Low-level Multi-INT Sensor Fusion using Entropic Measures of Dependence

Paul B. Deignan, Mark A. Wong, and Alexander B. Douglass
Technology Development
L-3 Communications / MID
Greenville, TX, U.S.A.

Abstract – An information-theoretic method of low-level multi-INT sensor fusion is presented, the end product of which is the entropic map, i.e. a collection of Gaussian clusters of information relevant to a given target signature formed over a geographical basis. The method is designed to be computationally efficient with minimal side-information. To that end, an unbiased estimate of information from finite data is derived along with a data-dependent, information-optimal measurement partition. A method for the determination of the information-optimal sensor suite is given for a possibly geographically dependent target signature. Finally, it is shown that a multi-relational entropic measure of dependence can be superior to suboptimal error-based techniques of estimation of multiple sensor measurements of a real process.

Keywords: Entropic map, entropy, sensor fusion, resource management.

1 Introduction

Collaborative intelligence, surveillance, and reconnaissance (ISR) presumes that there exists a common object of information as well as a common reference for sharing information among agents/sensors. The object may be described as a multi-dimensional hypercube with dimensions corresponding to distinct sensors over a common geographical basis (Fig. 1) and will be referred to as an entropic map. In general, the sensor space of the hypercube is a subspace of a larger dimensional space in which the entities and events of interest are fully described, i.e. there is not necessarily a functional relationship between measurements of an entity of interest and sensor coordinates of the hypercube. For this reason descriptors of inter-dimensional association based on measures of central tendency are not always sufficient to quantify all entity related information present in the measured sensor space. However, entropic measures do possess the ability to quantify multi-valued relationships and when estimated as proposed here, also have additional mathematical properties that allow entropy estimates to be identified with “information”.

Entropic measures provide a quantification of spatial coherence both within and between dimensions and can be considered to be a generalization of covariance measures with the price of generalization being increased computation. Operational optimization of collaborative distributed sensors networks involves the maximization of useful information in terms of some metric of the network. For a generic network, the cost function of optimization might be taken as a ratio of environmental information discovered to information transmitted between agents/sensors.

Interest in information-theoretic techniques of sensor fusion has markedly increased over the past several years. Hero et al. [1] use Fischer’s information measure as an optimization criterion for the scheduling of sensors in a sensor management algorithm. Hero notes that since the information-theoretic measures are model-independent and otherwise general, they decouple the risk/reward optimization from the collection task in the sensor management algorithm design. Hero cautions that the use of Fisher information is only justifiable when the underlying posterior distribution is smooth. The use of Fischer’s information measure is of limited value in for multi-INT fusion since sensor measurements may be categorical or otherwise nonsmooth. Varshey [2], had similarly advocated information-theoretic measures as general cost functions in a sensor optimization algorithm.
Varshney was interested in distributed sensor detection and did framed his problem as a communication channel between sensor readings and the binary detection condition whereby their Shannon’s mutual information was to be maximized.

Mahler [3] mentions central entropy and cross-entropy as a measure of statistical dispersion, but relies primarily on models for radar-centric sensor fusion strategy of developing “hard” and “soft” mixture model clusters. This distinction is not as critical in the case of multi-INT fusion where sensors returns are relatively independent in the character of the target each sensor identifies. On the other hand, Schuck et al. [4] go beyond the problem of detecting and locating the existence of a target and concentrate on its identification. In this context, while a single-INT sensor such as radar may be used to measure simultaneously several independent attributes of a target, Schuck therefore relies to a greater extent than Mahler in the use of information-theoretic measures for hypothesis discrimination since the identification problem is not only soft, but also multi-relational. However, it should be noted that Schuck limits himself to one-dimensional entropy measurements and considers the contributions of independent sensors or attributes only after the assignment of targets probabilities are made based on that sensor return alone. Thus, Schuck does not fuse measurements, but rather fuses hypothesis using information-theoretic measures.

The utility of entropic measures is more than their use as a metric of discrimination of multimodal probability distribution or as a measure of relative probability concentration. Entropic measures have special properties arising from the fact that the logarithm is a group isomorphism from \((R^+,\times)\) to \((R,+)\) in that they enforce Bayes’ law in the relation between the frequentist counting of measurements and the probabilistic interpretation of the measurements as they relate to each other. While probabilities may be inferred, it is not necessary to estimate probability density functions to directly apply entropic measures in the assessment of measured data. This property avoids the curse of the concentration of measure phenomenon to estimates that produce tailed probability distributions whereby the integral of the probability estimate in measurement voids may be greater than that over regions of the measurement space in which measurements exist. This is a common cause of overfitting.

The paper is organized as follows: a brief background is given on information-theoretic measure estimation in Section 2. In the absence of a priori information, estimates of the measures are calculated over uniform partitions of the sensor space; however, since the bias of the measures is a function of the particular distribution of sensor readings as well as the cardinality of the partition, Section 3 gives the estimate information-optimal partition for a particular sensor dimension. Section 4 demonstrates the utility of mutual information for the association of simple asynchronous sensed signals. In the case of many associations, the combinatorics of the estimates becomes a significant bottleneck, thus Section 5 presents a branch and bound algorithm for the selection of mutual information-optimal sensor suites of common partition cardinality. Finally, Section 6 presents methods for forming the entropic map from multi-INT sensors.

2 Entropic Measures

When restricted to finite probability spaces, normalized measures of mutual information satisfy all of the Rényi postulates for measures of dependence [5]. By this token, entropy is the most general measure of statistical certainty when also estimated on a finite probability space which always is the case when given finite data. The mutual information between two or more random variables, \(I(X;Y)\), quantifies the uncertainty in signal, \(Y\), conditioned on another, \(X\), i.e.

\[
I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)
\]

and is zero only if one signal’s distribution is independent of the other, e.g., \(H(Y|X) = H(Y)\) [6]. Entropy is defined by the equation:

\[
H(Y) = \log(N) - \frac{1}{N} \sum_{n=1}^{N} n \log n
\]

for fixed record size, \(N\), with binning frequencies, \(n\), forming a partition, \(\mathcal{X}\). Mutual information can be calculated over the partitions \(\mathcal{X}\) for \(X\) and \(\mathcal{Y}\) for \(Y\).

\[
I(X,Y) = \log(N) + \frac{1}{N} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} n_{x,y} \left( \log n_{x,y} - \log n_x - \log n_y \right)
\]

which is a root-\(N\) consistent estimator of self-information for some bin width under a mild assumption of bounded tail behavior [7]. Note that the measure can be easily aggregated for multiple scopes by uniformly binning the data so the estimates yield themselves well to parallel computing architectures with each thread scaling by \(N\). Uniform binning over the span of the data values gives estimates invariant to affine transformations of individual variables, such as those caused by improperly calibrated linear sensors.

3 Estimation of Entropic Measures

The choice of binning interval for mutual information estimation is a compromise between resolution and computing resources given the measurements at hand. Estimates of the entropic functional may be calculated by substitution of estimates of probability. The probability of \(j\) occurrences in a particular bin given \(N\) occurrences uniformly distributed over \(k\) bins is given by the binomial probability distribution, \(f(\cdot)\),

![Image]

2036
\[ p(j) = f\left(\frac{j}{N}, \frac{1}{k}\right); \quad j \in \{0,1,2,...,N\} \]
\[ = \binom{N}{j} k^{-j} \left(1 - \frac{1}{k}\right)^{N-j} \]  

(4)

In the case of unordered data, the maximum feasible binning is the number of categories of a categorical data dimension. However, it is possible that not all categories are informative with respect to a target output and also that categories are mutually dependent on each other within the data sensor. Of course, the number of combinations of categories for each number of bins, \(k\), is \(\binom{N}{k}\) where \(N\) is the number of categories. The computation of all combinations for may be infeasible so that the suboptimal aggregation of categories by linkage lengths measured by the Kullback-Leibler distance between dimensional categories may be necessary. If the data is ordered, the entropic optimal probabilistic structure can be directly estimated as the binning that maximizes the difference of the entropic functional and partition entropy estimated by

\[ H_k^* = -k \cdot \sum_{j=1}^{N} \left( \frac{j \cdot p(j)}{N} \right) \log \left( \frac{j \cdot p(j)}{N} \right) \]  

(5)

where \(p(j)\) is given by (4). If the measurements are derived from a uniformly sampled time series of a continuous process, a regularity condition can be combined with a computational constraint in a penalty function, \(G\), of,

\[ \text{bin}_{opt} = \arg \max_k (I(X;Y) - \lambda G) \]

(6)

where \(k\) is the number of uniform bins of the dimension of variation of the mutual information estimate as well as a parameter of \(G\) and where \(\lambda\), the penalty factor, is positive real. Following [8], \(G\) may be chosen as a bias-corrected estimator derived from a finite difference approximation of a composition of the sampling and continuous process functions,

\[ G = k^2 \left( \sum_{r=1}^{N} \left( \frac{n_r}{N} - \frac{2}{N} \right)^2 \right) \]

(7)

where \(n_r\) is the number of occurrences found in consecutively indexed bins. The roughness penalty is the sum of squares of the differences in estimated probabilities of adjacent bins.

4 Speaker Network Identification

In order to demonstrate the aforementioned information theoretic concepts, a discrete-event queuing model was created to generate artificial speech patterns which represented pairs of speakers conversing over publicly switched voice communication networks. This model was based on ITU-T Recommendation P.59 and is depicted in the probabilistic state machine diagram shown below (Figure 2).

The four states may be described as: Speaker A talking/B silent, Speaker B talking/A silent, both speakers engaging in conversation simultaneously (double talk), and both speakers refraining from speaking (mutual silence). The model was duplicated to represent two independent pairs of speakers (A & B, C & D). Each model ran for a simulated 1000 seconds and generated traffic patterns for each speaker (Figure 3).

These patterns were written to a file during model runtime for post analysis to determine which pairs of speakers were engaged in conversation. More specifically, the mutual information was calculated on the traffic patterns across each pair-wise combination of speakers over a varying time sample window (Figure 4).
It was observed that the mutual information between communicating pairs remains almost constant, while the mutual information between non-communicating pairs decreases as the length of observation time increases. Thus, the example shows promise with the small number of pairwise correlations for the asynchronous association of sensed signals over a network.

To examine the effect of more pairs with similar statistics, the model was configured to generate speech traffic patterns for 1000 conversations. The mutual information data was examined to determine which speakers were associated with which other speakers in a conversation. The plot of mutual information versus speaker id is shown in a small scale in Figure 5 and for all speakers shown oriented by its edge in Figure 6. The latter figure shows type 1 and type 2 errors as the number of chance correlations overwhelm the simple single attribute mutual information statistic with the number of speaker pairs. Although the correct pairs are easily discernable by a human, the number of type 1 and type 2 errors indicates that additional sensor measurements (or side information from a datastore) could be used to augment the discrimination. The search for informationally optimal collaborative sensors scales computationally by the combinatorics of the associations.

5 Sequential Network Determination

Given a characteristic signature of a target in a certain geographical region, an optimal sensor suite can be determined by the maximization of the bias corrected mutual information estimate between the set of sensors and the target. Optimal sensor suite selection with Shannon’s joint mutual information as the criterion of merit evaluated over the candidate set of sensors is a combinatorial optimization problem in integer programming. In general, a dimensional stopping criteria need not be calculated a priori, but can be found through the branch and bound algorithm for a fixed set of measurements as a result of the finite data bias correction for partition entropy. Therefore, while the spanning set might conceivably be n-dimensional, in practice, it is often rare that high dimensional sets are feasible. The monotonicity of the estimated mutual information follows from the convexity of the natural logarithm over any interval of the positive real numbers, i.e.,

$$I(X_i; Y) \leq I(X_i, X_j; Y) \leq I(X_i, X_j, X_k; Y) \ldots (8)$$

for any \( \{X_i, X_j, X_k, \ldots, X_n\} \in X \) and is not altered by the finite data bias correction. The algorithm enumerates integer combinations via a spanning tree, the branches of which are evaluated in order.

At any node in the search of the spanning tree, subordinate branches are evaluated and ordered by decreasing mutual information. If the inequality is not satisfied for the current maximal estimated mutual information value, the algorithm backtracks and selects the next unexplored branch under the current node. The critical aspect of this algorithm is the selection of an efficient bound. Using a Venn diagram [9], the following bound on mutual information can be shown:

$$I(X_i, X_j; Y) \leq I(X_i, Y) + I(X_j, Y) \ldots (9)$$

In general,

$$I(X_i, X_j, \ldots, X_n; Y) \leq I(X_i, \ldots, X_n; Y) + I(X_n; Y) \ldots (10)$$

To the first degree of interaction, the bound on a branch of dimensionality \( l \) in a problem of dimensionality \( n \) is conservatively given for ease of computation as
As a result of information-theoretic subset selection, covariance, and average mutual information was calculated with a division of all signals into 50 uniformly spaced bins and for two contemporaneous with the target signal. The joint indexes the Gaussian clusters, \( \Omega \), is the set of available inputs was found to be particularly efficient [10]. Note that since the average mutual information of a null input set is zero, this inequality also imposes an implicit bound for irrelevant sensor rejection.

### 6 The Entropic Map Algorithm

The expectation-maximization algorithm is adapted for the optimization of input/output space mutual information clusters. The work here demonstrates that adaptation in the input dimension only. Generalization to the total input-output space is simply a matter of inclusion of the output dimension into the optimization vector.

The Expectation Maximization (EM) algorithm used for clustering of local mutual information estimates is as described by [11] for the iterative optimization of the maximum likelihood estimates of the Gaussian mixture model for probability distribution estimation except that estimates of MI are the operand. The algorithm is distilled into the following equations:

\[
I(\mathbf{X}^{(n)}; Y) \leq I(\mathbf{X}^{(n)}; Y) + \\
\max_{\Omega} \sum_{j=1}^{k} \left\{ I(\mathbf{X}^{(n)}; Y) - I(\mathbf{X}^{(n)} \setminus X_j; Y) \right\}, \quad (11)
\]

where \( \mathbf{X}^{(n)} \) is the input set at level \( n \) and \( \Omega \) is the set of available inputs was found to be particularly efficient [10]. Note that since the average mutual information of a null input set is zero, this inequality also imposes an implicit bound for irrelevant sensor rejection.

\[
I(\mathbf{X}^{(n)}; Y) \leq I(\mathbf{X}^{(n)}; Y) + \\
\max_{\Omega} \sum_{j=1}^{k} \left\{ I(\mathbf{X}^{(n)}; Y) - I(\mathbf{X}^{(n)} \setminus X_j; Y) \right\}, \quad (11)
\]

where \( \mathbf{X}^{(n)} \) is the input set at level \( n \) and \( \Omega \) is the set of available inputs was found to be particularly efficient [10]. Note that since the average mutual information of a null input set is zero, this inequality also imposes an implicit bound for irrelevant sensor rejection.

The Entropic Map Algorithm

The expectation-maximization algorithm is adapted for the optimization of input/output space mutual information clusters. The work here demonstrates that adaptation in the input dimension only. Generalization to the total input-output space is simply a matter of inclusion of the output dimension into the optimization vector.

The Expectation Maximization (EM) algorithm used for clustering of local mutual information estimates is as described by [11] for the iterative optimization of the maximum likelihood estimates of the Gaussian mixture model for probability distribution estimation except that estimates of MI are the operand. The algorithm is distilled into the following equations:

\[
\mathbf{\mu}^{(n)}_j = \frac{\sum \mathbf{x}^{(n)} P^{(n)}(j | \mathbf{x}^{(n)})}{\sum \mathbf{x}^{(n)} P^{(n)}(j | \mathbf{x}^{(n)})} \quad (12)
\]

\[
\left( \mathbf{\sigma}^{(n)}_j \right)^{-1} = \frac{1}{d} \frac{\sum \mathbf{x}^{(n)} P^{(n)}(j | \mathbf{x}^{(n)}) \mathbf{x}^{(n)} \mathbf{x}^{(n)\top} - \mathbf{\mu}^{(n)}_j \mathbf{\mu}^{(n)}_j \mathbf{I}}{\sum \mathbf{x}^{(n)} P^{(n)}(j | \mathbf{x}^{(n)})} \quad (13)
\]

\[
P^{(n)}(j | \mathbf{x}^{(n)}) = \frac{1}{N} \sum \mathbf{x}^{(n)} P^{(n)}(j | \mathbf{x}^{(n)}) \quad (14)
\]

where \( d \) is the dimensionality, \( N \) the number of records, \( j \) indexes the Gaussian clusters, \( \mu \) is the mean vector, \( \sigma \) the covariance, and \( \mathbf{x} \) is the data vector.

Consider a system of two sensors and a target signal. As a result of information-theoretic subset selection, Sensors \( \mathbf{X} \) and \( \mathbf{Y} \) are jointly determined to provide the most information relating to the target signal. The distribution of mutual information is illustrated by Figure 8. The map should approximate the target signature well with the minimal number of parameters. For this example, the potential sensed signals will be limited to two contemporaneous with the target signal. The joint average mutual information was calculated with a division of all signals into 50 uniformly spaced bins and for examined with exponential transformations of the sensed signals: \( x^2 \) to \( x^2 \). To demonstrate the consistency and utility of the entropic map constructed by the E-M technique, we will also construct an improved forward regression – orthogonal least squares (FR-OLS) radial basis function network (RBFN) of Chen et al., [12] and a mutual information based Gaussian mixture model where the mutual information estimates are clustered using a linkage length minimization scheme, which produces irregularly shaped clusters as opposed to the Gaussian clusters of the EM method.

Empirical results indicated that the FR-OLS algorithm seemed to perform best if the data points were separated in the input hyperspace by a Euclidean distance of at least 0.01. The ratio of training to test set data was selected as 4:1 and was divided using a round-robin scheme. The FR-OLS RBFN which minimizes the root mean squared error (RMSE) of the target signal is chosen as one which best approximates the data set with a minimum of overfitting. At each instance of node addition, a locally optimal global spread constant was recalculated in order to minimize the combined test and training set error. The spread constant is the node width parameter defined as the Euclidean distance from the node center that generates node output of 0.5 before multiplication by the linear height factor. The minimum total error was found at the incremental addition of the seventh node for a spread constant of 1.4790 and a RMSE calculated over the entire data set and normalized to equal 100 (Table 1).

Two different methods of mutual information clustering for node center location are examined. The first method (Figure 9) clusters the values of mutual information estimates by minimal linkage lengths with Gaussian node centers located at the weighted mean along each dimension. An analysis of the mutual information distribution using an incremental nearest neighbor cluster scheme indicates that the input space may be naturally divided into only six clusters. The data set is normalized and a single spread constant is determined through minimization of the normalized target RMSE (842).

As can be seen by Figure 8, there is a high degree of mutual information for the singularly prominent sensor state with lower values of mutual information irregularly spread throughout regions of the input space. Note that the iterative FR-OLS RBFN method locates node centers in close proximity where they could effectively be replaced by one node at a slightly different location with no significant effect to the overall mapping. Furthermore, one FR-OLS RBFN node is placed outside of the range supported by the data.

The second method (expectation-maximization) specifically assumes Gaussian clusters that will later conform to the Gaussians of the RBFN. In order to check the assertion that forming a Gaussian mixture model based on mutual information is most efficient in this instance, only five nodes are allowed. However, after fitting the nonlinearily optimized global node width scaling factor to the unequally sized radial basis nodes, it is found that a higher accuracy is achieved. This result supports the claims of efficiency of this methodology. A summary of results is shown in Table 1.
Fig. 8. \( \hat{I}(\text{Target}; \text{Sensor } X, \text{Sensor } Y) \).

Fig. 9. Entropic map with cluster centers noted (○ FR-OLS; ● Linkage length). All cubes within a linkage length cluster are labeled similarly.

Fig. 10. Entropic clusters located and sized by Expectation-Maximization algorithm.

Table 1: Entropic Mapping Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Nodes</th>
<th>Relative RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR-OLLS</td>
<td>7</td>
<td>100.0</td>
</tr>
<tr>
<td>Linkage length clustering</td>
<td>6</td>
<td>84.2</td>
</tr>
<tr>
<td>EM clustering</td>
<td>5</td>
<td>81.9</td>
</tr>
</tbody>
</table>

The EM-MI clustering (Figure 10) interpolates with Gaussian radial basis functions centered on clusters of mutual information, a far superior map is obtained than that of the RBFN although it has a greater number of nodes. The results are further improved when the mutual information clusters are defined by the expectation-maximization of Gaussians rather than the more general minimal linkage method. Fewer parameters implies greater generality by Occam’s razor and so the mutual information method of sensor fusion appears to be not only better at identifying this particular target signal, but should generalize well to other similar targets.

Whereas there are many methods of fitting RBFN nodes that rely on the direct minimization of an error metric, we adopt a two-stage process. In the first stage, the statistical support for the final mapping is established thereby fixing the node locations and relative widths. Only then are the RBFNs scaled to minimize an error metric. The node locations and variances are achieved through the adaptation of a multi-modal Gaussian maximum likelihood estimate of local mutual information estimates by expectation-maximization which establishes the support for the interpolation. Since the Gaussian regression estimate is a maximum likelihood estimator, the transformation from mutual information clustering to output mapping only involves a scaling of the Gaussians. Only one nonlinear parameter must be found in the optimization procedure through a line search. Moreover, the mapping is statistically robust.

7 Conclusions

This paper presents a unified approach for the estimation of an entropic map over a common basis by multi-INT sensors. A partition entropy bias term for entropic estimates derived from finite data is presented together with quantization guidance for continuous data. The problem of optimal sensor suite selection is solved through a branch-and-bound algorithm. Finally, it is demonstrated that an entropic map generated by information-theoretic sensor fusion may outperform conventional estimation techniques for real target data. The consistency of this result with the previous derivations supports the theoretical observations that information-theoretic multi-INT sensor fusion techniques should provide a tangible benefit in application to our increasingly more complex and challenging ISR requirements.
References


