The Use of Transinformation in the Design of Data Sampling Schemes for Inverse Problems

S P Luttrell
Royal Signals and Radar Establishment, St Andrews Road, Great Malvern, Worcs, WR14 3PS, UK

We analyse the average useful information content of data samples by using the transinformation entropy (rate of transmission) of Shannon’s information theory. We derive a simple expression for the transinformation in linear experiments with gaussian a priori distributions. We use this expression to examine various schemes for sampling the image spaces of a translation invariant (sine) and a conformally invariant (Laplace) mapping. The optimum sampling scheme is found to be considerably better than the naive sampling scheme (e.g. Nyquist) when the number of samples is small and the a priori knowledge is non-trivial.

I. INTRODUCTION

The mathematical theory of communication as formulated principally by Shannon [1,2] is directly applicable to the measurement of information acquired by an experimenter in the course of conducting an experiment. We choose the Shannon information measure because it is unique in satisfying the axioms one assigns to additive information. Accordingly we shall regard an experiment as a mapping $T$ (which may be non-linear) from a state $f$ of the system (s) under study to a state $g$ of the data samples ($d$). The mapping will usually involve a random noise element. Let us introduce the a priori system probability distribution function (PDF) $P_s[f]$, and the conditional data PDF $P_d[g|f]$. We may then obtain the unconditional data PDF $P_d[g]$, the a posteriori system PDF $P_s[d|f,g]$, and the joint system/data PDF $P_{s,d}[f,g]$. $P_{s,d}$ summarises our state of knowledge of the current system state $f$ bearing in mind the current data observation $g$ and our a priori knowledge $P_s[f]$. Formally the ensemble of possible joint system/data states and the measure of the corresponding joint PDF forms a probability measure space [3]. The joint entropy (measured in bits) of the system and data is defined as [1,2]

$$H[s,d] = - \int [df] [dg] P_{s,d}[f,g] \log_2 \{ P_{s,d}[f,g] \}$$

where we are taking $f$ and $g$ to be continuous. Note that the argument of the logarithm depends on the density of elementary events. The useful information content of the data is given by the rate of transmission $I$ (or transinformation; see [3]) defined as [1,2]

$$I \equiv H[s] - H[s|d]$$

$$= H[d] - H[d|s]$$

$$= H[s] + H[d] - H[s,d]$$

where the entropies $H[s], H[d], H[s|d]$ and $H[d|s]$ are obtained from Equation (1.1) by suitably modifying the PDF inside the logarithm. $I$ does not depend on the density of elementary events [3], as required for a consistent measure of useful information content. The transinformation is the entropy of the source (system) as reduced by the conditional entropy characterising the channel of communication (data acquisition scheme) [4].

The simple situation outlined above assumes that prior knowledge in the form of a priori PDFs actually exists. This will be the case if we have a model of the system, and the model parameters ($f$ in Equation (1.1)) are assigned their a priori distributions. However, there is a large class of problems in which the roles of a priori information and information contained in data are inextricably intertwined. Such situations occur when the model of the system under observation is updated in the light of data already acquired; thus the status of prior data is elevated to that of prior knowledge. Our analysis does not apply to such adaptive models with any rigour.

With these restrictions in mind we shall explore how maximising transinformation optimises data acquisition in a variety of simple situations. In Section II we interpret what such an optimisation achieves. In Section III we introduce a simple model for which the transinformation may be calculated easily, and we show how the singular value decomposition method may be used to analyse this type of transinformation. In Section IV, Section V and Section VI we analyse specific examples where the model of Section III may be applied. The shift-invariant case is studied in Section V and its generalisation to include a support constraint on the object is studied in Section VI. The (conformally invariant) Laplace mapping with an object support constraint is studied in Section VI.

II. OPTIMISING DATA ACQUISITION

We have introduced the transinformation $I$ in order to measure the average additional (useful) information supplied by the sampled output of an experiment. We maximise $I$ in order to optimise the experiment with respect to the average amount of useful information which is acquired [5]. Note that such an optimisation procedure is designed to maximise the amount of useful information as measured with respect to the stated prior knowledge...

If in fact the prior knowledge is incomplete or incorrect, then the optimum data acquisition scheme will not be found. Also, the whole of the data set must be taken into consideration when calculating transinformation. This is important if we make more than one type of measurement on the system, by using a multispectral scanner for instance: the data from all the wavebands must be considered together when calculating the transinformation. The detail of this procedure is highly problem dependent, but the universality of the principle of maximising the transinformation (once it is measured) is preserved. However, we shall see that there are many situations in which this whole procedure may be carried out without difficulty.

We commented in Section 1 that $I$ does not depend on the density of elementary events in a continuous probability measure space. The variation of $I$ with sample position arises because the chosen sampling scheme (which includes the data quantisation scheme, i.e. the discretisation of the data values used by the experimenter) defines a mapping from an infinite-dimensional continuous probability measure space to a finite-dimensional discrete probability measure space. The particular discrete measure which is mapped to, and hence the transinformation, depends on the sampling scheme. Note that we shall assume that the quantisation levels are spaced sufficiently finely that they may be regarded as a continuum of levels, and so the transinformation will be unaffected by the exact details of the density of levels. However, a similar program of transinformation maximisation could be carried out in which the positions of the quantisation levels are varied. Note that maximising transinformation affects only the data acquisition stage of an experiment: it places no constraints on how the data should be interpreted.

It is instructive to interpret what sample position optimisation achieves. One extreme case is when the signal component of the sample values is highly redundant (loosely speaking, the signal is oversampled). Therefore there is not much potential signal information present (undesirable), but it is robust to the disruptive effects of data noise (desirable). At the other extreme we could place the samples so that the signal component of the sample values has very little redundancy. Therefore there is a lot of potential signal information present (desirable), but it is not robust to the disruptive effects of noise (undesirable). We would like to have the advantages of both sampling schemes, but sadly they are mutually exclusive: we must accept a trade-off between the two extremes. We achieve this by maximising the useful Shannon information contained in the data samples, which may be deduced from $P_s[f]$ and $P_{d,s}[g,f]$. This requires that sample positions must be chosen to ‘match’ the prior knowledge. The basic inequalities which summarise the effect of redundancy (and/or independence) on information content are

$$H[x] \text{ and } H[y] \leq H[x,y] \leq H[x] + H[y]$$  \hspace{1cm} (2.1)
The unconditional data PDF

\[ g = T f + n \]  \hspace{1cm} (3.1)\

where \( n \) is additive data noise in \( G \) which we assume to be distributed independently of the signal. The simplest prior PDFs which give non-trivial results are gaussian; these may be used to specify means and covariance alone. Accordingly we shall define \( P_s \) and \( P_{d|s} \) as

\[ P_s[f] = \frac{1}{\text{det}[\pi W]} \exp[-(f - f_0)^\dagger W^{-1} (f - f_0)] \]  \hspace{1cm} (3.2)\
\[ P_{d|s}[g|f] = \frac{1}{\text{det}[\pi N]} \exp[-(g - T f)^\dagger N^{-1} (g - T f)] \]  \hspace{1cm} (3.3)\

where \( W \) and \( N \) are covariance matrices

\[ W \equiv \langle (f - f_0) (f - f_0)^\dagger \rangle \]  \hspace{1cm} (3.4)\
\[ N \equiv \langle n n^\dagger \rangle \]  \hspace{1cm} (3.5)\

and \( f_0 \) is a mean

\[ f_0 \equiv \langle f \rangle \]  \hspace{1cm} (3.6)\

The choice of \( P_{d|s} \) in Equation 3.3 then specifies that the data have additive, zero mean, correlated gaussian noise present. The position of the mean of the additive data noise does not affect the information content of the data, and so we assume it to be zero for simplicity. We have modelled \( P_s \) in Equation 3.2 by using a gaussian PDF, which is more ad hoc than modelling data noise as gaussian random variables. However, if we have the prior knowledge that the possible system states \( f \) are clustered around state \( f_0 \) and that the covariance of this cluster is \( W \), then the chosen \( P_s \) is the least committal model [7]. The unconditional data PDF \( P_d[g] \) may be derived as

\[ P_d[g] = \frac{1}{\text{det}[\pi \overline{W}]} \exp[-(g - g_0)^\dagger \overline{W}^{-1} (g - g_0)] \]  \hspace{1cm} (3.7)\

where \( \overline{W} \) is the covariance matrix

\[ \overline{W} \equiv \langle (g - g_0) (g - g_0)^\dagger \rangle \]  \hspace{1cm} (3.8)\
\[ = TW T^\dagger + N \]\

and \( g_0 \) is the mean

\[ g_0 \equiv \langle g \rangle \]  \hspace{1cm} (3.9)\
\[ = T f_0 \]\

Again the position of the mean of \( P_d \) does not affect the information content of the data, and so we could have set \( f_0 = 0 \) in Equation 3.2.

We shall use Equation 1.3 to calculate \( I \). The entropies \( H[d] \) and \( H[d|s] \) are given by

\[ H[d] = \log_2(\pi + N) \]  \hspace{1cm} (3.10)\
\[ H[d|s] = \log_2(\pi N) \]  \hspace{1cm} (3.11)\

which gives the following expression for the transinformation:

\[ I = \log_2 \frac{\text{det}[TW T^\dagger + N]}{\text{det}[N]} \]  \hspace{1cm} (3.12)\

This is the central theoretical result [5] from which we shall calculate the transinformation in a variety of situations. The interpretation of Equation 3.12 is that the ratio of the determinants gives the average boost in the volume of data space within which the data can be found when the signal \( T f \) is added to the data noise \( n \). This ratio is directly related to the ability of the experimenter to discriminate which state \( f \) the system currently occupies. We could have anticipated this interpretation of \( I \) on more general grounds from Equation 1.3. If the noise entropy \( H[d|s] \) is held constant, then \( I \) increases with \( H[d] \). However, \( H[d] \) is a measure of the logarithm of the occupied volume of data space, and so the form of Equation 3.12 is not surprising.

In the following sections we shall obtain some results for particular types of \( T \). We shall assume there that the additive data noise is an isotropic tensor

\[ N_{a,b} \equiv \pi \delta_{a,b} \]  \hspace{1cm} (3.13)\
\[ N = \text{constant} \]

so that Equation 3.12 may be written in the form

\[ I \equiv \log_2(\text{det}[M]) \]  \hspace{1cm} (3.14)\

where

\[ M_{a,b} \equiv R_{a,b} + \delta_{a,b} \]  \hspace{1cm} (3.15)\

\[ R_{a,b} \] is the covariance matrix of the signal component of the data with the convention that the data are scaled so that the additive white noise has unit variance. Some expressions for transinformation which are derived from these formulae are contained in Appendix VIII. We shall use the following factorisation of \( R_{a,b} \)

\[ R_{a,b} \equiv R r_{a,b} \]  \hspace{1cm} (3.16)\

where \( R \) is a suitably defined signal-to-noise ratio (SNR).

We shall now examine how the expression for \( I \) given in Equation 3.12 may be decomposed into a sum over independent contributions to the transinformation. To achieve this we need to diagonalise simultaneously the matrices \( TW T^\dagger + N \) and \( N \); this is just a simultaneous Karhunen-Loève expansion. Let the diagonal elements of the matrices be given by

\[ TW T^\dagger + N \longrightarrow \text{diag} \{\lambda_1, \lambda_2, \cdots \} \]  \hspace{1cm} (3.17)\
\[ N \longrightarrow \text{diag} \{\nu_1, \nu_2, \cdots \} \]  \hspace{1cm} (3.18)
then $I$ may be expressed as

$$I \equiv \sum_a I_a$$

(3.19)

where

$$I_a = \log_2 \left( \frac{\lambda_a}{\nu_a} \right)$$

(3.20)

and $a$ labels the transformed data states. In general these states are not orthogonal and so we must be careful to use distinct covariant and contravariant components. However, the states do carry independent transinformation $I$, and so we have decomposed the mapping $T$ into a number of independent effective communication channels.

This analysis formalises the rationale behind part of the singular-value decomposition (SVD) method of data interpretation; namely the data decomposition step of the SVD. The SVD may be used to perform the diagonalisation in Equation 3.17 and Equation 3.18. Such an analysis has been performed for the sine imaging operation [6] [8-10], for the (sinc)$^2$ imaging operation [11], and for the Laplace transform [12,13]. The theoretical tools from which all these applications may be derived are contained in reference [6]. We assumed a specific pair of distributions $P_s$ and $P_{d_{ab}}$, and we assumed a linear mapping $T$ in order to derive Equation 3.12 in general the result is not so simple. The SVD rigorously decouples the data space $G$ into pieces carrying independent transinformation if, and only if, the two-point correlation function (and the mean) completely specifies the statistical properties of the data, i.e. when all higher-order correlations may be deduced from the two-point correlations. Thus the data PDFs (signal and signal with noise) must be gaussian for the SVD to provide a decomposition into independent partial transinformations. The basic reason for this restriction is that the SVD, which is a linear transformation, can diagonalise only two-point correlations.

The SVD is also used extensively (same references as the previous paragraph) to provide a means of inverting data to recover the object under observation: this corresponds to completing the interpretation of the data, which began with a decomposition using the SVD as discussed above. The object and image singular functions (or states) and the corresponding singular values provide a means for tracing how information passes through $T$. For situations where the SVD provides a rigorous decomposition of the transinformation (i.e. gaussian a priori PDFs) the object reconstruction which is obtained is the one which has the maximum a posteriori probability density [6]. For situations where the a priori PDFs are not all gaussian, the separate singular functions do not in general contain independent information, and the object reconstruction does not have such a simple interpretation. Nevertheless, the SVD method proves to be a very powerful analytical tool, as long as results obtained by using it are interpreted with care.

IV. SHIFT-INvariant MAPPING

Image formation may be modelled, to a first approximation, by a shift-invariant (isoplanatic) linear mapping $T$. For simplicity we shall restrict our attention to one-dimensional problems with a continuous object coordinate $x$ where $-\infty < x < +\infty$ and a discrete image coordinate $x_a$ where $a = 1, 2, \cdots, m$ (i.e. $m$ data samples). Furthermore, we shall assume that the covariance matrices $W$ and $N$ (Equation 3.4 and Equation 3.5) are isotropic tensors

$$W(x, y) \equiv W \delta(x - y)$$

$W = \text{constant}$

(4.1)

$$N_{a,b} \equiv N \delta_{a,b}$$

$N = \text{constant}$

(4.2)

The matrix elements of $TW T^\dagger$ are then given by

$$TW T^\dagger_{a,b} \equiv W \int_{-\infty}^{+\infty} dz T(x_a - z) T^\ast(x_b - z)$$

$$= \frac{W}{2\pi} \int_{-\infty}^{+\infty} dk \exp[i k (x_a - x_b)] \left| \hat{T}(k) \right|^2$$

(4.3)

where $\hat{T}$ is the Fourier transform of $T$ given by

$$\hat{T}(k) \equiv \int_{-\infty}^{+\infty} dz \exp[-i k z] T(z)$$

(4.4)

Note that $TW T^\dagger_{a,b}$ in Equation 4.3 is a function of the difference in coordinates, $x_a - x_b$, alone, and so the transinformation in Equation 3.12 is unchanged if all the data sample positions are displaced by the same amount. We shall define the signal-to-noise ratio of the system $R$ as the ratio of the expected signal energy (per sample) to the expected noise energy (per sample):

$$R \equiv \frac{W}{2\pi N} \int_{-\infty}^{+\infty} dk \left| \hat{T}(k) \right|^2$$

(4.5)

We shall also define the covariance matrix of the signal component of the data normalised to unit signal energy per sample

$$r_{a,b} \equiv \frac{\int_{-\infty}^{+\infty} dk \exp[i k (x_a - x_b)] \left| \hat{T}(k) \right|^2}{\int_{-\infty}^{+\infty} dk \left| \hat{T}(k) \right|^2}$$

(4.6)

We may now use Equation 3.14, Equation 3.15, and Equation 3.16 to calculate the transinformation. The maximum possible range for $I$ is

$$\log_2(m R + 1) \leq I \leq m \log_2(R + 1)$$

(4.7)

where the lower limit is obtained when the signals in the samples are completely redundant, and the upper limit is obtained when they are completely independent. To achieve the maximum possible transinformation we must place the samples in such a way that all their separations
Equation 4.8 is a special case of Equation 4.9 with $n = 0, 1, 2, 3, 4$ and 5 are depicted for a SNR $R = 100$.

lie at zeros of the signal covariance (see Equation 3.14); this is not usually possible.

The most common special case is when $|\hat{T}|$ takes the form

$$|\hat{T}(k)| = \begin{cases} 1 & |k| \leq c \\ 0 & |k| > c \end{cases} \quad (4.8)$$

In this case the samples can all be placed so that their separations correspond to zeros of the signal covariance, and so the upper limit in Equation 4.7 can be attained. This requires that all the intersample spacings be integer multiples of $\pi$ (the Nyquist length) $[5]$. Note that this result is true for any number of samples. However, it is important to note the assumptions which were used (Equation 3.1, Equation 3.2, Equation 3.3, Equation 4.1, Equation 4.2 and Equation 4.8) in order to derive this result. If any of these conditions is violated then the Nyquist sampling scheme does not achieve the upper limit in Equation 4.7 and more importantly it does not maximise the transinformation in general.

A useful class of $T$ is given by the cosine weighted filters

$$|\hat{T}(k)| = \begin{cases} \cos^n(\frac{\pi k}{2c}) & |k| \leq c \\ 0 & |k| > c \end{cases} \quad (4.9)$$

Equation 4.8 is a special case of Equation 4.9 with $n = 0$.

$r_{a,b}$ is then given by

$$r_{a,b} = \left(n!\right)^2 \sin[c (x_a - x_b)] \times \sum_{j=-n}^{n} \frac{(-1)^j}{(n-j)! (n+j)! c} \frac{1}{(x_a - x_b) - j \pi} \quad (4.10)$$

In Figure 1 we show how $I$ varies with the separation of the samples for the case $m = 2$ and $R = 100$. Note how the transinformation decreases as $n$ increases for a fixed sample separation. This is clearly a result of the progressively more drastic tapering of $\hat{T}$ as $n$ increases, which removes the higher frequencies from the response. As a consequence the samples have to be moved further apart in order to retain the same degree of independence.

V. RESTRICTED OBJECT SUPPORT FOR THE SINC MAPPING

In the previous section we derived some results for the transinformation obtained from a sampled image space of an isoplanatic mapping $T$ where the support of the continuous object space was assumed to be infinite. However, in practice we are more interested in situations where we have available some further constraints (prior knowledge) on the form of the object. Under such circumstances there is the possibility of obtaining a large gain in transinformation by optimising the sample positions. In order to keep the analysis as simple as possible we shall restrict our attention to the case studied in Equation 4.8 but with the additional restriction that the object support be limited. The generalisation to an arbitrary $W$ and $N$ is obvious.

Let us restrict the object support to the interval $[-1, +1]$, then $TW^\dagger$ is given by

$$TW^\dagger_{a,b} = W \int_{-\infty}^{+\infty} dz \frac{\sin[c (x_a - z)]}{\pi (x_a - z)} \frac{\sin[c (x_b - z)]}{\pi (x_b - z)} \quad (5.1)$$

which should be compared with Equation 4.1. Introducing the prolate spheroidal functions $\psi_k(z; c)$ [16] which have the properties

$$\sum_{k=0}^{\infty} \psi_k(x; c) \psi_k(y; c) = \frac{\sin[c (x - y)]}{\pi (x - y)} \quad (5.2)$$

$$\int_{-1}^{+1} dx \psi_k(x; c) \psi_l(x; c) = \delta_{k,l} \quad (5.3)$$

allows us to write $TW^\dagger$ as

$$TW^\dagger_{a,b} = W \sum_{k=0}^{\infty} \lambda_k(c) \psi_k(x_a; c) \psi_k(x_b; c) \quad (5.4)$$

We shall define the SNR as the ratio of the expected signal energy to the expected noise energy in a sample.
placed at the origin, which gives
\[
R \equiv \frac{W}{N} \sum_{k=0}^{\infty} \lambda_k(c) (\psi_k(0;c))^2 \tag{5.5}
\]

We may use Equation 3.14, Equation 3.15 and Equation 3.16 to calculate \( I \) if we define the normalised covariance \( r_{a,b} \) as
\[
r_{a,b} \equiv \frac{\sum_{k=0}^{\infty} \lambda_k(c) \psi_k(x_a;c) \psi_k(x_b;c)}{\sum_{k=0}^{\infty} \lambda_k(c) (\psi_k(0;c))^2} \tag{5.6}
\]

These matrix elements \( r_{a,b} \) depend on a single dimensionless combination of the object support width and system bandwidth \( c \). This is the well known Shannon number \( S \) which is the object support width measured in Nyquist lengths. \( S \) is given by the expression
\[
S \equiv \frac{2c}{\pi} \tag{5.7}
\]
for the support \([-1, +1]\).

We shall now select a simple class of image sampling schemes in order to study how the transinformation varies with Shannon number and sample position. For \( m \) data samples there are \( \frac{m^2}{2} \) (e.g. \( \frac{3^2}{2} = 1 \)) independent sample positions if we assume that the samples are to be symmetrically placed either side of the origin. We shall denote these positions \( 0, \pm x_1, \pm x_2, \ldots, \pm x_m \) (omitting zero when \( m \) is even). We can display \( I \) graphically for \( \frac{m^2}{2} = 1 \) and \( \frac{m^2}{2} = 2 \) alone; we shall choose \( \frac{m^2}{2} = 2 \) since this will show the greatest wealth of structure.

In Figure 2 we show contour plots of \( I \) against \((x_1, x_2)\) for various \( S \) and for \( R = 100 \), and we have marked the optimum configurations with a cross. The sample positions are measured in Nyquist lengths, and the contour plots are symmetric because the data space is unchanged by swapping \( x_1 \) and \( x_2 \). For large Shannon numbers (Figure 2(d)) the optimal scheme is very similar to the Nyquist scheme, but there is a marked departure from the Nyquist scheme at low Shannon numbers. The optimum may be explained as the result of an attempt by the samples to position themselves so that they receive a large amount of signal energy whilst at the same time sampling the signal at points which are as independent as possible. For large Shannon numbers the signal energy does not vary much with sample position, and so the sample values seek only to be independent: this leads to Nyquist sampling as the optimal scheme. For small Shannon numbers the signal energy is predominantly contained within a Nyquist length of the origin, and so the optimum position is a trade-off between maximising the signal energy and minimising the signal redundancy.

An interesting feature of Figure 2 is that the optimum sample positions are not always separate (Figure 2(a)). The meaning of two (or more) coincident samples is that the data measurement is taken two (or more) times.

The signal is the same in each such measurement, but the noise is distributed according to its PDF each time. Therefore such coincident samples enhance the signal energy with respect to the noise energy at the expense of measuring redundant signal information. Clearly such a situation becomes more attractive as the original SNR is decreased, and we should expect that the optimal sample positions will coalesce progressively. For a small enough SNR the situation is complementary to the case of the Nyquist limit discussed above, since the samples would then seek only to maximise the signal energy. In practice there are constraints on the allowed relative position of the samples, and on the amount of signal energy which each sample can acquire when in competition with other samples. However, the principle of transinformation maximisation is not affected by such considerations.

In Figure 3 we display the increase \( \Delta R \) in SNR which is required in order that the Nyquist sampling scheme (with samples centred on the object support) has a transinformation equal to that achieved by the optimum sampling scheme. Thus \( \Delta R \) is a measure of the effective power gain obtained by using the optimum scheme instead of the Nyquist scheme. No significant enhancement in the sampling scheme can be obtained when the Shannon number is greater than about five (the chosen number of samples). This is because when the Shannon number is large the support constraint does not significantly constrain the data. In the opposite limit as the Shannon number approaches zero the effective power gain approaches five (the number of samples). This is because the optimum scheme then has all the samples placed at the origin,
which leads to a transinformation of \( \log_2(mR + 1) \) bits, whereas the Nyquist sampling scheme leads to a transinformation of \( \log_2(R + 1) \) bits because only the sample at the origin receives any signal. Of course such a superposition of samples could lead to a depletion of the signal energy available to each sample in practice. If the signal energy which would be available to a single sample is shared out (unequally) amongst \( m \) coincident samples each with the original noise level, there is no increase in the transinformation. This completely removes the advantage of multiple sampling when the samples compete for signal energy. If such competition is only partial then the transinformation will be depressed from its nominal value (for non-competing samples). In practice the extent to which samples compete for signal energy is usually arranged to be minimal. Thus the problem of superimposed samples is obviously alleviated by taking sequential samples at the same point (in order to allow more signal energy to arrive). Another intuitively appealing result which can be read from Figure 3 is that \( \Delta R \) increases as \( R \) is decreased (for a constant \( S \)). Thus the extent to which sample position optimisation helps is greater the smaller the amount of signal energy present. More generally we observe that in order to preserve information pertaining to weak signals in a (hypothetical) continuous sampling scheme, we must be very careful which pieces of data we retain (i.e. which discrete sampling scheme we use). The uneven shape of the curves in Figure 3 is caused by secondary maxima in the transinformation surface (see Figure 2) falling close to the Nyquist sample point(s). This causes the Nyquist scheme to be fortuitously good, and so the optimisation gain is reduced in comparison with what it would have been had the secondary maxima been absent.

In Figure 4 we show how the maximum transinformation depends on the number of samples for \( S = 1 \) and various SNR. We have normalised \( I \) to be measured as a dimensionality \( D(m) \) (defined in Equation 2). For a given SNR each successive sample increases \( D(m) \) by less than its predecessor. For a given \( m \), \( D(m) \) increases as the SNR increases. The upper limit to \( D(m) \) which is suggested by these graphs is in accord with the number of components which may be extracted from (hypothetical) continuous data by using the SVD. Note that this comparison is ad hoc in so far as the SNR in the continuous case is defined in terms of energies per unit length, and not energies per sample as in the discrete case. However, we may identify an approximate correspondence between these two cases which justifies the comparison. We have not increased the number of samples to the point where the optimum solution requires that samples be superimposed, so our results do not require an analysis of how samples may compete.
VI. LAPLACE MAPPING

We shall now analyse the transinformation content of the data samples acquired in experiments which are described by the Laplace transform mapping \[12\] \[13\]. The mapping \( T \) is then given by

\[
T(x_a, z) \equiv \exp(-x_a z)
\]

\( \quad 0 \leq x_a < \infty \quad 0 \leq A \leq z < B < \infty \quad (6.1) \)

The Laplace transform has a conformal invariance which will be used later to simplify the results. We shall assume that the forms of \( W(x, y) \) and \( N_{a,b} \) are as in Equation 4.1 and Equation 4.2. The matrix elements of \( T W T^\dagger \) are then given by

\[
TW T^\dagger = \frac{W}{x_a + x_b} \left( \exp[-(x_a + x_b) A] - \exp[-(x_a + x_b) B] \right)
\]

\( \quad (6.2) \)

Let us define the signal-to-noise ratio \( R \) as the ratio of the expected signal energy to the expected noise energy in a sample at \( x_a = 0 \). This gives

\[
R \equiv \frac{(B - A) W}{N} \quad (6.3)
\]

We may use Equation 3.14, Equation 3.15 and Equation 3.16 to calculate \( I \) if we define the normalised covariance matrix \( r_{a,b} \) as

\[
r_{a,b} \equiv \frac{\exp[-(x_a + x_b) A] - \exp[-(x_a + x_b) B]}{(x_a + x_b) (B - A)} \quad (6.4)
\]

\[ \]

\[
I \equiv \log_2 \left| \frac{R + 1}{R} \left( e^{-x} - e^{-x \gamma} \right) \right| \quad (6.9)
\]

where intersample competition as \( X \rightarrow 0 \) has been ignored. In Figure 3 we show \( I \) plotted against \( X \) for \( R = 100 \) and various values of \( \gamma \). The graphs show that provided \( \gamma \) is large enough then there is a non-zero \( \gamma x \) which maximises \( I \). Clearly if \( X \rightarrow \infty \) then the second sample is wasted since it measures nothing but noise (Equation 6.11). On the other hand, as \( X \rightarrow 0 \) the signals measured by the two samples are completely redundant. Between these limits of \( X \) there is an optimum choice which is a trade-off between maximising the signal energy and minimising the signal redundancy. However, for a small enough \( \gamma \) (for a fixed SNR) the optimum choice is \( X = 0 \), where we have ignored the effect of intersample competition.

In order to make contact with the suggested exponential sampling method for Laplace transforms \[17\], we display the optimum sample positions as functions of \( \gamma \) for \( R = 10^6 \) in Figure 3. For a fixed \( \gamma \) we have used the largest number of samples for which the optimum positions are all distinct; thus there are several regions within each of which the number of samples remains fixed. As \( \gamma \) is increased the dimensionality of the data increases, and so a greater number of samples is permitted. The most important feature of this graph is that the separa-
The Use of Transinformation in the Design of Data Sampling Schemes for Inverse Problems

Figure 5: Plots of $I$ (in bits) for a two-sample image against $\log_{10}(\gamma X)$ for the Laplace filter of Equation 5.1. The first sample is placed at the origin, and $X$ is the position of the second sample in inverse units of the lower limit of the object support. $\gamma = 2, 3, 4, 5, 10$ and $\gamma \rightarrow \infty$ are depicted for $R = 100$.

The optimum data acquisition scheme requires that the samples be approximately exponentially spaced in accord with reference [17]. The exact choice of sample position depends on the prior knowledge which has been incorporated systematically into the transinformation.

In Figure 7 we show how the maximum transinformation (scaled to a dimensionality $D(m)$) depends on the number of samples for $\gamma = 5$ and various SNR. This figure is completely analogous to Figure 4 in all other respects, and the apparent upper limits to $D(m)$ in Figure 7 are consistent with the results of the SVD analysis of (hypothetical) continuous image data in reference [13]. Note that the same difficulties with relating a discrete SNR to a continuous SNR arise in this case as in the sinc case of Section V.

VII. CONCLUSIONS

Transinformation is the unique measure of additive information which is contained in a data set when prior knowledge is expressed in the form of PDFs. This quantity depends on both the prior knowledge and the way in which the data are acquired. We have restricted our attention to linear $T$ and to prior knowledge which can be expressed as gaussian PDFs, although the theory is completely general. Such situations have a transinformation which may be expressed in the form given by Equation 3.12. This form may be re-expressed as a sum of independent transinformations (Equation 3.19), which justifies the SVD method of data decomposition for gaussian prior PDFs.

The data themselves are a discrete representation of a continuous function, and the transinformation depends both on the sample positions and the sample value quantisation methods employed. We have assumed that the quantisation levels are of sufficient number and are sufficiently finely spaced that they constitute a continuum of levels, and so do not affect the transinformation. Instead we have investigated the dependence of the transinformation on the sample positions when there are only a few samples available; this is a realistic experimental situation. For linear $T$ and gaussian prior PDFs we have expressed the transinformation in terms of a normalised data covariance matrix, and also in terms of the intensity moments of the corresponding gaussian PDF (see Appendix VII). The numerical results which we have obtained from these expressions are in accord with our intuitive notions of how an information measure should behave.

We studied shift-invariant filters in Section IV and Section V. In Figure 1 we observed that when an object of infinite support is observed with a shift-invariant filter whose image space is sampled at two points, then
the number of data samples

Figure 7: Plots of the effective dimensionalitv $D(m)$ against the number of data samples $m$ for various SNRs and $\gamma = 5$. The Laplace filter of Equation 6.1 has been used.

The transinformation reflects the fact that the correlation length of the image increases as the filter is tapered more strongly. Thus the two samples must be separated by a distance greater than or of order of the correlation length in order that the signal redundancy is reduced to a negligible level. For the sinc filter the optimum sample positions are always separated by integer multiples of the Nyquist length whatever the number of samples (though note the restrictions mentioned after Equation 1.8). In Figure 2 we observed how a non-trivial form of prior knowledge, a support constraint on the object, modifies the optimum sample positions for a sinc filter. The optimum positions are modified progressively as the prior knowledge becomes stronger (Shannon number decreases). We may interpret the modification as arising through a trade-off between two conflicting conditions which the data samples are trying to satisfy: minimising the redundancy of the signal component of the data, and maximising the signal-to-noise ratio in each sample value. In Figure 3 we have shown the increase in the effective SNR which is obtained when one chooses the optimal sampling scheme instead of the Nyquist sampling scheme for the sinc filter. The gain in effective SNR increases as the Shannon number decreases. We may interpret this gain as arising from the assistance which the prior knowledge provides, and we could use the extent of the gain to define the strength of the prior knowledge. The gain exists because the Nyquist sampling scheme is optimal only for the limiting case when the support is of infinite extent. In Figure 4 we have shown how the dimensionality of the data increases as we increase the number of optimally placed samples, until it attains a ceiling beyond which no more transinformation is obtainable. This dimensionality also increases with SNR for a fixed number of optimally placed samples, as expected. Our dimensionality results are consistent with published results for the number of significant data components (i.e. the size of the signal space) which may be extracted by using the SVD.

We studied the Laplace transform in Section VI. The Laplace transform is an example of a conformally invariant mapping; there is a choice of scalings of the variables which leaves the kernel invariant. In order to maximise the transinformation we require that at least one data sample is located at zero (of the Laplace transform variable). In Figure 5 we have shown how the transinformation for a two-sample data space varies as the position of the second sample is varied. As for the case depicted in Figure 2 (sinc mapping) the optimum sample position is modified progressively as the strength of the prior knowledge is increased ($\gamma$ is decreased). In Figure 6 we have seen that the sampling scheme which is selected by maximising the transinformation corresponds to the exponential sampling suggested in reference [17]. This follows because Figure 6 shows that the logarithms of the optimum sample positions are approximately uniformly spaced. The data dimensionality results which we have shown in Figure 7 (analogous to Figure 4 for the sinc mapping) are consistent with SVD results.

We should not expect there to be gross differences between our results and those obtained by using the SVD, because the model which we have used to generate our numerical results (Equation 3.12) is closely related to the SVD method as explained in Section III and in references [5, 6]. However, use of the transinformation measure is not restricted to such simple situations. Indeed any data which are not distributed in a (correlated) gaussian manner cannot be analysed for their transinformation content by a simple diagonalisation. For more complicated distributions where correlations between more than two points (samples) must be separately specified we must use the basic equations of Section I to measure the transinformation. An estimate of the dimensionality of the data which is obtained from the SVD method will be an over-estimate if there are important higher-order correlations present. This is because any multi-point correlations (in addition to those implied by a correlated gaussian distribution) will decrease the transinformation content of the data. In general we should perform a full stochastic (Monte-Carlo) simulation of the data measurement process and the system under observation in order to derive the transinformation in complicated cases. We may then determine its dependence on sample position, sample type (quantisation levels, etc) and any other relevant parameters.
APPENDIX

In this appendix we shall summarise some formulae which are too long to include in the main text. We shall assume that the transinformation has been expressed in the form

\[ I \equiv \log_2(\det[M]) \]  

where

\[ M_{a,b} \equiv R_{a,b} + \delta_{a,b} \]  

This form corresponds to that used in the text if we identify

\[ R_{a,b} \equiv R r_{a,b} \]  

where \( R \) is a suitably defined signal-to-noise ratio. Thus \( R_{a,b} \) is the covariance matrix of the signal component of the data with the convention that the data are scaled so that the additive white noise has unit variance. We may evaluate the determinant in Equation (1) for various dimensions \( m \) of the data space. Thus we have

\[
\begin{align*}
m = 1: & \quad \det[M] = (R_{1,1} + 1) \\
m = 2: & \quad \det[M] = (R_{1,1} + 1) (R_{2,2} + 1) - |R_{1,2}|^2 \\
m = 3: & \quad \det[M] = (R_{1,1} + 1) (R_{2,2} + 1) (R_{3,3} + 1) - (R_{1,1} + 1)|R_{2,3}|^2 - 2 \text{ perms} \\
& \quad + 2 \text{ Re}(R_{1,2} R_{2,3} R_{3,1}) \\
m = 4: & \quad \det[M] = (R_{1,1} + 1) (R_{2,2} + 1) (R_{3,3} + 1) (R_{4,4} + 1) - (R_{1,1} + 1)(R_{2,2} + 1)|R_{3,4}|^2 - 5 \text{ perms} \\
& \quad + |R_{1,2} R_{3,4}|^2 + 2 \text{ perms} \\
& \quad + 2 (R_{1,1} + 1) \text{ Re}(R_{2,3} R_{3,4} R_{4,2}) + 3 \text{ perms} \\
& \quad - 2 \text{ Re}(R_{1,2} R_{2,3} R_{3,4} R_{4,1}) - 2 \text{ perms}
\end{align*}
\]

Note that it is important to retain the cyclic order of the variables when permuting terms which are not modulus squared.

On the other hand, we may define a gaussian PDF which has covariance \( R_{a,b} \)

\[ P[z] \equiv \frac{1}{\det[\pi R]} \exp(-z^\dagger R^{-1} z) \]  

where \( z \) is an \( m \)-dimensional complex vector and \( R \) is a matrix with elements \( R_{a,b} \). Using the factorisation properties of gaussian PDFs we may derive the expectation values of products of the intensities (not to be confused with transinformation) where

\[ I_k \equiv |z_k|^2 \]  

Thus we obtain

\[
\begin{align*}
I_1 & \equiv \langle 1 \rangle = R_{1,1} \\
I_1 I_2 & \equiv \langle 1 \cdot 2 \rangle = R_{1,1} R_{2,2} + |R_{1,2}|^2 \\
I_1 I_2 I_3 & \equiv \langle 1 \cdot 2 \cdot 3 \rangle = R_{1,1} R_{2,2} R_{3,3} \\
& \quad + R_{1,1} |R_{2,3}|^2 + 2 \text{ perms} \\
& \quad + 2 \text{ Re}(R_{1,2} R_{2,3} R_{3,1}) \\
I_1 I_2 I_3 I_4 & \equiv \langle 1 \cdot 2 \cdot 3 \cdot 4 \rangle = R_{1,1} R_{2,2} R_{3,3} R_{4,4} \\
& \quad + R_{1,1} R_{2,2} |R_{3,4}|^2 + 5 \text{ perms} \\
& \quad + |R_{1,2} R_{3,4}|^2 + 2 \text{ perms} \\
& \quad + 2 R_{1,1} \text{ Re}(R_{2,3} R_{3,4} R_{4,2}) + 3 \text{ perms} \\
& \quad - 2 \text{ Re}(R_{1,2} R_{2,3} R_{3,4} R_{4,1}) - 2 \text{ perms}
\end{align*}
\]
The quantities which appear in Equation 4 may be expressed in terms of the expected values of the intensity products to give

\[
\begin{align*}
m &= 1: & \det[M] &= ((1) + 1) \\
&\quad m = 2: & \det[M] &= ((1) + 1)( (2) + 1) - ((1 \cdot 2) - (1) (2)) \\
&\quad m = 3: & \det[M] &= ((1) + 1)( (2) + 1)((3) + 1) \\
&\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad - (2 (1) + 1)((2 \cdot 3) - (2) (3)) - 2 \text{ perms} \\
&\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad + ((1 \cdot 2 \cdot 3) - (1) (2) (3)) \\
&\quad m = 4: & \det[M] &= ((1) + 1)( (2) + 1)((3) + 1)( (4) + 1) \\
&\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad - (6 (1) (2) + 2 (1) + 2 (2) + 1)((3 \cdot 4) - (3) (4)) - 5 \text{ perms} \\
&\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad + 2 ((1 \cdot 2) (3 \cdot 4) - (1) (2) (3) (4)) + 2 \text{ perms} \\
&\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad + (2 (1) + 1)((2 \cdot 3 \cdot 4) - (2) (3) (4)) + 3 \text{ perms} \\
&\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad - ((1 \cdot 2 \cdot 3 \cdot 4) - (1) (2) (3) (4)) \tag{8}
\end{align*}
\]

These formulae allow us to express the transinformation either in terms of covariance matrix elements (as in Equation 4) or in terms of intensity moments (as in Equation 5).

---


