Hypothesis testing over factorizations for data association

Alexander T. Ihler, John W. Fisher III, and Alan S. Willsky

ihler@mit.edu, fisher@ai.mit.edu, willsky@mit.edu
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract. The issue of data association arises frequently in sensor networks; whenever multiple sensors and sources are present, it may be necessary to determine which observations from different sensors correspond to the same target. In highly uncertain environments, one may need to determine this correspondence without the benefit of an *a priori* known joint signal/sensor model. This paper examines the data association problem as the more general hypothesis test between factorizations of a single, learned distribution. The optimal test between known distributions may be decomposed into model-dependent and statistical dependence terms, quantifying the cost incurred by model estimation from measurements compared to a test between known models. We demonstrate how one might evaluate a two-signal association test efficiently using kernel density estimation methods to model a wide class of possible distributions, and show the resulting algorithm's ability to determine correspondence in uncertain conditions through a series of synthetic examples. We then describe an extension of this technique to multi-signal association which can be used to determine correspondence while avoiding the computationally prohibitive task of evaluating all hypotheses. Empirical results of the approximate approach are presented.

1 Introduction

Data association describes the problem of partitioning observations into like sets. This is a common problem in networks of sensors – multiple signals are received by several sensors, and one must determine which signals at different sensors correspond to the same source.

In many collaborative sensing scenarios, the signal models are assumed to be known and fully specified *a priori*. With such models, it is possible to formulate and use optimal hypothesis tests for data association. However, real-world uncertainty often precludes strong modelling assumptions. For example, it is difficult to analytically quantify dependence between data of different modalities. Additionally, nonlinear effects and inhomogenous media create complex interactions and uncertainty. When applicable, a learning/estimation based approach is appealing, but in the online case requires that one learn the signal distributions while simultaneously performing the test. For example, this is possible for data association because it is a test described in terms of the distribution *form*, in particular as a test over factorization and independence.

We show that the optimal likelihood test between two factorizations of a density learned from the data can be expressed in terms of mutual information. Furthermore, the analysis results in a clear decomposition of terms related to statistical dependence (i.e. factorization) and those related to modelling assumptions. We propose the use of kernel density methods to estimate the distributions and mutual information from data. In the case of high-dimensional data, where learning a distribution is impractical, this can be done efficiently by finding *statistics* which capture its interaction. Furthermore, the criterion for learning these statistics is also expressed in terms of mutual information. The estimated mutual information of these statistics can be used as an approximation to the optimal likelihood ratio test, by training the statistics to minimize a bound on the approximation error.

We will begin by describing a data association example between a pair of sensors, each observing two targets. We show first how the optimal hypothesis test changes in the absence of a known signal model and express the resulting test in terms of information. We then discuss how one may use summarizing features to estimate the mutual information efficiently and robustly using kernel methods. This can yield a tractable estimate of the hypothesis test when direct estimation of the observations' distribution is infeasible. Finally, we present an algorithmic extension of these ideas to the multiple target case.

2 An Information-Theoretic Interpretation of Data Association

Data association can be cast as a hypothesis test between density factorizations over measurements. As we will show, there is a natural information-theoretic interpretation of this hypothesis test, which decomposes the test into terms related to statistical dependency and terms related to modelling assumptions. Consequently, one can quantify the contribution of prior knowledge as it relates to a known model; but more importantly, in the absence of a prior model one can still achieve a degree of separability between hypotheses by estimating statistical dependency only. Furthermore, as we show, one can do so in a low-dimensional feature space so long as one is careful about preserving information related to the underlying hypothesis.

Consider the following example problem, which illustrates an application of data association within tracking problems. Suppose we have a pair of widely spaced acoustic sensors, where each sensor is a small array of many elements. Each sensor produces an observation of the source and an estimate of bearing, which in itself is insufficient to localize the source. However, triangulation of bearing measurements from multiple sensors can be used to estimate the target location. For a single target, a pair of sensors is sufficient to perform this triangulation.

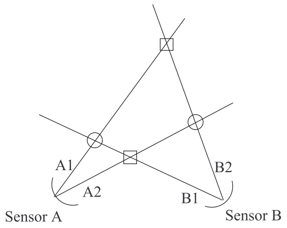


Fig. 1. The data association problem: two pairs of measurements results in estimated targets at either the circles or the squares; but which remains ambiguous.

However, complications arise when there are multiple targets within a pair of sensors' fields of view. Each sensor determines two bearings; but this yields four possible locations for only two targets, as depicted in Figure 1. With only bearing information, there is no way to know which one of these target pairs is real, and which is the artifact. We will show that it is possible to address this ambiguity under the assumption that the sources are statistically independent, without requiring a prior model of the relationship between observations across sensors.

2.1 Mutual Information

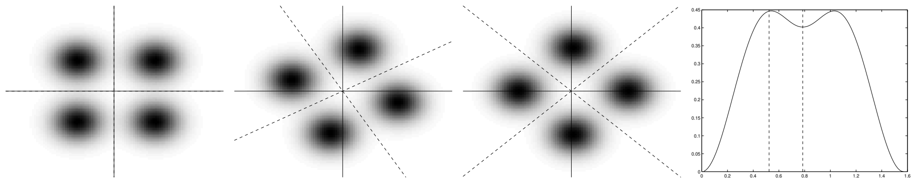
Mutual information is a quantity characterizing the statistical dependence between two random variables. Although most widely known for its application to communications (see e.g. [1]), here it arises in the context of discrimination and hypothesis testing [2].

Correlation is equivalent to mutual information only for jointly Gaussian random variables. The common assumption of Gaussian distributions and its computational efficiency have given it wide applicability to association problems. However, there are many forms of dependency which are not captured by correlation.

For example, Figure 2(a-c) shows three non-Gaussian joint distributions characterized by a single parameter θ, indicating an angle of rotation with respect to the random variables x, y. Although the correlation between x and y is zero for all θ, the plot of mutual information as a function of θ (Figure 2(d)) demonstrates that for many θ, x and y are far from independent. This illustrates how mutual information as a measure of dependence differs from correlation.

2.2 Data Association as a Hypothesis Test

Let us assume that we receive N *i.i.d.* observations of each source at each of the two sensors. When a full distribution is specified for the observed signals, we



(a) $\theta = 0$ (b) $\theta = \frac{\pi}{6}$ (c) $\theta = \frac{\pi}{4}$ (d) $I(\theta)$

Fig. 2. Two variables x, y with joint distributions (a-c) are uncorrelated but not necessarily independent – (d) shows mutual information as a function of the angle of rotation θ.

have a hypothesis test over *known*, factorized models

$$\begin{aligned} H_1 : [A_1, B_1, A_2, B_2]_k &\sim p_{H_1}(A_1, B_1) p_{H_1}(A_2, B_2) \\ H_2 : [A_1, B_1, A_2, B_2]_k &\sim p_{H_2}(A_1, B_2) p_{H_2}(A_2, B_1) \\ &\text{for } k \in [1 : N] \end{aligned} \tag{1}$$

with corresponding (normalized) log-likelihood ratio

$$\frac{1}{N} \log L = \frac{1}{N} \sum_{k=1}^{N} \left[\log \frac{p_{H_1}([A_1, B_1]_k) p_{H_1}([A_2, B_2]_k)}{p_{H_2}([A_1, B_2]_k) p_{H_2}([A_2, B_1]_k)} \right] \tag{2}$$

As N grows large, the (normalized) log-likelihood approaches its expected value, which can be expressed in terms of mutual information (MI) and Kullback-Leibler (KL) divergence. Under H_1 this value is

$$\begin{aligned} E_{H_1}[\log L] =& I_{H_1}(A_1; B_1) + I_{H_1}(A_2; B_2) + \\ & D(p_{H_1}(A_1), \ldots, p_{H_1}(B_2) \| p_{H_2}(A_1, \ldots, B_2)) \end{aligned} \tag{3}$$

and similarly when H_2 is true:

$$\begin{aligned} E_{H_2}[\log L] =& - I_{H_2}(A_1; B_2) - I_{H_2}(A_1; B_2) - \\ & D(p_{H_2}(A_1), \ldots, p_{H_2}(B_2) \| p_{H_1}(A_1, \ldots, B_2)) \end{aligned} \tag{4}$$

The expected value of Equation (3) can be grouped in two parts – an information part (the two MI terms) measuring statistical dependency across sensors, and a model mismatch term (the KL-divergence) measuring difference between the two models. We begin by examining the large-sample limits of the likelihood ratio test, expressed in terms of its expected value; when this likelihood ratio is not available we see that another estimator for the same quantity may be substituted.

Often the true distributions p_{H_i} are unknown, e.g. due to uncertainty in the source densities or the medium of signal propagation. Consider what might be done with estimates of the densities based on the empirical data to be tested. Note that this allows us to learn densities *without* requiring multiple trials under similar conditions. We can construct estimates assuming the factorization under either hypothesis, but because observations are only available for the true hypothesis our estimates of the other will necessarily be incorrect. Specifically, let $\hat{p}_{H_i}(\cdot)$ be a consistent estimate of the joint distribution assuming the factorization under H_i and let $\tilde{p}_{H_i}(\cdot)$ denote its limit; then we have

$$
\begin{aligned}
&\text{if } H_1 \text{ is true,}\\
&\quad \hat{p}_{H_1} \to \tilde{p}_{H_1} = p_{H_1}(A_1, B_1)p_{H_1}(A_2, B_2)\\
&\quad \hat{p}_{H_2} \to \tilde{p}_{H_2} = p_{H_1}(A_1)p_{H_1}(B_1)p_{H_1}(A_2)p_{H_1}(B_2)\\
&\text{if } H_2 \text{ is true,}\\
&\quad \hat{p}_{H_1} \to \tilde{p}_{H_1} = p_{H_2}(A_1)p_{H_2}(B_1)p_{H_2}(A_2)p_{H_2}(B_2)\\
&\quad \hat{p}_{H_2} \to \tilde{p}_{H_2} = p_{H_2}(A_1, B_2)p_{H_2}(A_2, B_1)
\end{aligned}
\tag{5}
$$

Thus when $\hat{p}_{H_i}$ assumes the correct hypothesis we converge to the correct distribution, while assuming the incorrect hypothesis leads to a fully factored distribution. This is similar to issues arising in generalized likelihood ratio (GLR) tests [3].

We proceed assuming that our estimates have negligible error, and analyze the behavior of their limit $\tilde{p}(\cdot)$; we will examine the effect of error inherent in finite estimates $\hat{p}(\cdot)$ later. Now the expectation of the log-likelihood ratio can be expressed solely in terms of the mutual information between the observations. Under H_1 this is

$$
\begin{aligned}
E_{H_1}[\log \tilde{L}] &= E_{H_1}\left[\log \frac{\tilde{p}_{H_1}(A_1, B_1)\tilde{p}_{H_1}(A_2, B_2)}{\tilde{p}_{H_2}(A_1, B_2)\tilde{p}_{H_2}(A_2, B_1)}\right]\\
&= I(A_1; B_1) + I(A_2; B_2)
\end{aligned}
$$

and similarly under H_2,

$$
E_{H_2}[\log \tilde{L}] = -I(A_1; B_2) - I(A_2; B_1)
$$

Notice in particular that the KL-divergence terms stemming from model mismatch in Equation (3) have vanished. This is due to the fact that both models are estimated from the same data, and quantifies the increased difficulty of discrimination when the models are unknown. We can write the expectation *independent* of which hypothesis is true as

$$
E[\log \tilde{L}] = I(A_1; B_1) + I(A_2; B_2) - I(A_1; B_2) - I(A_2; B_1) \tag{6}
$$

since for either hypothesis, the other two terms above will be zero; this casts the average log-likelihood ratio as an estimator of mutual information.

We have not assumed that the true distributions $p(\cdot)$ have any particular form, and therefore might consider using nonparametric methods to ensure that our estimates converge under a wide variety of true distributions. However, if the observations are high-dimensional such methods require an impractical number of samples in order to obtain accurate estimates. In particular, this means that the true likelihood ratio cannot be easily calculated, since it involves estimation and evaluation of high-dimensional densities. However, the log-likelihood ratio is acting as an estimator of the mutual information, and we may instead substitute another, more tractable estimate of mutual information if available.

Direct estimation of the MI terms above using kernel methods also involves estimating high-dimensional distributions, but one can express it succinctly using features which summarize the data interaction. We explore ways of learning such features, and shall see that the quality criterion for summarization is expressed as the mutual information between features estimated in a low-dimensional space.

Let us suppose initially that we possess low-dimensional sufficient statistics for the data. Although finding them may be difficult, we know that for the data association problem sufficient statistics should exist, since the true variable of interest, correspondence, is summarized by a single scalar likelihood. More precisely, let $f_i^{A_j}$ be a low-dimensional feature of A_j and $\bar{f}_i^{A_j}$ its complement, such that there is a bijective transformation between A_j and $[f_i^{A_j}, \bar{f}_i^{A_j}]$ (and similarly for B_k). If the following relation holds,

$$\begin{aligned} p_{H_i}(A_j, B_k) &= p_{H_i}(f_i^{A_j}, \bar{f}_i^{A_j}, f_i^{B_k}, \bar{f}_i^{B_k}) \\ &= p_{H_i}(f_i^{A_j}, f_i^{B_k}) p_{H_i}(\bar{f}_i^{A_j} | f_i^{A_j}) p_{H_i}(\bar{f}_i^{B_k} | f_i^{B_k}) \end{aligned} \tag{7}$$

then the log-likelihood ratio of Equation (6) can be written exactly as

$$E[\log \tilde{L}] = I(f_1^{A_1}; f_1^{B_1}) + I(f_1^{A_2}; f_1^{B_2}) - I(f_2^{A_1}; f_2^{B_2}) - I(f_2^{A_2}; f_2^{B_1}) \tag{8}$$

Although sufficient statistics are likely to exist, it may be difficult or impossible to find them exactly. If the features $f_i^{A_j}$ and $f_i^{B_k}$ are *not* sufficient, several divergence terms must be added to Equation (8). For *any* set of features satisfying $p_{H_i}(A_j, B_k) = p_{H_i}(f_i^{A_j}, \bar{f}_i^{A_j}, f_i^{B_k}, \bar{f}_i^{B_k})$, we can write

$$E[\log \tilde{L}] = I_1^{1;1} + I_1^{2;2} - I_2^{1;2} - I_2^{2;1} + D_1^{1;1} + D_1^{2;2} - D_2^{1;2} - D_2^{2;1} \tag{9}$$

where for brevity we have used the notation

$$\begin{aligned} I_i^{j;k} &= I(f_i^{A_j}; f_i^{B_k}) \\ D_i^{j,k} &= D(\tilde{p}(A_j, B_k) \| \tilde{p}(f_i^{A_j}, f_i^{B_k}) \tilde{p}(\bar{f}_i^{A_j} | f_i^{A_j}) \tilde{p}(\bar{f}_i^{B_k} | f_i^{B_k})) \end{aligned}$$

The data likelihood of Equation (9) contains a difference of the divergence terms from each hypothesis. Notice, however, that only the divergence terms involve high-dimensional data; the mutual information is calculated between low-dimensional features. Thus if we discard the divergence terms we can avoid all calculations on the high-dimensional compliment features $\bar{f}$. We would like to minimize the effect on our estimate of the likelihood ratio, but cannot estimate the terms directly without evaluating high-dimensional densities. However, by nonnegativity of the KL-divergence we can bound the difference by the sum of the divergences:

$$\left| D_1^{1;1} + D_1^{2;2} - D_2^{1;2} - D_2^{2;1} \right| \quad \leq \quad D_1^{1;1} + D_1^{2;2} + D_2^{1;2} + D_2^{2;1} \tag{10}$$

We then minimize this bound by minimizing the individual terms, or equivalently maximizing each mutual information term (which can be done in the low-dimensional feature space). Note that these four optimizations are decoupled from each other.

Finally, it is unlikely that with finite data our estimates $\hat{p}(\cdot)$ will have converged to the limit $\tilde{p}(\cdot)$. Thus we will also have divergence terms from errors in the density estimates:

$$\begin{aligned} E[\log \tilde{L}] =& \hat{I}_1^{1;1} + \hat{I}_1^{2;2} - \hat{I}_2^{1;2} - \hat{I}_2^{2;1} \\ &+ D(\tilde{p}_{H_1} \| \hat{p}_{H_1}) - D(\tilde{p}_{H_2} \| \hat{p}_{H_2}) \end{aligned} \tag{11}$$

where the $\hat{I}$ indicate the mutual information of the density estimates. Once again we see a difference in divergence terms; in this case minimization of the bound means choosing density estimates which converge to the true underlying distributions as quickly as possible. Note that if $\hat{p}_{H_1}(\cdot)$ is not a consistent estimator for the distribution $\tilde{p}_{H_i}(\cdot)$, the individual divergence terms above will never be exactly zero.

Thus we have an estimate of the true log-likelihood ratio between factorizations of a learned distribution, computed over a low-dimensional space:

$$\begin{aligned} E[\log \tilde{L}] =& \hat{I}(f_1^{A_1}; f_1^{B_1}) + \hat{I}(f_1^{A_2}; f_1^{B_2}) - \hat{I}(f_2^{A_1}; f_2^{B_2}) \\ &- \hat{I}(f_2^{A_2}; f_2^{B_1}) + \textit{divergence terms} \end{aligned} \tag{12}$$

where maximizing the $\hat{I}$ with regard to the features $f_i^{X_j}$ minimizes a bound on the ignored divergence terms. We can therefore use estimates of the mutual information over learned features as an estimate of the true log-likelihood ratio for hypothesis testing.

3 Algorithmic Details

The derivations above give general principles by which one may design an algorithm for data association using low-dimensional sufficient statistics. Two primary elements are necessary:

1. a means of estimating entropy, and by extension mutual information, over samples, and
2. a means of optimizing that estimate over the parameters of the sufficient statistic.

We shall address each of these issues in turn.

3.1 Estimating Mutual Information

In estimating mutual information, we wish to avoid strong prior modelling assumptions, i.e. jointly Gaussian measurements. There has been considerable research into useful nonparametric methods for estimating information-theoretic quantities; for an overview, see e.g. [4].

Kernel density estimation methods are often used as an appealing alternative when no prior knowledge of the distribution is available. Similarly, these kernel-based methods can be used to estimate mutual information effectively. Using estimates with smooth, differentiable kernel shapes will also yield simple calculations of a gradient for mutual information, which will prove to be useful in learning. An issue one must consider is that the quality of the estimate degrades as the dimensionality grows; thus we perform the estimate in a low-dimensional space.

To use kernel methods for density estimation requires two basic choices, a kernel *shape* and a *bandwidth* or smoothing parameter. For the former, we use Gaussian kernel functions $K_\sigma(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-x^2/2\sigma^2\}$, where σ controls the bandwidth. This ensures that our estimate is smooth and differentiable everywhere. There are a number of ways to choose kernel bandwidth automatically (see e.g. [5]). Because we intend to use these density estimates for likelihood evaluation and maximization, it is sensible to make this the criterion for bandwidth as well; we therefore make use of a leave-one-out maximum likelihood bandwidth, given by

$$\arg\max_\sigma \left[-\frac{1}{N} \sum_j \log \left(\frac{1}{N-1} \sum_{i \neq j} K_\sigma(x_j - x_i) \right) \right] \tag{13}$$

Because our variables of interest are continuous, it is convenient to write the mutual information in terms of joint and marginal entropy, as:

$$I(f_i^{A_j}; f_i^{B_k}) = H(f_i^{A_j}) + H(f_i^{B_k}) - H(f_i^{A_j}, f_i^{B_k}) \tag{14}$$

There are a number of possible kernel-based estimates of entropy available [4]. In practice we use either a leave-one-out resubstitution estimate:

$$\hat{H}_{RS}(x) = -\frac{1}{N} \sum_j \log \left(\frac{1}{N-1} \sum_{i \neq j} K_\sigma(x_j - x_i) \right) \tag{15}$$

or an integrated squared error estimate from [6]:

$$\hat{H}_{ISE} = H(\mathbf{1}) - \frac{1}{2} \int (\mathbf{1} - \hat{p}(x))^2 \, dx \tag{16}$$

where $\mathbf{1}$ is the uniform density on a fixed range, and

$$\hat{p}(x) = \frac{1}{N} \sum_j K_\sigma(x - x_j)$$

These methods have different interpretations – the former is a stochastic estimate of the true entropy, while the latter can be considered an exact calculation of an entropy approximation. In practice both of these estimates produce similar results. Both estimates may also be differentiated with respect to their arguments, yielding tractable gradient estimates useful in learning.

3.2 Learning Sufficient Statistics

In order to learn sufficient or relatively sufficient statistics, we must define a function from our high-dimensional observation space to the low-dimensional space over which we are able to calculate mutual information. By choosing a function which admits a simple gradient-based update of the parameter values, we can use gradient ascent to train our function towards a local information maximum [7, 8].

Often, quite simple statistic forms will suffice. For example, all of the examples below were performed using a simple linear combination of the input variables, passed through a hyperbolic tangent function to threshold the output range:

$$f(x = [x^1 \dots x^d]) = \tanh(\sum_i w_i x^i) \tag{17}$$

That is, using the method of [7, 8] we apply gradient ascent of mutual information between the associated features with respect to the weight parameters w_i.

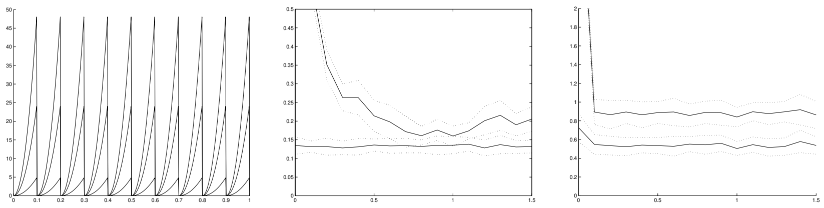
However, the methods are applicable to any function which can be trained with gradient estimates, allowing extension to much more complex functional forms. In particular, *multiple layer perceptrons* are a generalization of the above form which, allowed sufficient complexity, can act as a universal function approximator [9].

We may also wish to impose a capacity control or complexity penalty on the model (e.g. regularization). In practice, we put a penalty on the absolute sum of the linear weights (adding to the gradient a constant bias towards zero) to encourage sparse values.

4 Data Association of Two Sources

We illustrate the technique above with two examples on synthetic data. The first is a simulation of dispersive media – an all-pass filter with nonlinear phase characteristics controlled by an adjustable parameter α. The phase response for three example values of α are given in Figure 3(a). Sensor A observes two independent signals of bandpassed *i.i.d.* Gaussian noise, while sensor B observes the allpass-filtered versions of A.

If the filter characteristics are known, the optimal correspondence test is given by applying the inverse filter to B followed by finding its correlation with A. However when the filter is not known, estimating the inverse filter becomes a source reconstruction problem. Simple correlation of A and B begins to fail as the phase becomes increasingly nonlinear over the bandwidth of the sources. The upper curve of Figure 3(b) shows the maximum correlation coefficient between correct pairings of A and B over all time shifts, averaged over 100 trials. Dotted lines indicate the coefficient's standard deviation over the trials. To determine significance, we compare this to a baseline of the maximum correlation coefficient between incorrect pairings. The region of overlap indicates nonlinear phases for which correlation cannot reliably determine correspondence.



(a) (b) (c)

Fig. 3. Data association across a nonlinear phase all-pass filter: tunable filter (a) yields correlations (b) and mutual information (c).

Figure 3(c) shows an estimate of mutual information between the Fourier spectra of A and B, constructed in the manner outlined above. As α increases, the mutual information estimate assumes a steady-state value which remains separated from the baseline estimate and can accurately determine association.

The second example relates observations of non-overlapping Fourier spectra. Suppose that we observe a time series and would like to determine whether some higher-frequency observations are unrelated, or are a result of observing some nonlinear function (and thus harmonics) of the original measurements. We simulate this situation by creating two independent signals, passing them through a nonlinearity, and relating high-passed and low-passed observations. Sensor A observes the signals' lower half spectrum, and sensor B their upper half.

Synthetic data illustrating this can be seen in Figures 4-5. For Figure 4 we create a narrowband signal whose center frequency is modulated at one of two different rates, and pass it through a cubic nonlinearity. In the resulting filtered spectra (shown in Figure 4(a-d)), the correct pairing is clear by inspection. Scatterplots of the trained features (see Figure 4(e-h)) show that indeed, features of the correct pairings have high mutual information while incorrect pairings have nearly independent features.

Figure 5 shows the same test repeated with wideband data – Gaussian noise is passed through a cubic nonlinearity, and the resulting signal is separated into high- and low-frequency observations, shown in Figure 5(a-d). The resulting structure is less obvious, both visually and to our estimates of mutual information (Figure 5(e-h)), but the correct pairing is still found.

5 Extension to Many Sources

For the problem described above, the presence of only two targets means the data association problem can be expressed as a test between two hypotheses. However, as the number of targets is increased, the combinatorial nature of the hypothesis test makes evaluation of each hypothesis infeasible. Approximate methods which determine a correspondence without this computational burden

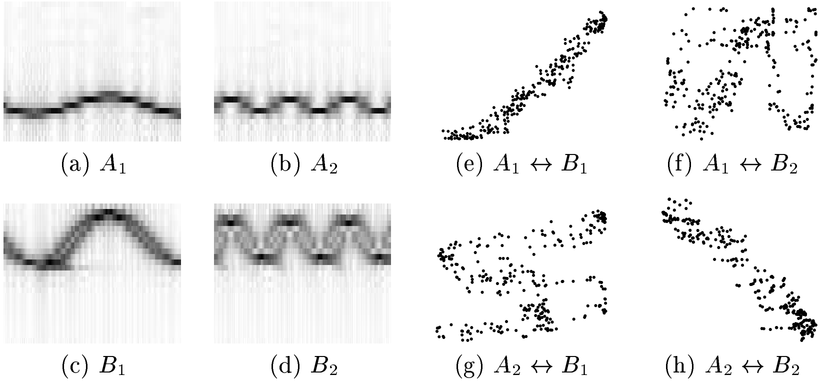


Fig. 4. Associating non-overlapping harmonic spectra: the correct pairing of data sets (a-d) is easy to spot; the learned features yield MI estimates which are high for correct pairings (e,h) and low for incorrect pairings (f,g).

offer an alternative which may be particularly attractive in the context of sensor networks. We describe an extension of the above method to perform data association between many targets without requiring evaluation of all hypotheses.

Let us re-examine the problem of Section 2, but allow both sensors to receive separate observations from M independent targets, denoted $A_1, \ldots, A_M$ and $B_1, \ldots, B_M$. One may still apply estimates of MI to approximate the hypothesis test as described in Section 2.2, but direct application will require that mutual information be estimated for each of the M^2 data pairs – a potentially costly operation.

However, we suggest an approximate means of evaluating the same test which does not compute each MI estimate. We can solve the data association problem by finding features which summarize *all* the signals received at a particular sensor. A test can then be performed on the learned feature coefficients directly, rather than computing all individual pairwise likelihoods.

Let us denote the concatenation of all signals from sensor A by $[A_1, \ldots, A_M]$. One can learn features which maximize mutual information between this concatenated vector and a particular signal B_j; we denote the feature of B_j by $f_A^{B_j}$, and the feature of $[A_1, \ldots, A_M]$ by $f_j^{[A_1,\ldots,A_M]}$.

Again, let us consider the linear statistics of Section 3.2:

$$f_A^{B_j} = \tanh(\sum_i w_i B_j^i) \tag{18}$$

$$f_j^{[A_1,\ldots,A_M]} = \tanh(\sum_{i,k} w_{i,A_k} A_k^i) \tag{19}$$

where A_k^i (B_j^i) indicates the i^{th} dimension of the signal A_k (B_j).

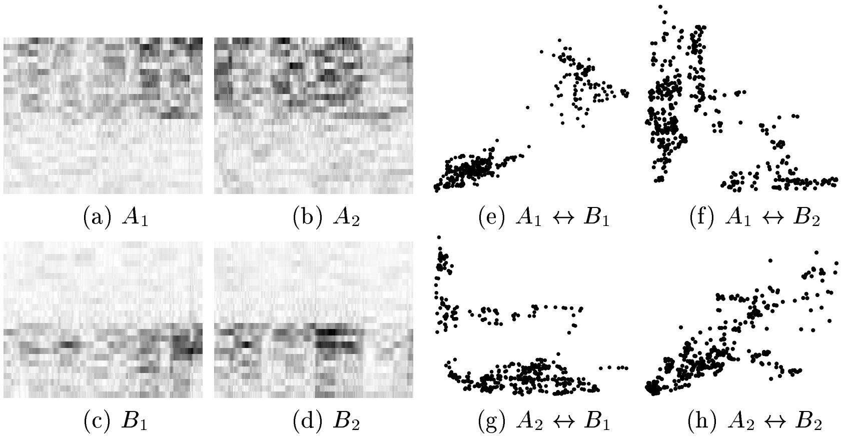


Fig. 5. Associating non-overlapping wideband harmonic spectra: though the correct pairing is harder to see than Figure 4, the estimated MI is still higher for the correct hypothesis (e,h).

We now consider tests based on the absolute deviation of the feature coefficients for each signal A_k:

$$\sum_i |w_{i,A_k}|$$

Under the assumption of independent sources, mutual information exists only between the correctly associated signals; i.e. if A_s and B_t represent a correct association, we have

$$\begin{aligned} I(A_s; B_t) &= I([A_1, \ldots, A_M]; B_t) \\ &= I(A_s; [B_1, \ldots, B_M]) \end{aligned}$$

We may then analyze the mutual information of a particular feature

$$\begin{aligned} I(f_t^{[A_1,\ldots,A_M]}; f_A^{B_t}) &= I(\tanh(\sum_{i,k} w_{i,A_k} A_k^i); f_A^{B_t}) \\ &= I(\sum_{i,k} w_{i,A_k} A_k^i; f_A^{B_t}) \\ &= \sum_k I(\sum_i w_{i,A_k} A_k^i; f_A^{B_t}) \\ &= I(\sum_i w_{i,A_s} A_s^i; f_A^{B_t}) \end{aligned}$$

Thus, for $k \neq s$ the weights w_{i,A_k} have no contribution to the mutual information. This tells us that among all features with maximal MI, the one with minimum absolute deviation $\sum_{i,k} |w_{i,A_k}|$ has support *only* on A_s. Whether distributions exist such that no linear feature captures dependence (i.e. $I(f_t^A; f_A^{B_t}) = 0$ for all linear f) is an open question.

As a means of exploiting this property, we impose a regularization penalty on the feature coefficients during learning. In particular, we augment the information gradient on the concatenated vector feature with a sparsity term, giving

$$\frac{\partial I(f_j^{[A_1,\ldots,A_M]}; f_A^{B_j})}{\partial w_{i_0,A_{k_0}}} - \alpha \max_{i,k} |w_{i,A_{k \neq k_0}}| \tag{20}$$

where the parameter α controls the strength of the regularization. This imposes a penalty on the absolute deviation of the weights which is proportional to the maximum weight from a *different* signal, giving sparse selection of signals – if only one of the M signals has nonzero coefficients, it has no regularization penalty imposed.

A decision can be reached more efficiently using the coefficient deviations, since only a few ($\mathcal{O}(M)$) statistics must be learned; a simple method such as greedy selection or the auction algorithm may be applied to determine the final association.

In the following example, we show the application of this technique to associating harmonics of wideband data passed through a nonlinearity; each of four signals is created in the same manner as those of the final example in Section 4. The signals' Fourier coefficients are shown in Figure 6; sensor A observes the lower half-spectrum and sensor B the upper. For demonstration purposes, we calculate statistics both for each B_k with $[A_1, \ldots, A_M]$, and each A_k with $[B_1, \ldots, B_M]$. Again, we use the ISE approximation of Equation (16) to calculate the information gradient.

Statistics trained in this way are shown in the upper half of Figure 7. To see how one would use these statistics to determine association, we can write the total absolute deviation of the statistic coefficients grouped by observation, and normalize by its maximum. This gives us the pairwise values shown in the lower part of Figure 7. In this example, a greedy method on either set of statistics is sufficient to determine the correct associations. More sophisticated methods might compute and incorporate both sets into a decision.

6 Discussion

We have seen that the data association problem may be characterized as a hypothesis test between *factorizations* of a distribution. An information-theoretic analysis led to a natural decomposition of the hypothesis test into terms related to prior modelling assumptions and terms related to statistical dependence. Furthermore, this analysis yielded insight into how one might perform data association in a principled way in the absence of a prior model. The approach described is similar to a nonparametric generalized likelihood ratio test.

In addition, we have presented an algorithm which utilizes these principles for the purposes of performing data association. This allows us to perform correspondence tests even when the source densities are unknown or there is uncertainty in the signals' propagation by learning statistics which summarize the

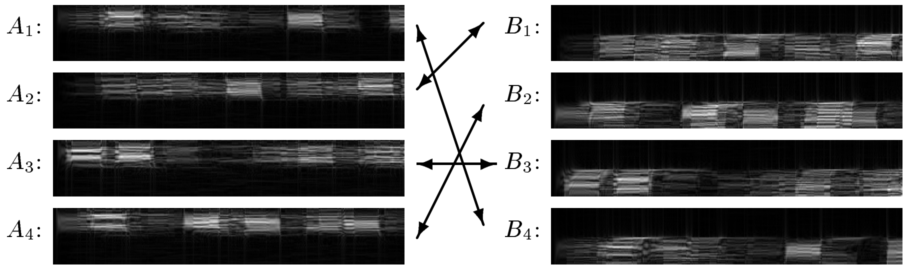
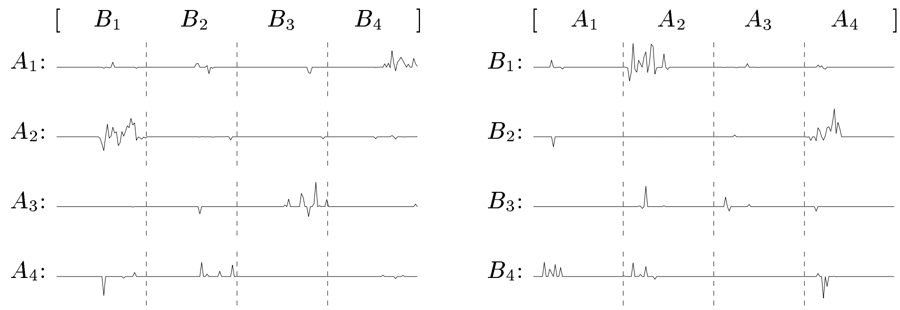


Fig. 6. Associating many signal pairs: a naive approach to finding the association above would require 4^2 estimates of mutual information.



0.097	0.202	0.144	**1.000**
1.000	0.029	0.013	0.040
0.005	0.094	**1.000**	0.033
0.746	**1.000**	0.000	0.131

0.058	**1.000**	0.029	0.041
0.084	0.001	0.016	**1.000**
0.000	1.000	**0.666**	0.185
1.000	0.629	0.000	0.758

Fig. 7. Statistics learned on concatenated signals (above); each feature's region of support indicates probable associations. The row-normalized absolute sum (L^1 norm) of the statistics subdivided by signal index (below) may be used to determine correspondence; bold type indicates the correct association

mutual information between observed data vectors in a compact form. This was equivalent to approximating the likelihood ratio test with mutual information estimates in a low-dimensional space.

We have also suggested an approximate method of determining correspondence between larger signal sets based on the same techniques. Although this does not correspond directly to the optimal hypothesis test, it has the advantage that it does not require that mutual information be estimated for all M^2 signal pairs. Finally, we demonstrated the efficacy of this method with experiments on synthetic data.

References

1. T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.
2. S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.
3. E. J. Kelly. An adaptive detection algorithm. *IEEE Transactions on Aerospace and Electrical Systems*, 22(1):115–127, 1986.
4. J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Math. Stat. Sci.*, 6(1):17–39, June 1997.
5. B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York, 1986.
6. J.W. Fisher III and J.C. Principe. A methodology for information theoretic feature extraction. In A. Stuberud, editor, *International Joint Conference on Neural Networks*, pages ?–?, 1998.
7. J. W. Fisher III, A. T. Ihler, and P. Viola. Learning informative statistics: A nonparametric approach. In S. A. Solla, T. K. Leen, and K-R. Müller, editors, *Neural Information Processing Systems 12*, 1999.
8. A. Ihler, J. Fisher, and A. S. Willsky. Nonparametric estimators for online signature authentication. In *International Conference on Acoustics, Speech, and Signal Processing*, May 2001.
9. C. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, 1995.