

Tina Memo No. 2001-013  
Internal Report

# Computing Covariances for Mutual Information Co-registration.

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

Last updated  
23 / 3 / 2004



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

# Computing Covariances for Mutual Information Co-registration.

N.A. Thacker, P.A. Bromiley and M. Pokric  
 Imaging Sciences and Biomedical Engineering  
 email: neil.thacker@manchester.ac.uk

## Approach

The problem of co-registration for inter-modality clinical volumes is often solved by maximising the so-called mutual information measure. This document presents a derivation of an analytical expression for the covariance between the parameters for mutual information coregistration. Only the primary results are presented: a more detailed mathematical analysis proceeding from first principles is presented in the companion memo [1].

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

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

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

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

via.

$$-\log P(I|J) = N(H(I) - \mathcal{I}(I, J))$$

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.

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

We would like to be able to compute a covariance matrix for the estimated parameters from such an optimisation. To this end we can split the expression for the mutual information into two terms

$$\begin{aligned} -\log P(I|J) &= - \sum_v \log \frac{p(i, j)p(i_{max}, j)}{p(i_{max}, j)p(j)} \\ &= - \sum_v \log \frac{p(i, j)}{p(i_{max}, j)} - \sum_v \log \frac{p(i_{max}, j)}{p(j)} \end{aligned}$$

where  $p(i_{max}, j)$  is the maximum probability within the distribution along each column of the joint image histogram. 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  $\chi^2$  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. This second term depends on the marginal distribution in image  $J$ , which will change as optimisation proceeds, and it is therefore capable of introducing bias into the results. However, it is legitimate to ignore bias due to this term provided the likelihood term has sufficient information to generate an accurate estimate of the coregistration parameters<sup>1</sup>. It is the second differential at its maximum of the first term, approximated as a quadratic, that defines the covariance matrix.

As a consequence of this relationship we can 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

$$\chi^2 = \sum_v \chi_v^2 = \sum_v -2 \log(L_v) \Rightarrow \chi_v = \sqrt{-2 \log L_v}$$

---

<sup>1</sup>This assumption may not be satisfied for approaches which have a large number of parameters, such as non-rigid co-registration

where  $L_v$  is the underlying continuous distribution<sup>2</sup>. that generates the quantised estimate  $p(i, j)/p(i_{max}, j)$

If  $L_v$  is Gaussian, then this relationship reproduces the conventional *chi-square* under the normalisation outlined above. For non-Gaussian data it is the more general log-likelihood, from which we can estimate covariance.

We can now use conventional techniques from the numerical literature [3] as a basis for the estimation of an inverse covariance  $C_{\Theta}^{-1}$  on a set of coregistration parameters  $\Theta$ . In particular, since the log-likelihood is derived from probability terms normalised such that their peak value is unity, we can apply the expression [1]

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

Alternatively, we can expand this expression using the chain rule as

$$C_{\Theta}^{-1} = \sum_v \left( \frac{\partial \chi_v}{\partial L_v} \right)^2 \left( \frac{\partial L_v}{\partial J_v} \right)^2 (\nabla_{\Theta} J_v)^T \otimes (\nabla_{\Theta} J_v)$$

which expresses the covariance in terms of image derivatives  $\nabla_{\Theta} J_v$ , derivatives of the likelihood estimation  $\partial \chi_v / \partial L_v$  and the derivative of the likelihood function  $\partial L_v / \partial J_v$ . Where the use of  $\nabla_{\Theta} J_v$  ensures consistent treatment of signs during the formation of the outer product. 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  $\chi_v$  we obtain

$$\frac{\partial \chi_v}{\partial L_v} = \frac{-1}{L_v \sqrt{-2 \log(L_v)}}$$

and thus

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

which can be considered a general result for the calculation of covariances on parameters  $\theta$  for any image based bootstrapped likelihood [4] in the case where the likelihood has been generated from probabilities normalised such that their peak value is unity.

We can check that this result is sensible by applying it to a naive Gaussian model, using the normalisation outlined above

$$L_v = \exp\left[-\frac{(J_v - J_M)^2}{2\sigma_j^2}\right]$$

The estimated covariance is then given as

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

as expected. The Gaussian model results in a pure quadratic form for the log-likelihood function, and so the covariance estimate is exact in this case. For non-Gaussian data the estimated log-likelihood functions  $L_v$  must be quadratic over a range determined by the stability of the estimated parameters  $\theta$ . This can be expected to be true for smoothly varying likelihood functions and large quantities of voxel data.

Conversely, we can show that this new interpretation of normalisation ( $p(i, j)/p(i_{max}, j)$ ) is needed for this calculation of the covariance by observing what happens with the derivation of covariance if we keep the original MI form for the likelihood terms  $L$ . Using  $L = p(i, j)/p(j)$  results in a covariance estimate of;

$$C_{\Theta}^{-1} = \sum_v \frac{-(\partial p(i, j) / \partial I_v)^2}{4 p(i, j)^2 \log(p(i, j) / p(j))} (\nabla_{\Theta} I_v)^T \otimes (\nabla_{\Theta} I_v)$$

This could not be correct as the value of the estimated covariance depends directly upon our choice of histogram binning due to the presence of a  $\log(p(i, j) / p(j))$  term in the denominator. Although it is true that the binning process could have an effect on the accuracy of estimated results, it is to be hoped that these problems will be removed for sufficient approximation of the probability distribution with small enough bins and large samples. Any absolute estimate of parameter accuracy should therefore not contain terms which still depend directly upon bin scale. The presence of a  $\log(p(i, j) / p(i_{max}, j))$  term resulting from our preferred normalisation, converges to a constant estimate in the required way.

<sup>2</sup>this function needs to be defined as continuous and differentiable in order to make any attempt at covariance estimation

As the true probability distributions  $L_v$  are unknown, they are generally bootstrapped from the data itself using a process such as

$$L_v = \frac{n(i, j) + 1}{n(i_{max}, j) + 1}$$

where  $n(i, j)$  is the number of pixels with joint grey level values  $i$  and  $j$  and  $n(i_{max}, j)$  is estimated to second order. At this point it also makes sense to relax the requirement of grey-level binning and use the underlying grey level values to interpolate the estimated likelihood function if they have more accuracy. Derivatives of the likelihood function can also be estimated to second order using finite differences.

$$\frac{\partial L_v}{\partial J_v} = \frac{n(i, j + 1) - n(i, j - 1)}{2n(i_{max}, j) + 2}$$

Substituting these results into our expression for the inverse covariance gives

$$C_{\Theta}^{-1} = \sum_v \frac{-(n(i, j + 1) - n(i, j - 1))^2}{8(n(i, j) + 1)^2 \log((n(i, j) + 1)/(n(i_{max}, j) + 1))} (\nabla_{\Theta} J_v)^T \otimes (\nabla_{\Theta} J_v)$$

which also 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 inherent image noise, as this information is already encoded in the sampled likelihood distribution.

## Comments

This document has presented the derivation of a discrete solution for the covariance matrix associated with mutual information coregistration, and has illustrated several important features of the method. In particular, 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 co-incidental. 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 construction 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. Although this term has sensible behaviour its presence may bias the coregistration result. 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 observant reader may have noticed that the presented method yields a result which is asymmetric under interchange of the data sets under alignment, ie: there is a second estimate of the parameter covariances which can be obtained by swapping images  $I$  and  $J$ . This does not invalidate this result as potentially, given the assumptions and the approximations, both are valid estimates. We can choose to model the data distributions with respect to either data set. It would therefore be legitimate to compute both and compare them for consistency. However, the covariance expression is derived assuming uncorrelated data terms  $\chi_v$  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. This scaling may be different for the two images but reproducible enough for calibration of these estimates.

We have no data as yet to back up our assertion that this approach will give the correct estimation of covariance, but we may like to predict what we might find when we have. A 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. Therefore, estimates of covariances for rigid coregistration which predict voxel alignments with accuracy much greater than this do not reflect the ability to determine image alignments on a voxel-by-voxel basis.

## References

1. P.A. Bromiley and N.A. Thacker. Computing Covariances for Mutual Information Coregistration 2. TINA Memo No. 2003-002, <http://www.tina-vision.net/docs/memos.php>, 2003.
2. A. Roche, G.Malandain, N.Ayache and S.Prima. Towards a Better Comprehension of Similarity Measures Used in Medical Image Registration. MICCAI, 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.