

Tina Memo No. 2003-002
Internal Report

Computing Covariances for Mutual Information Coregistration 2

P.A. Bromiley and N.A. Thacker

Last updated
26 / 8 / 2004



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

Computing Covariances for Mutual Information Coregistration 2

P.A. Bromiley and N.A. Thacker
Imaging Science and Biomedical Engineering Division
Medical School, University of Manchester
Manchester, M13 9PT, UK
paul.bromiley@man.ac.uk

1 Introduction

This document is intended to act as a companion to TINA memo no. 2001-013 “Computing Covariances for Mutual Information Coregistration”, expanding the theoretical treatment such that all results are derived from first principles. Section 1 provides a proof of the minimum variance bound in terms of the likelihood function, demonstrating that the inverse covariance matrix is bounded by the second derivative of the log-likelihood function with respect to the parameters at its maximum, and that this bound becomes an equality in the case of a Gaussian likelihood function. The validity of the results is then confirmed through assuming a Gaussian distribution for the likelihood function. Section 2 develops similar proofs in terms of the χ^2 metric. Section 3 then applies these results to the mutual information metric. Basic proofs for covariance matrices, for the relationships between entropy, mutual information, and the Kullback-Leibler divergence, and for the Schwartz inequality are given in the Appendices.

2 Covariances and Log-Likelihood Functions

2.1 The Minimum Variance Bound

The proofs presented in this section are a concatenation and extension of those presented in [1] and [2]. Suppose we have a likelihood function L for a set of independent observations x where $f(x|\theta)$ is the probability density function of x , and θ are some parameters. The likelihood function is given by

$$L(x_1 \dots x_n | \theta) = f(x_1 | \theta) f(x_2 | \theta) \dots f(x_n | \theta)$$

By definition the total probability is normalised to unity ¹

$$\int \dots \int L dx_1 \dots dx_n = 1$$

so, differentiating w.r.t. one of the model parameters θ_r , and assuming that the order of integration and differentiation can be interchanged, gives

$$\int \dots \int \frac{\partial L}{\partial \theta_r} dx_1 \dots dx_n = 0 = \int \dots \int \frac{\partial \log L}{\partial \theta_r} L dx_1 \dots dx_n \quad (1)$$

Now the expectation value of a function $h(x)$ of a random variate x is given by integrating over all possible values of all the x_i weighted by the total probability i.e. the likelihood function

$$\int \dots \int h(x_1 \dots x_n) L(x_1 \dots x_n | \theta) dx_1 \dots dx_n = \langle h(x_1 \dots x_n) \rangle \quad (2)$$

and so comparing Eq. 1 with Eq. 2. gives

$$\left\langle \frac{\partial \log L}{\partial \theta_r} \right\rangle = 0$$

¹Due to the arbitrary specification of the parameter space this means that likelihood is not a true statistic and can only be used for parameter estimation for problems where the normalisation is fixed a-priori.

In addition, the expectation value of any estimator $\hat{\theta}_r$ of the model parameter θ_r will be given by

$$\int \dots \int \hat{\theta}_r L dx_1 \dots dx_n = \langle \hat{\theta}_r \rangle = \theta_r$$

where we can equate to the true value θ_r in the limit of large n if the estimator is consistent. The estimator depends only on the data, whereas the likelihood function depends on both the data and the model, and so differentiating w.r.t. θ_r , and again assuming that we can interchange the order of integration and differentiation,

$$\int \dots \int \hat{\theta}_r \frac{\partial L}{\partial \theta_r} dx_1 \dots dx_n = 1 = \int \dots \int \hat{\theta}_r \frac{\partial \log L}{\partial \theta_r} L dx_1 \dots dx_n$$

Multiplying this by θ_r and subtracting from Eq. 1 gives

$$\int \dots \int \hat{\theta}_r \frac{\partial \log L}{\partial \theta_r} L dx_1 \dots dx_n - \int \dots \int \theta_r \frac{\partial \log L}{\partial \theta_r} L dx_1 \dots dx_n = 1$$

and so

$$\int \dots \int (\hat{\theta}_r - \theta_r) \frac{\partial \log L}{\partial \theta_r} L dx_1 \dots dx_n = 1$$

and equivalently for another model parameter θ_s

$$\int \dots \int (\hat{\theta}_s - \theta_s) \frac{\partial \log L}{\partial \theta_s} L dx_1 \dots dx_n = 1$$

Now, differentiating both w.r.t. X gives

$$(\hat{\theta}_r - \theta_r) \frac{\partial \log L}{\partial \theta_r} L = 0$$

and

$$(\hat{\theta}_s - \theta_s) \frac{\partial \log L}{\partial \theta_s} L = 0$$

and so we observe that the two expressions must at least be proportional (equal to within a multiplicative constant). Therefore, we can apply the Schwartz inequality in its limit (see Appendix 3) using

$$u = \sqrt{(\hat{\theta}_r - \theta_r) \frac{\partial \log L}{\partial \theta_r} L} \quad \text{and} \quad v = \sqrt{(\hat{\theta}_s - \theta_s) \frac{\partial \log L}{\partial \theta_s} L}$$

to obtain

$$1 = \int \dots \int \sqrt{(\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s) \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} L^2} dx_1 \dots dx_n$$

Now apply the Schwartz inequality again using

$$u = \sqrt{(\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s) L} \quad \text{and} \quad v = \sqrt{\frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} L}$$

This gives

$$\left[\int \dots \int (\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s) L dx_1 \dots dx_n \right] \left[\int \dots \int \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} L dx_1 \dots dx_n \right] \geq 1$$

Comparing with equation 2 again, and to the definition for the covariance between two quantities given in Appendix 1, we observe that the first term on the L.H.S. is the covariance of θ_r and θ_s , and that the second term is the expectation value of the product of first differentials of the log-likelihood function w.r.t. these two parameters. Therefore, we can express the inverse of the covariance matrix terms as

$$C_{\theta_r, \theta_s}^{-1} \leq \left\langle \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} \right\rangle$$

In the limit of large n we can replace the expectation value of the estimator with its true value at the maximum

$$C_{\theta_r, \theta_s}^{-1} \leq \left. \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} \right|_{\theta = \theta_{max}}$$

and we can express this in the convenient matrix form

$$C_{\theta}^{-1} \leq (\nabla_{\theta} \log L)^T \otimes (\nabla_{\theta} \log L) \Big|_{\theta=\theta_{max}}$$

This expression therefore provides a minimum bound on the covariance of the parameters. It is variously known as the Minimum Variance Bound (MVB), the Cramer-Rao bound, or the Frechet inequality. The expression will become an equality in the limit of large n for any consistent estimator.

In addition, proceeding from Eq. 1. and differentiating again w.r.t. θ_s we obtain

$$\int \dots \int \left[\frac{\partial \log L}{\partial \theta_r} \frac{\partial L}{\partial \theta_s} + L \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right] dx_1 \dots dx_n = 0$$

so

$$\int \dots \int \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} L dx_1 \dots dx_n = - \int \dots \int \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} L dx_1 \dots dx_n$$

giving

$$\left\langle \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} \right\rangle = - \left\langle \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right\rangle$$

and

$$C_{\theta}^{-1} \leq - \left\langle \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right\rangle = - \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \Big|_{\theta=\theta_{max}}$$

This expression elucidates the behaviour of the terms of the inverse covariance matrix: they are bounded by the second differential of the log-likelihood function with respect to the parameters at its maximum.

2.2 The Condition for Achieving the Minimum Variance Bound

The utility of these expressions is severely limited unless we know the condition under which the equations become equalities i.e. the properties that the log-likelihood function must exhibit in order that we can achieve the minimum variance bound. Consider the first differentials of the log-likelihood function w.r.t. the parameters. Clearly these differentials must be equal to zero at the maximum, so

$$\frac{\partial \log L}{\partial \theta_r} \Big|_{\theta=\theta_{max}} = 0$$

and

$$\frac{\partial \log L}{\partial \theta_s} \Big|_{\theta=\theta_{max}} = 0$$

Now take Taylor expansions about the maximum, postulating that all derivatives higher than the second derivative are zero, so that

$$\frac{\partial \log L}{\partial \theta_r} + (\hat{\theta}_r - \theta_r) \frac{\partial^2 \log L}{\partial \theta_r^2} = 0$$

and

$$\frac{\partial \log L}{\partial \theta_s} + (\hat{\theta}_s - \theta_s) \frac{\partial^2 \log L}{\partial \theta_s^2} = 0$$

are exact. Multiplying these expressions gives

$$\frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} + (\hat{\theta}_r - \theta_r) \frac{\partial^2 \log L}{\partial \theta_r^2} \frac{\partial \log L}{\partial \theta_s} + (\hat{\theta}_s - \theta_s) \frac{\partial^2 \log L}{\partial \theta_s^2} \frac{\partial \log L}{\partial \theta_r} + (\hat{\theta}_r - \theta_r) \frac{\partial^2 \log L}{\partial \theta_r^2} (\hat{\theta}_s - \theta_s) \frac{\partial^2 \log L}{\partial \theta_s^2} = 0$$

Now take expectation values, remembering that expectation values add

$$\langle (f + g) \rangle = \sum_r (f + g)P(r) = \sum_r fP(r) + \sum_r gP(r) = \langle f \rangle + \langle g \rangle$$

but do not multiply

$$\langle (fg) \rangle \neq \langle f \rangle \langle g \rangle$$

unless the quantities f and g are independent. We obtain

$$\begin{aligned} & \left\langle \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} \right\rangle + \left\langle (\hat{\theta}_r - \theta_r) \frac{\partial^2 \log L}{\partial \theta_r^2} \frac{\partial \log L}{\partial \theta_s} \right\rangle + \left\langle (\hat{\theta}_s - \theta_s) \frac{\partial^2 \log L}{\partial \theta_s^2} \frac{\partial \log L}{\partial \theta_r} \right\rangle = \\ & - \left\langle (\hat{\theta}_r - \theta_r) \frac{\partial^2 \log L}{\partial \theta_r^2} (\hat{\theta}_s - \theta_s) \frac{\partial^2 \log L}{\partial \theta_s^2} \right\rangle \end{aligned}$$

Now, since we have assumed that the second differentials of the log-likelihood function w.r.t. the parameters are the highest non-zero differentials, they must be constant and therefore independent of the terms of the form $(\hat{\theta}_s - \theta_s)$. Also observe that the expectation value for terms of this form is zero in the limit of large numbers. Therefore the second and third terms on the L.H.S. have expectation values of zero, and we are left with

$$- \frac{\left\langle \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} \right\rangle}{\left. \frac{\partial^2 \log L}{\partial \theta_r^2} \right|_{\theta=\theta_{max}} \left. \frac{\partial^2 \log L}{\partial \theta_s^2} \right|_{\theta=\theta_{max}}} = \left\langle (\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s) \right\rangle$$

Using the previous result

$$\left\langle \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_r} \right\rangle = - \left\langle \frac{\partial^2 \log L}{\partial \theta_r^2} \right\rangle$$

and remembering that in the limit of large numbers expectation values are replaceable with the true values at the maximum, we can expand the denominator in terms of first differentials, cancel with the numerator, and re-express in terms of second differentials to leave

$$\left\langle (\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s) \right\rangle = - \frac{1}{\left\langle \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right\rangle}$$

which is just the minimum variance bound. The only assumption made was that there were no differentials higher than second order. This implies that the log-likelihood function is quadratic, and thus that exponentiating will produce a Gaussian form. Therefore, the minimum variance bound provides an exact expression for the covariance to whatever extent the log-likelihood function is quadratic, and this will always be a good approximation over a sufficiently small range around the maximum for any smooth and continuous function. From this point onwards, we will modify the previous statement concerning the covariance matrix to state that the inverse covariance matrix is the second differential of the log-likelihood function *approximated as a quadratic* w.r.t. the parameters at its maximum, and assume that we are always working within a sufficiently small range of the maximum that this approximation is satisfied.

2.3 An Example using a Gaussian Likelihood Function

The validity of these results can be demonstrated by assuming a Gaussian form for the likelihood function

$$L = \prod_i A_i e^{-\frac{(I_i - I_M)^2}{2\sigma_i^2}}$$

where I_i are the data, I_M is the model, and A_i is some normalisation constant that depends only on σ_i , the error on measurement i . Taking logs and differentiating gives

$$\log L = \sum_i -\frac{(I_i - I_M)^2}{2\sigma_i^2} + \log A_i$$

and, since the data are not a function of the model parameters,

$$\frac{\partial \log L}{\partial \theta_r} = \sum_i \frac{I_i - I_M}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r}$$

Notice that applying these two operations in turn has removed the normalisation of the original probability function. This will apply to any distribution composed of a shape term and a multiplicative or additive normalisation constant as long as the errors are not correlated with the data. Differentiating again gives

$$\frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} = \sum_i \frac{1}{\sigma_i^2} [(I_i - I_M) \frac{\partial^2 I_M}{\partial \theta_r \partial \theta_s} - \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s}]$$

Since the Gaussian distribution is symmetric about the mean, the term $(I_i - I_M)$ will sum to zero, leaving

$$\left. \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right|_{\theta=\theta_{max}} = \sum_i -\frac{1}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s} \Big|_{\theta=\theta_{max}} \Rightarrow C_\theta^{-1} = \sum_i \frac{1}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s} \Big|_{\theta=\theta_{max}}$$

Comparing to the propagation of errors formula derived in Appendix 1, this is the correct result for a Gaussian distribution, again showing that the minimum variance bound is equal to the covariance for a Gaussian likelihood i.e. a quadratic log-likelihood.

Alternatively, we can proceed from the product of first differentials of the log-likelihood function w.r.t. the parameters to obtain

$$\frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} = \sum_i \frac{I_i - I_M}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \sum_i \frac{I_i - I_M}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_s} = \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s} \sum_i \frac{I_i - I_M}{\sigma_i^2} \sum_i \frac{I_i - I_M}{\sigma_i^2}$$

Expanding the product of the sums will produce two types of terms: squared terms in a single i and cross terms in pairs of the i . Therefore, we can write

$$\sum_i \frac{I_i - I_M}{\sigma_i^2} \sum_i \frac{I_i - I_M}{\sigma_i^2} = \sum_i \left(\frac{I_i - I_M}{\sigma_i^2} \right)^2 + \text{cross terms}$$

The cross terms will have the form

$$\frac{(I_i - I_M)(I_j - I_M)}{\sigma_i^2 \sigma_j^2}$$

i.e. will be the weighted product of two samples from Gaussian distributions, and so will also have a Gaussian distribution. Therefore, their distribution will be symmetric and they will sum to zero. Considering the remaining terms in the expansion, and using the definition for the standard deviation

$$\sum_v (I_v - I_M)^2 = \langle I_v^2 \rangle - \langle I_v \rangle^2 = \sigma_i^2$$

we obtain

$$\sum_i \frac{(I_i - I_M)^2}{\sigma_i^4} = \sum_i \frac{\sigma_i^2}{\sigma_i^4} = \sum_i \frac{1}{\sigma_i^2}$$

and so we have

$$\left. \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} \right|_{\theta=\theta_{max}} = \sum_i \frac{1}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s} \Big|_{\theta=\theta_{max}}$$

This is again the expected result for a Gaussian, confirming that

$$C_\theta^{-1} = \left. \frac{\partial \log L}{\partial \theta_r} \frac{\partial \log L}{\partial \theta_s} \right|_{\theta=\theta_{max}} = - \left. \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right|_{\theta=\theta_{max}}$$

In most cases, and certainly in the case of coregistration where the data consists of images containing of the order of 100,000 voxels, the likelihood function will be built up from a large number of individual data terms. In that case the Central Limit Theorem should ensure that the distribution of the log-likelihood function is Gaussian, even when the probability density function for the individual data terms is not.

3 Covariance Matrices, χ^2 and χ_i

3.1 Covariance Matrices in terms of χ^2

The χ^2 metric² is a true statistic which can be compared across different data sets for equivalent degrees of freedom (the significance of this will be made clear in section 4.3), and is conventionally written [3]

$$\chi^2 = \sum_i \frac{(I_i - I_M)^2}{\sigma_i^2}$$

²The χ^2 used to here is in the *general* sense as statistical tests for assessing the adequacy of fitting results for a number of degrees of freedom, not tests associated with comparing histograms or tables, for which there are *specific* computational forms.

Following the derivations used in Section 2 we can rewrite the results from that section in terms of the χ^2 metric

$$\frac{\partial \chi^2}{\partial \theta_r} = \sum_i -2 \frac{(I_i - I_M)}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \Rightarrow \frac{\partial \chi^2}{\partial \theta_r} \frac{\partial \chi^2}{\partial \theta_s} = \sum_i \frac{4}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s}$$

and

$$\frac{\partial^2 \chi^2}{\partial \theta_r \partial \theta_s} = \sum_i -\frac{2}{\sigma_i^2} [(I_i - I_M) \frac{\partial^2 I_M}{\partial \theta_r \partial \theta_s} - \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s}] = \sum_i \frac{2}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s}$$

allowing us to write

$$C_\theta^{-1} = \frac{1}{4} \frac{\partial \chi^2}{\partial \theta_r} \frac{\partial \chi^2}{\partial \theta_s} \Big|_{\theta=\theta_{max}} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \theta_r \partial \theta_s} \Big|_{\theta=\theta_{max}}$$

It is conventional to remove these factors of two by defining

$$\alpha_{rs} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \theta_r \partial \theta_s} \quad \text{and} \quad \beta_r = -\frac{1}{2} \frac{\partial \chi^2}{\partial \theta_r}$$

in order to facilitate comparisons with the results from Section 2.

Instead, we now take a different route and find an expression for the covariance matrix in terms of the quantity known as the χ of the χ^2 , allowing us to identify the contributions of individual data terms. We can write

$$\chi^2 = \sum_i \frac{(I_i - I_M)^2}{\sigma_i^2} = \sum_i \chi_i^2 \quad \Rightarrow \quad \chi_i = \frac{(I_i - I_M)}{\sigma_i}$$

Therefore

$$\frac{\partial \chi_i}{\partial \theta_r} = -\frac{1}{\sigma_i} \frac{\partial I_M}{\partial \theta_r}$$

and

$$\frac{\partial \chi_i}{\partial \theta_r} \frac{\partial \chi_i}{\partial \theta_s} = \frac{1}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s}$$

allowing us to write

$$\sum_i \frac{\partial \chi_i}{\partial \theta_r} \frac{\partial \chi_i}{\partial \theta_s} = \sum_i \frac{1}{\sigma_i^2} \frac{\partial I_M}{\partial \theta_r} \frac{\partial I_M}{\partial \theta_s}$$

Since the R.H.S. is equal to the previous expression for the covariance matrix of a Gaussian, we can write

$$C_\theta^{-1} = \sum_i (\nabla_{\theta} \chi_i)^T \otimes (\nabla_{\theta} \chi_i) \Big|_{\theta=\theta_{max}}$$

3.2 χ^2 and $\log L$

Note that, under the assumption of a Gaussian likelihood function, the χ of the χ^2 can be generated from the likelihood function relatively easily. We have

$$\chi^2 = \sum_i \frac{(I_i - I_M)^2}{\sigma_i^2} = \sum_i \chi_i^2 \quad \Rightarrow \quad \chi_i^2 = \frac{(I_i - I_M)^2}{\sigma_i^2}$$

and

$$\log L = \sum_i -\frac{(I_i - I_M)^2}{2\sigma_i^2} + \log A_i \quad \Rightarrow \quad \log L_i = -\frac{(I_i - I_M)^2}{2\sigma_i^2} + \log A_i$$

The results in Section 2 demonstrated that the normalisation constants A_i have no effect on the covariance calculation. It is also clear that they will not contribute to any differential process such as optimisation of the likelihood in a maximum likelihood technique. Therefore, we are free to choose any (constant) normalisation of the individual data terms without affecting the final result of either model parameter estimation or covariance calculation. If we choose to normalise such that the peak value of the distribution is 1, rather than the usual normalisation such that the area is 1, we obtain

$$\log L_i = -\frac{(I_i - I_M)^2}{2\sigma_i^2}$$

allowing us to write³

$$\chi^2 = \sum_i \chi_i^2 = \sum_i -2 \log(L_i) \Rightarrow \chi_i = \sqrt{-2 \log L_i}$$

The assumption of a specific normalisation here does not conflict with the previous normalisation assumption in Section 2: normalising the likelihood function to its peak, rather than its area, still results in the integral of the likelihood function over all space being a constant. Therefore, the differentials w.r.t. the data will be equal to zero and the derivations given in Section 2 still hold.

4 The Covariance Matrix for Mutual Information Coregistration

4.1 Mutual Information as a Biased Maximum Likelihood Method

As shown in the Appendices, the mutual information measure is defined as

$$\mathcal{I}(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x) \cdot p(y)}$$

Suppose we have two images I and J and produce an intensity histogram described by a probability distribution $p(i, j)$ where i and j are bin coordinates and the number of samples in that bin is N_{ij} . Following the derivation given in [4] the mutual information of the images I and J can be written

$$\begin{aligned} \mathcal{I}(I; J) &= \sum_{i,j} p(i, j) \log \frac{p(i, j)}{p(i) \cdot p(j)} = \sum_{i,j} p(i, j) \log \frac{1}{p(i)} + p(I, J) \log \frac{p(i, j)}{p(j)} \\ &= \sum_i p(i) \log \frac{1}{p(i)} + \sum_{i,j} p(i, j) \log \frac{p(i, j)}{p(j)} \end{aligned}$$

Recognising the first term as the entropy [8] of image I , and that

$$p(i, j) = \frac{N_{ij}}{N}$$

where N is the total number of voxel pairs in the image pair, we can write

$$N[\mathcal{I}(I; J) - H(I)] = \sum_{i,j} N_{ij} \log p(i|j)$$

or, summing over the voxels in the images instead of over the histogram bins,

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

At this point we must consider the way in which the joint histogram is built from the images. One of the images is fixed, and corresponds to the data in a conventional maximum likelihood method: let this be called the reference image. The other is manipulated by the parameters of the coregistration and therefore corresponds to the model: let this be called the floating image. Each image consists of a sampling of some common scene defined at a set of grid points. As the coregistration proceeds, the floating image is sampled on the grid of the reference image in order to build the joint histogram. Therefore the entropy of the floating image will not be constant. However, since the histogram is built only from the overlapping portions of the images, if regions of the reference image are allowed to pass outside the floating image, then the entropy of the reference image may also change. This circumstance can be avoided by defining a border around the reference image that represents the maximum expected movement of the floating image, and not sampling the data in that region. The entropy and marginal probability distribution of the reference image will then be fixed⁴, We can then make the arbitrary definition that I is the reference image, and write

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

³Note that this is **not** the same as writing

$$\chi = \sqrt{\sum_i -\log L_i}$$

as we are taking the square root over individual terms, not over the sum.

⁴The number of voxels from which the joint histogram is built will also be fixed: this is essential for maximum likelihood techniques.

Clearly, the constant and the multiplicative term N can have no effect on differential processes, including both optimisation and covariance matrix estimation. Therefore, the mutual information measure is a monotonic function of the log-likelihood of image I given image J , implying that mutual information coregistration is in fact a maximum-likelihood process.

4.2 Bias in the Mutual Information Metric

For all processes that involve only derivatives of the mutual information measure, including both covariance estimation and optimisation in a coregistration process, we can ignore the additive and multiplicative constant and concern ourselves only with the term

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

In mutual information coregistration, we are using a so called “bootstrapped” likelihood, where the likelihood is estimated from the behaviour of the joint histogram, rather than from a model in some explicit functional form. Quantitative analysis of the variation of the log-likelihood around the minimum generally requires that the normalisation of the likelihood distribution is fixed for each data point. However, changes in the sample of data and coregistration parameters guarantee that the normalisation will change during co-registration and that the shape (specifically width) of the distribution will also change. We must therefore choose a normalisation for the data which can be expected to be quantitatively equivalent during the optimisation process and “correct” at the minimum. Where we define correct here to be equivalent to a likelihood normalisation, as that is the one for which we know we can define a covariance. Fortunately, the solution is immediately accessible in the form of the χ^2 metric. From the discussion in Section 3, we can convert the likelihood to a χ^2 metric simply by imposing a new normalisation to the peak, rather than the area, of the distribution. Therefore, the metric we wish to use becomes

$$\sum_v \log \frac{p(i, j)}{p(i_{max}|j)} = \sum_v \log \frac{p(i, j)}{p(i_{max}, j)}$$

where $p(i_{max}|j)$ is the peak of the probability distribution along some row of the joint histogram specified by a given j .

The actual effect this will have on choosing to minimise a function like mutual information is difficult to assess, but the main point can be illustrated using a simple model which deliberately adjusts the width of the sample distribution. The difference between the two statistics is illustrated in Fig.1. A joint histogram was generated from 2000 random samples from a 2D Gaussian distribution, shown in Fig.1(a). Log-likelihoods were calculated from the data normalised both to the peak (i.e. a χ^2 statistic) and the area of the distributions i.e.

$$\log L_{area} = \sum \log \frac{p(i, j)}{p(j)} \quad \text{and} \quad \log L_{peak} = \sum \log \frac{p(i, j)}{p(i_{max}, j)}$$

Both forms were calculated from joint histograms of various bin sizes. Fig.1(b) shows the log-likelihoods, plotted against bin size quoted in terms of the standard deviation of the original Gaussian distribution. The upper curve shows $\log L_{peak}$. It is constant with bin size, varying only at the point that the bins become so small that they contain on average less than one data point each, so that it becomes impossible to estimate the peak height reliably. The lower curve shows $\log L_{area}$, which is clearly dependent on bin size over the entire range. Such effects would be expected during the minimisation of a mutual information measure and in the variation of that measure around its optimum. This could produce anomalous estimates of covariance if the Mutual Information measure were taken as a quantitative estimate of the likelihood.

The χ^2 metric can be related back to mutual information by splitting the expression

$$\sum_v \log \frac{p(i, j)}{p(j)} = \sum_v \log \frac{p(i, j)p(i_{max}, j)}{p(i_{max}, j)p(j)} = \sum_v \log \frac{p(i, j)}{p(i_{max}, j)} + \log \frac{p(i_{max}, j)}{p(j)} \quad (3)$$

The first term on the R.H.S. is the χ^2 metric. The second term optimises the “peakiness” of the joint histogram in order to maximise the correlation between equivalent structures in the image pair. Since covariance estimates are not dependent on the normalisation of the likelihoods used, the covariance estimate based on the first term will be correct to whatever extent the second term is constant near the optimum. Since the second term is dependent only on positions of the peak of the distribution and on the marginal in the floating image, it would be expected to vary more slowly than the first term in cases where the data is sufficient to support a stable estimate of the

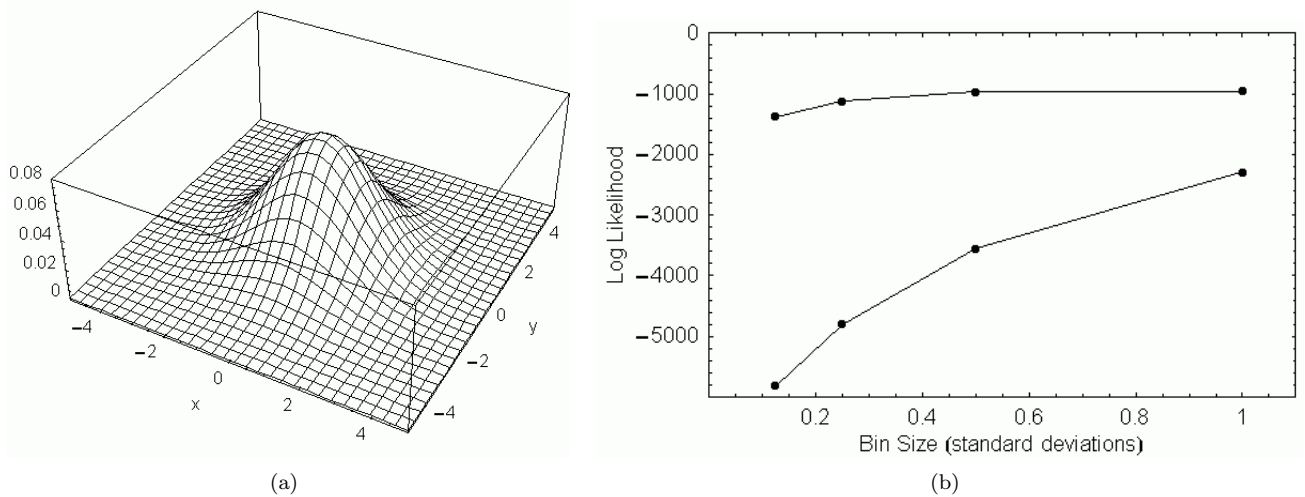


Figure 1: Log-likelihoods (b) calculated for a 2D Gaussian distribution joint histogram (a). The upper curve in (b) shows the log-likelihood normalised to the distribution peak, and the lower curve the log-likelihood normalised to the distribution area. As the bin size of the joint histogram is changed, the likelihood normalised to the area also changes, whereas that normalised to the peak is roughly constant up to the point where there are too few samples available to populate the histogram, and thus the peak estimation begins to fail. See main text for explanation.

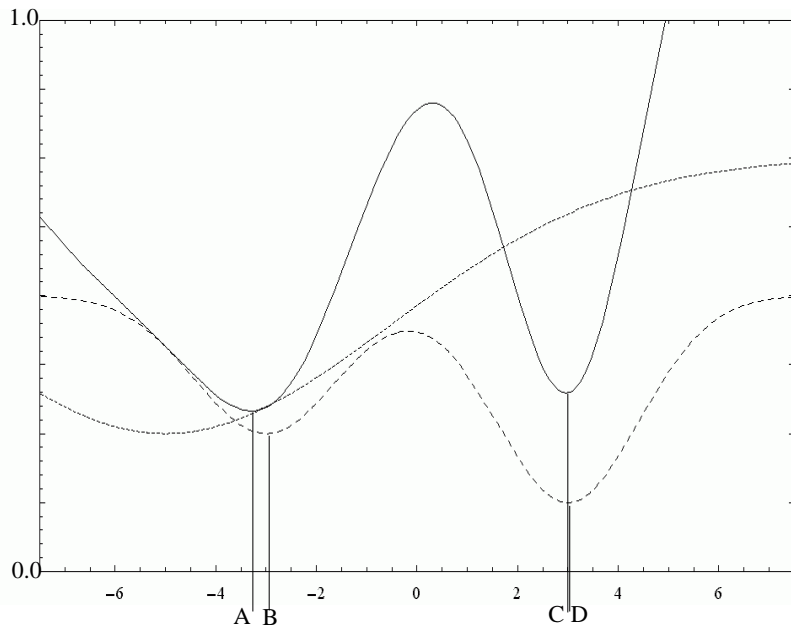


Figure 2: Potential results of the inclusion of a non-uniform bias term, demonstrated using some Gaussian distributions. If the dashed line represents the negative likelihood, and the dotted line represents a non-uniform bias, then the product is given by the solid line. The original likelihood had two minima, points B and D, of which D was the global minimum. After multiplication by the prior, both minima have been biased to lower values and now point A, the biased equivalent of B, has become the global minimum.

covariance parameters. This is likely to be true in the case of rigid coregistration, but may be questionable in the case of non-rigid coregistration with large numbers of parameters.

In the general case, the second term may change as the coregistration parameters are manipulated. Since it is non-uniform, this term has the capability to bias the results or even, in the case of multiple local optima, to change which of the optima is the global optimum, as shown in Fig. 1. Situations of this kind are not uncommon in medical image analysis [7]. In general, it is not acceptable to minimise a biased quantity, since this will produce biased estimates of the optimal parameters. If the bias represents useful information e.g. priors, then the correct

approach is to minimise the true statistic and use the priors to select between competing local or global minima at the end of the optimisation.

4.3 Computing Covariances for Bootstrapped Likelihoods

Having identified a suitable χ^2 metric for the coregistration, normalised in a way that is consistent with the results in Section 3, we can proceed to calculate the covariance using the χ of the χ^2 . Starting from

$$\chi_v = \sqrt{-2 \log L_v}$$

and

$$C_\theta^{-1} = \sum_v (\nabla_\theta \chi_v)^T \otimes (\nabla_\theta \chi_v) \Big|_{\theta=\theta_{max}}$$

we can expand the latter using the chain rule in terms of derivatives w.r.t. the likelihood L_v and the image J_v giving

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

and using

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

we can write

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

This can be considered as a general result for the calculation of covariances on parameters θ for any image bootstrapped likelihood, where the likelihood has been generated from data terms with probabilities normalised such that their peak value is unity.

This result can now be applied to the modified mutual information metric

$$L_v = \frac{p(i, j)}{p(i_{max}, j)}$$

This gives

$$\frac{\partial L_v}{\partial J_v} = \frac{1}{p(i_{max}, j)} \left[\frac{\partial p(i, j)}{\partial J_v} - \frac{p(i, j)}{p(i_{max}, j)} \frac{\partial p(i_{max}, j)}{\partial J_v} \right]$$

and

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

The validity of this procedure can be tested by assuming a simple Gaussian model for the conditional probability, remembering to implement the required normalisation,

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

so

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

giving a covariance estimate of

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

the familiar result for the covariance of a Gaussian likelihood distribution.

5 Conclusion

This document has provided a derivation of a discrete solution for the covariance matrix for mutual information (MI) coregistration. Knowledge of measurement errors is essential for any scientific measurement, but we believe that this expression will be particularly useful for non-rigid coregistration techniques, where the errors will vary across the data. In that case, and particularly where further analysis purports to use pixel-by-pixel differences between coregistered image pairs, knowledge of the covariance matrix is essential in order to express the extent to which displacement errors between voxels in uniform regions of the image data allow analyses of this type.

The approach has been to identify the MI metric as a biased form of a maximum likelihood technique. We have explained that the use of such bias terms may be justified for the selection of candidate minima but not for parameter estimation. Removal of the bias term, by identification with χ^2 measures, allows the use of standard formulae related to covariances for maximum likelihood techniques. The covariance matrix is then valid for standard MI coregistration to whatever extent the bias term has zero derivative around the optimal coregistration parameters. Since the bias term depends only on the marginal distribution, the variation around the optimum would be expected to be small, especially for medical data sets, which typically feature large areas of uniform grey-levels. Therefore we anticipate that the results presented here should be applicable to standard MI coregistration.

Identification of the relationship between MI and maximum likelihood also raises the issue of the theoretical basis of the technique. The MI metric is usually justified in terms of Shannon entropy, as shown in the Appendices. However, we believe that its true theoretical justification lies in its connection to likelihood and statistics, and indeed Shannon's concepts of entropy were partially based on Fisher's concept of information [1]

$$I(\theta_s) = \left\langle \left(\frac{\partial \log L}{\partial \theta_s} \right)^2 \right\rangle$$

which is the denominator in the expression for the MVB on a maximum likelihood estimator given in Section 2.

We can also make some general comments on covariance estimation in maximum likelihood techniques. Most important is the observation that the normalisation of the probability distributions used to generate the likelihood function has no effect on the covariance calculation as long as the normalisation has zero derivative w.r.t. the model parameters at the optimum. It should be noted that the process of optimising the likelihood function w.r.t. the parameters of the model and the process of covariance estimation, where we are measuring the variation of the likelihood function around the optimum, are two independent processes. This provides some freedom in the choice of normalisation, which has been used here to correct for the bias resulting from the use of a bootstrapped likelihood, thus allowing step-to-step comparison of values of the optimisation metric across the optimisation. The requirement of a quadratic log-likelihood function for achieving the Minimum Variance Bound implies that we should always work in an equal variance space: [5] and [6] discuss some common variance-normalising transforms.

References

- [1] R.J. Barlow. *Statistics: A Guide to the use of Statistical Methods in the Physical Sciences*. John Wiley and Sons, U.K., 1989.
- [2] K. Ord and S. Arnold. *Kendall's Advanced Theory of Statistics: Classical Inference*. Arnold, 1998.
- [3] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd. Ed. Cambridge University Press, 1992.
- [4] A. Roche, G. Malandain, N. Ayache and S. Prim. Towards a Better Comprehension of Similarity Measures used in Medical Image Registration. Proc. MICCAI 1999, p555-566, Springer-Verlag, 1999.
- [5] P.A. Bromiley and N.A. Thacker. The Effects of a Square Root Transform on a Poisson Distributed Quantity. TINA Memo No. 2001-010, <http://www.tina-vision.net/docs/memos.php>, 2001.
- [6] P.A. Bromiley and N.A. Thacker. The Effects of an Arcsin Square Root Transform on a Binomial Distributed Quantity. TINA Memo No. 2002-007, <http://www.tina-vision.net/docs/memos.php>, 2002.
- [7] P.A. Bromiley, N.A. Thacker, M.L.J. Scott, M. Pokrić, A.J. Lacey and T.F. Cootes. Bayesian and Non-Bayesian Probabilistic Models for Medical Image Analysis. TINA Memo No. 2001-014, <http://www.tina-vision.net/docs/memos.php>, 2001. To be published in *Image and Vision Computing*, 2003.
- [8] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications, John Wiley and Sons, U.K., 1991.

A Basic Definitions for Covariance Matrices

Starting from the most basic definition, the covariance of a pair of quantities x and y can be calculated from a set of measurements x_i and y_i as

$$\text{cov}(x_i, y_i) = \overline{x_i \cdot y_i} - \overline{x_i} \cdot \overline{y_i}$$

If there are more than two variables, the set of covariances can be represented as a matrix \mathbf{V} where

$$V_{ij} = \text{cov}(x_i, x_j)$$

in which the diagonal elements are the variances of the individual variables

$$V_{ii} = \overline{(x_i^2)} - (\overline{x_i})^2 = \sigma_i^2$$

where σ is the standard deviation. For a joint distribution function we can write the equivalent

$$\text{cov}(x_i, x_j) = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle = \langle x_i x_j \rangle - \mu_i \mu_j$$

where μ_i is the mean of x_i , and $\langle x_i \rangle$ represents the expectation value of x_i . Now, suppose we have a set of m functions f_m of n variables x_n . Even if the variables are independent, the functions in general will not be since they depend on the same variables. Using the above definitions the variances of the functions can be written as

$$V(f_i) = \langle f_i^2 \rangle - \langle f_i \rangle^2$$

Expanding the f_i in a Taylor series gives

$$f_i = f_i(\mu_1, \mu_2, \dots, \mu_n) + \left(\frac{\partial f_i}{\partial x_1}\right)(x_1 - \mu_1) + \left(\frac{\partial f_i}{\partial x_2}\right)(x_2 - \mu_2) \dots$$

so

$$\begin{aligned} V(f_i) &= \left(\frac{\partial f_i}{\partial x_1}\right)^2 \langle (x_1 - \mu_1)^2 \rangle + \dots + 2\left(\frac{\partial f_i}{\partial x_1}\right)\left(\frac{\partial f_i}{\partial x_2}\right) \langle (x_1 - \mu_1)(x_2 - \mu_2) \rangle + \dots \\ &= \sum_j \left(\frac{\partial f_i}{\partial x_j}\right)^2 V(x_j) + \sum_j \sum_{k \neq j} \left(\frac{\partial f_i}{\partial x_j}\right)\left(\frac{\partial f_i}{\partial x_k}\right) \text{cov}(x_j, x_k) \end{aligned}$$

The covariances can be found in the same way

$$\langle f_k f_l \rangle - \langle f_k \rangle \langle f_l \rangle = \langle (x_1 - \mu_1)(x_1 - \mu_1) \rangle \left(\frac{\partial f_k}{\partial x_1}\right)\left(\frac{\partial f_l}{\partial x_1}\right) + \dots + \langle (x_1 - \mu_1)(x_2 - \mu_2) \rangle \left(\frac{\partial f_k}{\partial x_1}\right)\left(\frac{\partial f_l}{\partial x_2}\right) + \dots$$

or

$$\text{cov}(f_k, f_l) = \sum_i \sum_j \left(\frac{\partial f_k}{\partial x_i}\right)\left(\frac{\partial f_l}{\partial x_j}\right) \text{cov}(x_i, x_j)$$

So using the notation

$$G_{ki} = \left(\frac{\partial f_k}{\partial x_i}\right)$$

the covariance matrix \mathbf{V}_f of the functions f can be related to the covariance matrix \mathbf{V}_x of the variables x by

$$\mathbf{V}_f = \mathbf{G}^T \mathbf{V}_x \mathbf{G}$$

This is the most general form of the propagation of errors formula.

B Entropy, Kullback-Leibler Divergence, and Mutual Information

This section summarises some basic definitions and results related to the concepts of entropy and mutual information. More complete treatments can be found in [8]. It will use the following results from basic probability theory:

$$\text{Marginal probability: } P(x) = \sum_y P(x, y)$$

$$\text{Conditional probability: } P(x|y) = \frac{P(x, y)}{P(y)}$$

$$\text{Product rule: } P(x, y) = P(x|y)p(y)$$

$$\text{Sum rule: } P(x) = \sum_y P(x, y) = \sum_y P(x|y)p(y)$$

The entropy $H(X)$ of $X = x_1 \dots x_n$ is defined as

$$H(X) = \sum_x P(x) \log \frac{1}{P(x)}$$

which extends immediately to a joint distribution

$$H(X, Y) = \sum_{x, y} P(x, y) \log \frac{1}{P(x, y)}$$

and to a conditional distribution $P(X|y = y_i)$

$$H(X|y) = \sum_{x, y} P(x|y) \log \frac{1}{P(x|y)}$$

The conditional entropy of X given Y is then the average of the above, weighted by the probability $P(y)$

$$H(X|Y) = \sum_y P(y) \left[\sum_x P(x|y) \log \frac{1}{P(x|y)} \right] = \sum_{x, y} P(x, y) \log \frac{1}{P(x|y)}$$

Starting from the expression for a joint distribution

$$H(X, Y) = \sum_{x, y} P(x, y) \log \frac{1}{P(x, y)}$$

we can apply the product rule $P(x, y) = P(x|y)P(y)$ to obtain

$$H(X, Y) = \sum_{x, y} P(x, y) \left[\log \frac{1}{P(y|x)} + \log \frac{1}{P(x)} \right]$$

and then apply the product rule $P(x, y) = (P(y|x)P(x))$ to obtain

$$H(X, Y) = \sum_{x, y} P(x, y) \log \frac{1}{P(y|x)} + \sum_{x, y} P(x|y)P(y) \log \frac{1}{P(x)}$$

Then using the sum rule

$$\sum_y P(x|y)P(y) = P(x)$$

we obtain

$$H(X, Y) = \sum_{x, y} P(x, y) \log \frac{1}{P(y|x)} + \sum_x P(x) \log \frac{1}{P(x)} = H(Y|X) + H(X)$$

This expression is known as the chain rule for entropy

$$H(X, Y) = H(Y|X) + H(X) = H(X|Y) + H(Y)$$

The Kullback-Leibler divergence [8] is a well-known measure of the difference between two probability distributions $P(x)$ and $Q(x)$ defined over the same x

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Note that in the general case it is not symmetric under interchange of $P(x)$ and $Q(x)$ and so is not a true distance metric. The K-L divergence obeys Gibb's inequality

$$D_{KL}(P||Q) \geq 0$$

$$D_{KL}(P||Q) = 0 \quad \text{if} \quad P \equiv Q$$

and so the measure grows as $P(x)$ and $Q(x)$ diverge.

Given two random variates x and y the K-L divergence can be used to measure their degree of independence by comparing the joint distribution $P(x, y)$ to the product of the marginal distributions $P(x).P(y)$ since the two would be equal if $P(x)$ and $P(y)$ were independent

$$D_{KL}(P(x, y)||P(x).P(y)) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x).P(y)}$$

As before, the measure will grow as the two variates become more dependent. Therefore, if $P(x)$ and $P(y)$ represent the probability distribution of two images defined over the same scene, the K-L divergence in this form can be maximised as a function of some coregistration parameters to achieve alignment. Note that in this formulation the K-L divergence is symmetric under interchange of $P(x)$ and $P(y)$. It can also be linked back to the concept of entropy

$$\begin{aligned} D_{KL}(P(x, y)||P(x).P(y)) &= \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x).P(y)} = \sum_{x,y} P(x, y) \log \frac{P(x|y)}{P(x)} \\ &= \sum_{x,y} P(x, y) \log \frac{1}{P(x)} - \sum_{x,y} P(x, y) \log \frac{1}{P(x|y)} = \sum_x P(x) \log \frac{1}{P(x)} - \sum_{x,y} P(x, y) \log \frac{1}{P(x|y)} \end{aligned}$$

so

$$D_{KL}(P(x, y)||P(x).P(y)) = H(X) - H(X|Y)$$

This expression is also called the mutual information $\mathcal{I}(X; Y)$ of X and Y , since it measures the average reduction in the uncertainty of X that arises from knowing the value of Y . Note that it is still symmetric under interchange of X and Y . Using the chain rule we can also write

$$\mathcal{I}(X; Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y)$$

and this is the form in which the mutual information measure is most easily implemented in a coregistration routine, since it requires only sums over a joint image histogram.

C Proof of the Schwartz Inequality

Following the procedure suggested in [1], consider the integral

$$\int (\lambda u + v)^2 dX = \lambda^2 \int u^2 dX + 2\lambda \int uv dX + \int v^2 dX$$

where λ is a constant and u and v are both functions of the X . From the L.H.S. it is clear that the integral must be positive or zero whatever the value of λ . Therefore, there can be no non-zero roots in lambda. Applying the quadratic formula

$$b^2 - 4ac \leq 1$$

for no non-zero roots, where

$$a = \int u^2 dX \quad b = 2 \int uv dX \quad c = \int v^2 dX$$

gives

$$\int u^2 dX \int v^2 dX \geq (\int uv dX)^2$$

This is known as the Schwartz inequality. The condition for equality is that u is proportional to v for all X , which can be demonstrated by putting $v = au$ where a is the constant of proportionality, giving

$$\begin{aligned} \int u^2 dX \int (au)^2 dX &= a^2 (\int u dX)^2 \\ (\int au^2 dX)^2 &= a^2 (\int u dX)^2 \end{aligned}$$