-vision.net), using simplex minimisation, and allowing the coregistration to optimise the rotation (as Euler angles), translation and scaling of the images. A rotation offset of 5° was added to the floating images before coregistration, but the coregistration was started from the correct alignment. This followed the suggestion by Pluim et. al. [7] regarding the suppression of interpolation artefacts. These artefacts arise at points where large portions of the voxel grid for both images coincide, and so large numbers of voxels from the source image are used without interpolation. Since interpolation inevitably smooths the data, such points lead to sudden jumps in the value of the similarity measure.

Monte-Carlo simulations were run by adding random Gaussian noise to the reference image at levels of 0.25 to 2.5 times the original image noise, in ten steps of 0.25σ . One thousand coregistrations were performed at each noise level, and the results used to estimate the covariance matrix of the coregistration parameters. Then, the above estimate was applied at each noise level, taking the median of 1000 estimate of the covariances over a range around the minimum 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 two covariance estimates at each noise level were compared. Since each covariance matrix is prepared from a set of $1 \times n$ vectors of parameters, it has only

n degrees of freedom despite containing n^2 parameters. Therefore, it is sufficient to examine only n parameters, and so we will limit the discussion of the results to the n diagonal elements (the variances) alone.

3 Results

Fig. 1. shows the standard deviations on the parameters for the Brainweb data. In each case, the Monte-Carlo estimates scale linearly with the addition of noise as expected. Linear least-squares fits to the data are shown. The two points for the highest levels of added noise show a departure from the trend. This was due to bimodality in the Monte-Carlo results i.e. the added noise destabilised the coregistration enough that a local minimum close to the global minimum 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 covariance estimates on the translation parameters are identical between the Monte-Carlo results and the analytical estimate to within the noise on the data. The results for the rotational parameters show some divergence, and are also notably noisier, due to the non-linear nature of rotational transformations. The results 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 minimum. The Monte-Carlo results tend to fall into the local minima generated by these oscillations, leading to underestimates of covariances, whereas the estimated covariance was stabilised against this effect by taking the median value over 1000 points around the global minimum. 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 of the Monte-Carlo results.

Fig. 2 shows the standard deviations on the coregistration parameters for the genuine MR data. Again, all results scale linearly with noise as expected. The image content and noise were roughly equivalent to the Brainweb data, but the genuine MR volumes contained only half as many voxels, implying that the variances should be roughly twice as large, and this can indeed be seen in the results. The other features of the results are all broadly similar: the estimated covariances either match the Monte-Carlo results or converge with them at low noise levels. An exception is seen in the scaling parameter in the z direction: here the estimated covariances are significantly higher. This was due to artefacts in the similarity metric around the minimum, which made it impossible to produce a stable covariance estimate.

4 Conclusion

This paper has provided a derivation of an analytical expression for the covariances in the parameters of mutual information (MI) coregistration. The valid-

ity of the result has been demonstrated through comparison with the results of Monte-Carlo simulations on both simulated and genuine MR images of the brain. The estimated variances are consistent between the two techniques, confirming that the equation for variance estimation is valid and that our assumption that the bias term is negligible is justified.

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 than its relationship to concepts of entropy. It is the link to maximum likelihood that allows the theory to support calculation of a covariance matrix. The likelihood interpretation may also provide new perspectives on MI 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, in which the variation in sample size can be accommodated as a variation 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. Viola, P., Wells, W.M.: Alignment by maximisation of mutual information. *International Journal of Computer Vision* **24** (1997) 137–154
2. Crum, W.R., Griffin, L.D., Hill, D.V.G., Hawkes, D.J.: Zen and the art of medical image registration: correspondence, homology, and quality. *Neuroimage* **20** (2003) 1425–1437
3. Cover, T.M., Thomas, J.A.: *Elements of Information Theory*. John Wiley and Sons, New York (1991)
4. Roche, A., Malandain, G., Ayache, N., Prima, S.: Towards a better comprehension of similarity measures used in medical image registration. In: *Proceedings MICCAI'99*. (1999) 555–566
5. Barlow, R.J.: *Statistics: A Guide to the use of Statistical Methods in the Physical Sciences*. John Wiley and Sons Ltd., UK (1989)
6. Cocosco, C.A., Kollokian, V., Kwan, R.K.S., Evans, A.C.: Brainweb: Online interface to a 3D MRI simulated brain database. *Neuroimage* **5** (1997) S425
7. Pluim, J.P.W., Antoine Maintz, J.B., Viergever, M.A.: Interpolation artefacts in mutual information-based image registration. *Computer Vision and Image Understanding* **77** (2000) 211–232

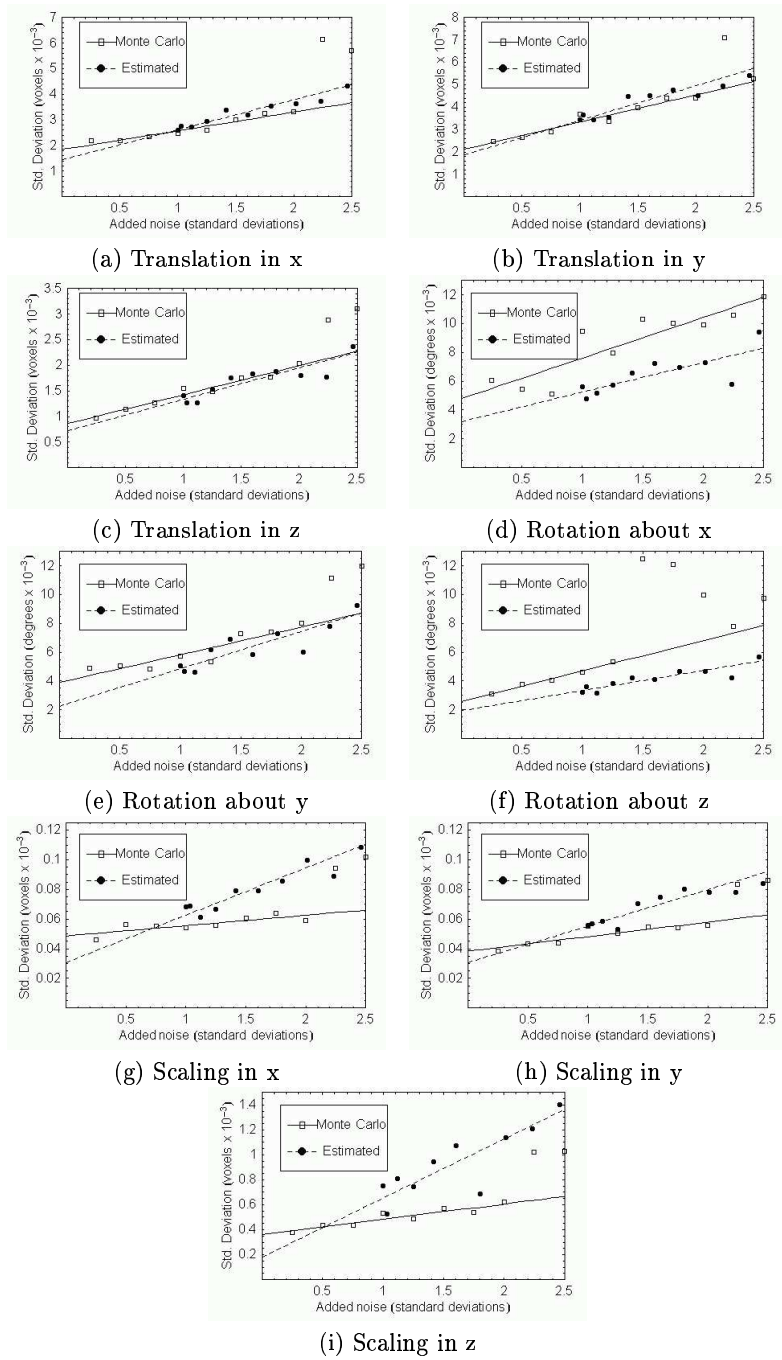


Fig. 1. The standard deviations on the coregistration parameters for the Brainweb data. The lines show least-squares linear fits to the data, omitting the top two points from the Monte-Carlo experiments due to evidence of bimodality around the minimum (see main text).

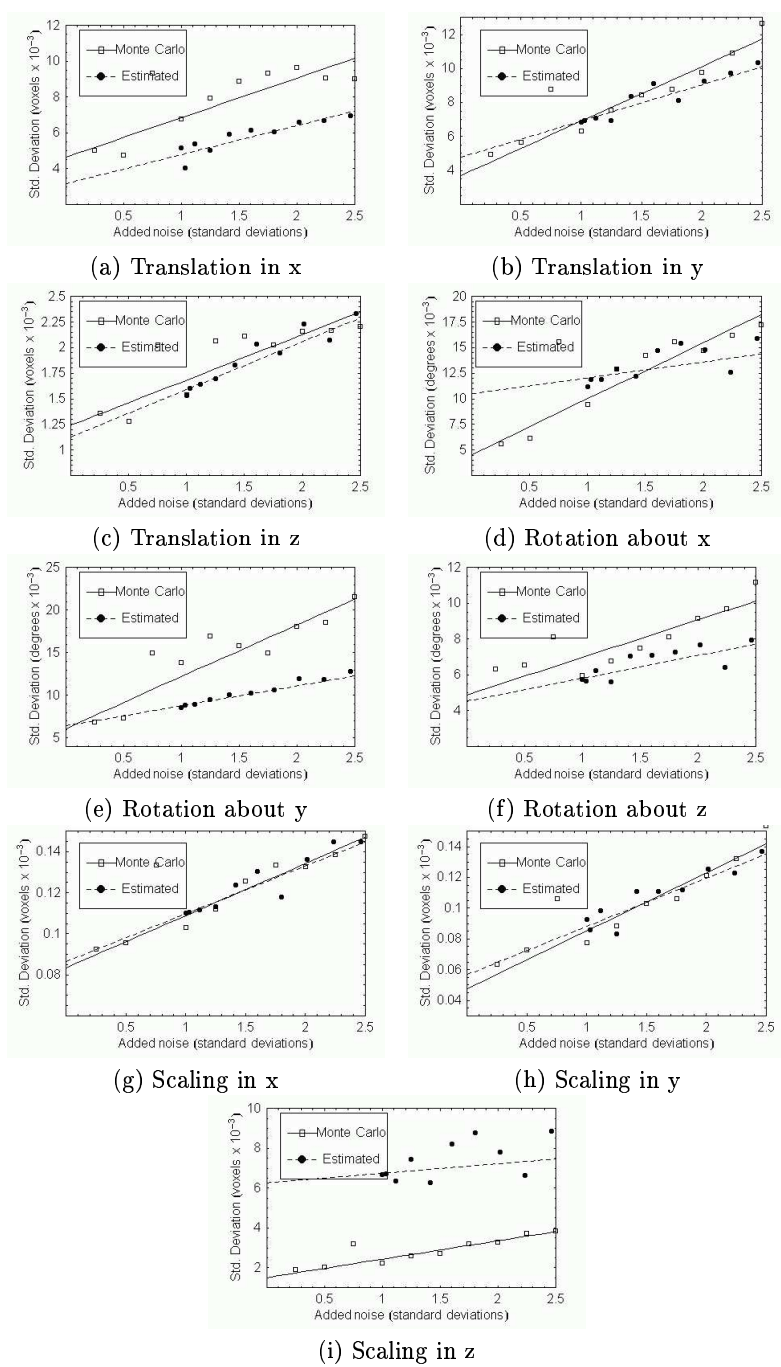


Fig. 2. The standard deviations on the coregistration parameters for the genuine MR data. The lines show least-squares linear fits to the data.