Nonparametric Hypothesis Tests for Statistical Dependency

Alexander T. Ihler, *Student Member, IEEE*, John W. Fisher, *Member, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—Determining the structure of dependencies among a set of variables is a common task in many signal and image processing applications, including multitarget tracking and computer vision. In this paper, we present an information-theoretic, machine learning approach to problems of this type. We cast this problem as a hypothesis test between factorizations of variables into mutually independent subsets. We show that the likelihood ratio can be written as sums of two sets of Kullback–Leibler (KL) divergence terms. The first set captures the structure of the statistical dependencies within each hypothesis, whereas the second set measures the details of model differences between hypotheses. We then consider the case when the signal prior models are unknown, so that the distributions of interest must be estimated directly from data, showing that the second set of terms is (asymptotically) negligible and quantifying the loss in hypothesis separability when the models are completely unknown. We demonstrate the utility of nonparametric estimation methods for such problems, providing a general framework for determining and distinguishing between dependency structures in highly uncertain environments. Additionally, we develop a machine learning approach for estimating lower bounds on KL divergence and mutual information from samples of high-dimensional random variables for which direct density estimation is infeasible. We present empirical results in the context of three prototypical applications: association of signals generated by sources possessing harmonic behavior, scene correspondence using video imagery, and detection of coherent behavior among sets of moving objects.**

***Index Terms*—Data association, factorization, hypothesis testing, independence tests, kernel density estimates, Kullback–Leibler divergence, mutual information, nonparametric.**

I. Introduction

DETERMINING the structure of statistical dependencies among a set of variables is a task common to many signal and image processing applications, including multitarget tracking, perceptual grouping, and multisensor data fusion. In many of these applications, it is difficult to specify a model for the data *a priori* due to lack of calibration, unknown environmental conditions, and complex or nonstationary interrelationships among sources. Estimating the dependency structure from the observed data without a prior model is of importance not only as an end in itself for applications such as data association in multitarget tracking and the correspondence problem in computer vision but also as an initial step in constructing signal models. In this paper, we present an information-theoretic, machine learning approach to structure discovery problems, focusing on the issues that arise when prior signal models are unavailable. We suggest an approach for learning informative statistics from data that is particularly applicable when the data is high dimensional, yet highly structured.

To be precise, we consider a class of hypothesis tests that we refer to as *factorization tests*. The primary goal of such tests is to determine the grouping of variables into a dependency structure. These tests have the following characteristics.

- Individual hypotheses specify a partitioning of the full set of variables into disjoint subsets, where
  - — the variables within each subset are dependent, and
  - — the subsets are mutually independent.
- The parameters of each component distribution may be different under each hypothesis.

A consequence of the first characteristic is that for each hypothesis, the distribution over the full set of variables factors into a product of the distributions of each subset of variables specified by the hypothesis. It will be useful to distinguish the cases in which the distribution parameterizations are known from the cases in which they are not. We will refer to the former as *parametric* factorization tests and the latter as *nonparametric* factorization tests.

Tests of this type commonly arise in problems such as data association [1] and perceptual grouping [2]. We show that this class of hypothesis tests naturally decomposes into two parts. The first captures statistical dependency within each subset, whereas the second summarizes differences between the parameterizations of hypotheses. Additionally, we show that in the absence of a parameterized model, nonparametric approaches may be utilized, leading to a general framework for determining and distinguishing between dependency structures and quantifying the increase in difficulty of such factorization tests for highly uncertain environments.

Finally, application of such methods to high-dimensional data (e.g., imagery or spectrograms) presents an additional problem, not only in terms of computational complexity but the infeasibility of estimating high-dimensional distributions as well. We address these issues by developing bounds on the log-likelihood ratio using low-dimensional statistics of the observed signals. This leads to a machine learning procedure for learning informative statistics. We demonstrate our approach on three prototypical applications: association of signals generated by sources possessing harmonic structure, scene correspondence

Manuscript received July 2, 2003; revised December 29, 2003. This work was supported in part by the Air Force Office of Scientific Research under grant F49620-00-0362 and by ODDR&E MURI through the ARO under grant DAAD19-00-0466. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chin-Hui Lee.

A. T. Ihler and A. S. Willsky are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology Cambridge, MA 02139 USA (e-mail: ihler@mit.edu; willsky@mit.edu).

J. W. Fisher is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: fisher@csail.mit.edu).

Digital Object Identifier 10.1109/TSP.2004.830994

1053-587X/04$20.00 © 2004 IEEE

using video imagery, and detection of coherent behavior among sets of moving objects.

II. Asymptotic Analysis of Factorization Tests

The general problem we consider is as follows: We have a set of d signals or data sources $S = \{x_1, \ldots, x_d\}$ and a number of hypotheses for how these signals are partitioned into mutually independent groups. As in all likelihood-based methods, the key quantities are likelihood ratios associated with pairs of hypotheses. In the discussion to follow, we denote such a generic pair of hypotheses by H_1, H_2.

Each of these two hypotheses has associated with it a partition of the variables into dependency sets. For hypothesis $H_i (i = 1, 2)$, we denote the number of subsets by M_i and the jth subset by S_j^i, with

$$S_j^i \cap S_k^i = \emptyset, k \neq j \quad \text{and} \quad \bigcup_j S_j^i = S.$$

The joint statistical model under H_i is expressed as a product

$$p_{H_i}(x_1, \ldots, x_d) = \prod_{j=1}^{M_i} p_{H_i}\left(S_j^i\right). \tag{1}$$

In analyzing the likelihood ratio between hypotheses H_1 and H_2, it is also useful to identify those subsets of variables that are dependent under *both* hypotheses. Specifically, define the *intersection sets* by

$$S_{jk}^{12} = S_j^1 \cap S_k^2 \quad \forall\, j, k. \tag{2}$$

Note that these subsets also form a (generally finer) partition of S and, thus, can be thought of as implicitly specifying yet a third possible factorization of the joint probability distribution. While this factorization is in general not one of the hypotheses itself, this factorization plays an important role in the analysis, especially in the case in which we do not have prior models for the distributions under any of the hypotheses [e.g., the distributions on the right-hand side of (2)].

In one simple but important context (discussed further in Section III), namely, that in which the dependency subsets S_j^i under each hypothesis consist of pairs of signals, the intersection sets take particularly simple forms. For example, consider a set of four signals and two hypotheses (and the sets defined by their factorizations)

$$\begin{aligned} H_1&: p(x_1, x_2, x_3, x_4) = p_{H_1}(x_1, x_2) p_{H_1}(x_3, x_4) \\ &\Rightarrow S_1^1 = \{x_1, x_2\}, S_2^1 = \{x_3, x_4\} \\ H_2&: p(x_1, x_2, x_3, x_4) = p_{H_2}(x_1, x_4) p_{H_2}(x_2, x_3) \\ &\Rightarrow S_1^2 = \{x_1, x_4\}, S_2^2 = \{x_2, x_3\}. \end{aligned} \tag{3}$$

In this case, the factorization implied by the resulting intersection sets

$$\begin{aligned} S_{11}^{12} &= \{x_1\} \quad S_{12}^{12} = \{x_2\} \\ S_{21}^{12} &= \{x_4\} \quad S_{22}^{12} = \{x_3\} \end{aligned} \tag{4}$$

is that of complete independence of the variables, i.e., $p_{H_i}(x_1) p_{H_i}(x_2) p_{H_i}(x_3) p_{H_i}(x_4)$.

It is important to emphasize that in general, not only is each hypothesis distinguished by the partitioning into dependency sets but also by any assumed probability distribution for the variables in each set. In classic hypothesis-testing problems, the differences in those distributions provides useful information (e.g., the problem of deciding if a single Gaussian variable is zero mean or has mean two is a well-defined hypothesis testing problem). In the case on which we focus here, such prior models of distributions are unknown, and *all* we seek is to determine the dependency structure. The key, as we will see, is distinguishing the part of the likelihood ratio that depends on dependency structure alone from the part that exploits differences in assumed models.

A. Parametric Factorization Tests

We are primarily interested in the properties of nonparametric factorization tests. However, it is instructive to first consider the asymptotic properties of the fully specified, parametric factorization test.

If the model parameters under each hypothesis are known, we can write the (normalized) log-likelihood ratio test between H_1 and H_2, given N i.i.d. observations of the x, indexed by $t \in \{1 \ldots N\}$ as

$$\begin{aligned} \frac{1}{N} \log L &= \frac{1}{N} \sum_t \log \frac{\prod_j p_{H_1}\left(S_{j,t}^1\right)}{\prod_j p_{H_2}\left(S_{j,t}^2\right)} \\ &= \frac{1}{N} \sum \log \frac{\prod_j p_{H_1}\left(S_{j,t}^1\right)}{\prod_{j,k} p_{H_1}\left(S_{jk,t}^{12}\right)} \\ &\quad \cdot \frac{\prod_{j,k} p_{H_1}\left(S_{jk,t}^{12}\right)}{\prod_j p_{H_2}\left(S_{j,t}^2\right)}. \end{aligned} \tag{5}$$

If H_1 is true, this quantity approaches the following limit:

$$\begin{aligned} \lim_{N \to \infty} \frac{1}{N} \log L &= D\left(\prod_j p_{H_1}\left(S_j^1\right) \middle\| \prod_{j,k} p_{H_1}\left(S_{jk}^{12}\right)\right) \\ &\quad + D\left(\prod_{j,k} p_{H_1}\left(S_{jk}^{12}\right) \middle\| \prod_j p_{H_2}\left(S_j^2\right)\right) \end{aligned} \tag{6}$$

(proof given in Appendix II-A), and similarly, if H_2 is true, it approaches the limit

$$\begin{aligned} \lim_{N \to \infty} \frac{1}{N} \log L &= -D\left(\prod_j p_{H_2}\left(S_j^2\right) \middle\| \prod_{j,k} p_{H_2}\left(S_{jk}^{12}\right)\right) \\ &\quad - D\left(\prod_{j,k} p_{H_2}\left(S_{jk}^{12}\right) \middle\| \prod_j p_{H_1}\left(S_j^1\right)\right) \end{aligned} \tag{7}$$

where $D(p_{H_1} \| p_{H_2})$ denotes the Kullback–Leibler (KL) divergence between p_{H_1} and p_{H_2}. While it is not a surprise that KL-divergence terms arise (cf. [3]), the result expressed in (6) and (7) allows us both to separate the parts of the log-likelihood that deals with dependency structure exclusively from the part that takes advantage of differences in assumed statistical models (and hence will be unavailable to us in the case in which prior models are not given). In particular, under either hypothesis, the limit of the likelihood ratio test can be decomposed into two KL divergence terms. The first term captures differences in the dependency structure between hypotheses (e.g., under H_1, the set of variables S_j^i are dependent, but under H_2, this set

may be further decomposed into mutually independent sets S_{jk}^{12}, $k = 1, 2, \ldots$). In contrast, the second terms in (6) and (7) capture differences stemming from different assumed statistical models under the two hypotheses (for example, one model having mean zero versus another with mean two). This decomposition is illustrated with a specific example in Section II-C; but first, we highlight the differences that arise for a *nonparametric* factorization test.

B. Nonparametric Factorization Tests

Nonparametric factorization tests are distinguished from parametric tests in that although the factorization under each hypothesis is specified, the model parameters are not. If we replace each density in (5) with a distribution $\hat{p}_{H_i}$ estimated from the data, use a consistent density estimator, and have the luxury of collecting samples under *each* hypothesis, then (6) and (7) still hold in the limit.

A more interesting case, however, is when we estimate the models and perform the hypothesis test *at the same time*, i.e., we must both learn the models and distinguish between the hypotheses using the same set of data. Of course, in this case, the available data will come from one hypothesis or the other (assuming, as we do, that "truth" does indeed correspond to one of our hypothesized factorizations). Consequently, while the model estimates formed for the correct hypothesis will be asymptotically accurate, the estimates for the incorrect hypothesis will not (since we are not basing them on data corresponding to that hypothesis) but will be biased in a manner that best matches the available data. This fact makes the hypothesis testing problem more difficult in this case.

In particular, if the data are generated from a hypothesis under which variables x_1 and x_2 are independent, then the estimate of their joint distribution under any other hypothesis—in particular one that allows these two variables to be dependent—will asymptotically converge to the product of their marginal distributions. More generally, if the data are generated under hypothesis $H_1(H_2)$, but we estimate densities assuming the *factorization* of $H_2(H_1)$, then the resulting estimate will converge to the factorization described by the *intersection set*, as defined previously. Specifically, assuming consistent density estimates (e.g., kernel density estimates; see Appendix I), we have

$$\begin{aligned}
\text{if } H_1 \text{ is true,} \quad & \hat{p}_{H_1} \to p_{H_1}(S^1) = \prod_j p_{H_1}\left(S_j^1\right) \\
& \hat{p}_{H_2} \to p_{H_1}(S^2) = \prod_{j,k} p_{H_1}\left(S_{jk}^{12}\right) \\
\text{if } H_2 \text{ is true,} \quad & \hat{p}_{H_1} \to p_{H_2}(S^1) = \prod_{j,k} p_{H_2}\left(S_{jk}^{12}\right) \\
& \hat{p}_{H_2} \to p_{H_2}(S^2) = \prod_j p_{H_2}\left(S_j^2\right) \qquad (8)
\end{aligned}$$

(where we have used the shorthand $p_{H_i}(S^l) = \prod_j p_{H_i}\left(S_j^l\right)$ and similarly for S^2). Thus, when $\hat{p}_{H_i}$ models the correct factorization, the estimate converges to the true distribution; conversely, when enforcing the structure under the incorrect hypothesis, the estimate converges to a factorization consistent with the intersection set.

This is perhaps best illustrated with a short example drawn from (3). Suppose that H_1 is true; then, if we assume H_2, our estimate is $\hat{p}_{H_2}(x_1, x_2, x_3, x_4) = \hat{p}_{H_2}(x_1, x_4)\hat{p}_{H_2}(x_2, x_3)$. In the limit, we have, e.g., $\hat{p}_{H_2}(x_1, x_4) \to p_{H_1}(x_1, x_4) = p_{H_1}(x_1)p_{H_1}(x_4)$ (as these are independent under H_1), and we obtain a factorization described by the intersection sets.

In fact, the intersection set can be thought of as a null hypothesis for the test between factorizations H_1 and H_2; we therefore denote this by H_0. When the factorization of one hypothesis is entirely contained within the other, H_0 corresponds to the more factored of the two possibilities, as is typical for an independence test.

The loss of test power when distributions must be learned is similar to issues that arise in generalized likelihood ratio (GLR) tests [4]. However, a nonparametric factorization test based on kernel methods makes assumptions only about the smoothness of the distributions and not their form.

As a consequence of estimating densities from samples drawn from a single hypothesis, the limit of the likelihood ratio between *estimated* densities ($\hat{L}$) is expressed solely in terms of the hypotheses' divergence from the intersection factorization. Under H_1, this is

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N} \log \hat{L} &= E_{H_1}\left[\log \frac{\prod_j \hat{p}_{H_1}\left(S_j^1\right)}{\prod_j \hat{p}_{H_2}\left(S_j^2\right)}\right] \\
&= D\left(\prod_j p_{H_1}\left(S_j^1\right) \middle\| \prod_{j,k} p_{H_1}\left(S_{jk}^{12}\right)\right) \qquad (9)
\end{aligned}$$

and similarly under H_2

$$\lim_{N\to\infty} \frac{1}{N} \log \hat{L} = -D\left(\prod_j p_{H_2}\left(S_j^2\right) \middle\| \prod_{j,k} p_{H_2}\left(S_{jk}^{12}\right)\right). \qquad (10)$$

Note that as a result of estimating both models from the same data, the KL-divergence terms stemming from model mismatch in (6) and (7) have vanished. The value of these divergence terms quantifies the increased difficulty of discrimination when the models are unknown and illustrates the primary distinction between parametric and nonparametric factorization tests.

The limits in (9) and (10) can be expressed independent of which hypothesis is correct (assuming one of the H_i is correct) as

$$\begin{aligned}
& D\left(\prod_j p\left(S_j^1\right) \middle\| \prod_{j,k} p\left(S_{jk}^{12}\right)\right) \\
& \qquad - D\left(\prod_j p\left(S_j^2\right) \middle\| \prod_{j,k} p\left(S_{jk}^{12}\right)\right) \qquad (11)
\end{aligned}$$

since one of these two terms will be zero.

While many issues arise in the context of maximum likelihood tests for dependence structure, for example model complexity [5] and significance [4], our primary focus is on *estimating* these KL-divergence terms (equivalently likelihoods) in problems with high-dimensional, complex joint distributions.

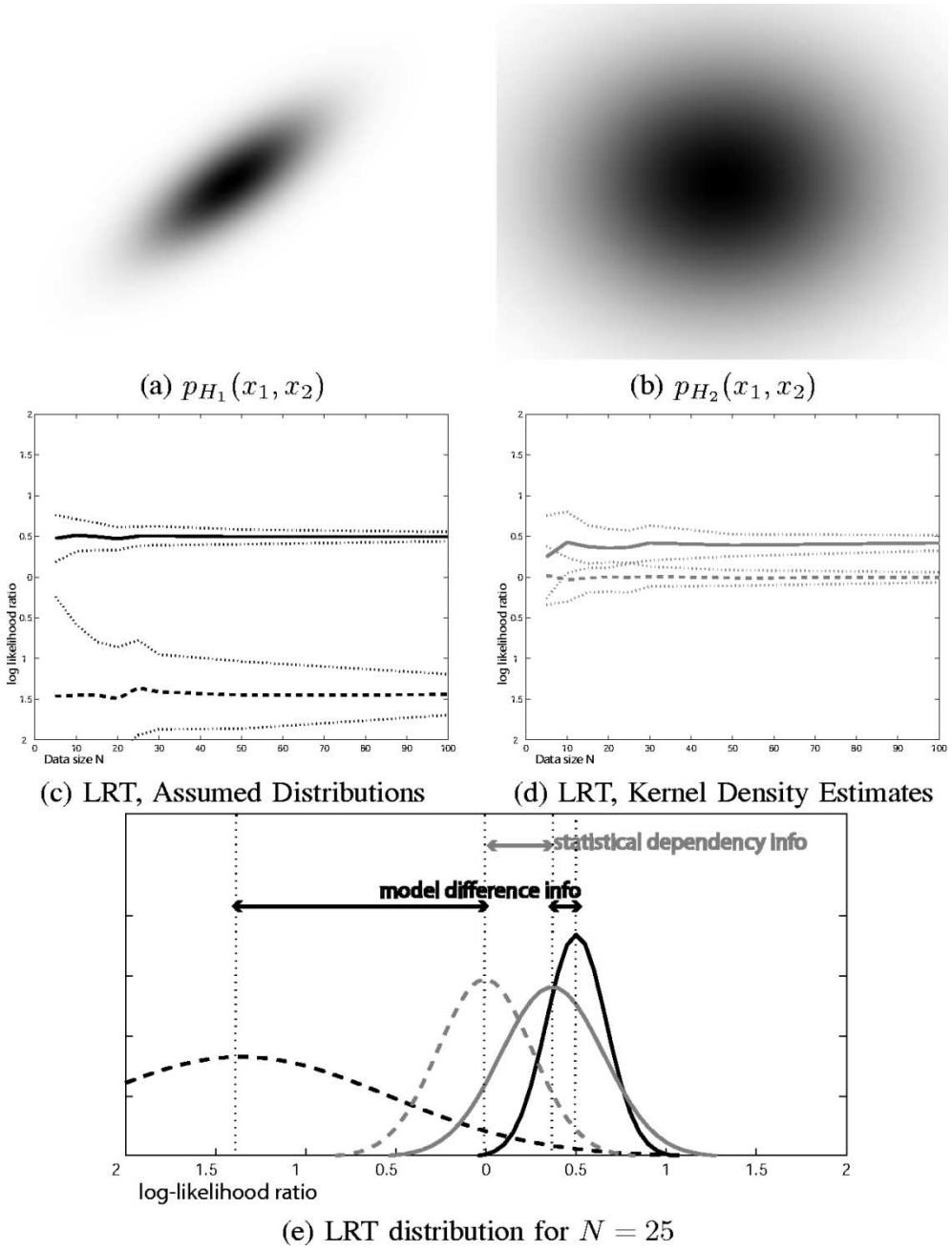


Fig. 1. Performance as a function of parameter assumptions I. When the two alternatives p_{H_1} and p_{H_2} are known *a priori*, the (c) parametric likelihood ratio tests under H_1 (solid) and H_2 (dashed) benefit from the differences between the model parameters. Using (d) a nonparametric estimate with only the test data has less separation (only that due to the factorization information). (e) Respective contributions are shown in the cross-section.

C. Comparison: Parametric versus Nonparametric Factorization Tests

We illustrate the previous analysis with a simple bivariate Gaussian example. Suppose we have two hypotheses

$$\begin{aligned} H_1&: p(x_1, x_2) = \mathcal{N}(0, \Sigma_1) \\ H_2&: p(x_1, x_2) = \mathcal{N}(0, \Sigma_2) \end{aligned} \tag{12}$$

where

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1^2 \\ \rho\sigma_1^2 & \sigma_1^2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}. \tag{13}$$

Note that this is a parametric factorization test, in that we specify both a *factorization* (independence in H_2, dependence in H_1) and a *parameterization* (that the distributions are Gaussian and have parameters σ_1, σ_2, ρ). These two distributions are shown in Fig. 1(a) and (b).

For this case, the expected log-likelihood ratio can be computed in closed form. When H_1 is true, the result is

$$\begin{aligned} D(p_{H_1} \| p_{H_2}) &= I(x_1; x_2) + D\left(\prod_j p_{H_1}(x_j) \| p_{H_2}(x_1, x_2) \right) \\ &= \left[-\frac{1}{2} \log(1 - \rho^2) \right] + \left[\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log \frac{\sigma_1^2}{\sigma_2^2} \right] \end{aligned} \tag{14}$$

and when H_2 is true

$$-D(p_{H_2}||p_{H_1}) = [0] - \left[\frac{\sigma_2^2}{(1-\rho^2)\sigma_1^2} - 1 - \log\frac{\sigma_2^2}{\sigma_1^2} + \frac{1}{2}\log(1-\rho^2)\right]. \quad (15)$$

Let us now consider the equivalent nonparametric factorization test, in which only the factorization is specified. In this case, we learn densities that reflect the factorization under each hypothesis from the observed samples. Necessarily, these samples are collected under a single hypothesis, and as shown by the previous analysis, the model divergence terms disappear, resulting in decreased separability between the two hypotheses. For this case, under H_1, the expected log-likelihood ratio converges to

$$D(\hat{p}_{H_1}||\hat{p}_{H_2}) = \left[-\frac{1}{2}\log(1-\rho^2)\right] \quad (16)$$

and under H_2

$$-D(\hat{p}_{H_2}||\hat{p}_{H_1}) = [0]. \quad (17)$$

Monte Carlo simulations of both cases are shown in Fig. 1(c) (parametric tests) and Fig. 1(d) (nonparametric tests). In both cases, we compare the mean log-likelihood ratio under H_1 (solid) and H_2 (dashed). Variance estimates are given by the dotted lines; the separability of the two hypotheses increases as a function of the number of observed data N. In this example, the parametric test benefits greatly from its (correct) knowledge of the distribution. A cross-section from both kinds of tests taken at $N = 25$ is shown in Fig. 1(e), illustrating the separability of both tests and the relative contribution of dependency information and model parameters.

As Fig. 1 indicates, if the assumed Gaussian models for these two hypotheses are in fact correct, we gain some performance by using these models. However, if these detailed models are incorrect, there can be significant loss in performance for the primary goal of interest to us, namely, that of determining the correct *factorization* of the distribution. Indeed, for data that have one or the other of these factorization structures but have distributions that differ from the Gaussian models, the use of a test based on these models may fail catastrophically.

To illustrate this latter point, we intentionally choose two particularly difficult densities as the true underlying distributions. Specifically, let the true data distributions be Gaussian sums, located such that under either hypothesis, the two components x_1, x_2 are uncorrelated. However, in one case [Fig. 2(a)], the variables are dependent, whereas in the other [Fig. 2(b)], they are independent. Moreover, the parameters have been chosen so that the variances of the variables under the *dependent* distribution match the variance σ_2 in (13), whereas the variances under the independent distribution match the variance σ_1.

Using a nonparametric estimate of the likelihood, we correctly estimate the statistical dependency and thus determine the factorization [Fig. 2(d)]. Again, we show the likelihood ratio under H_1 as solid, and under H_2 as dashed, along with their respective variance estimates. The model-based test [Fig. 2(c)], however, not only fails to find any statistical dependency (both means are less than zero) but also rates the model with the correct factorization as having a lower likelihood.

III. Pairwise Association of High-Dimensional Data

In the face of model uncertainty, machine learning and data-based methods are appealing; however, the practical aspects of using nonparametric density methods raise a number of important issues. In particular, when observations are high-dimensional or there are many observed variables, direct estimation of the probability densities in (9)–(11) becomes impractical due to sample and computational requirements. To render this problem tractable, we apply a learning approach to estimate optimized information-theoretic quantities in a low-dimensional space.

Data association between pairs of observations is a special case of factorization tests. We illustrate aspects of this problem with the following example. Suppose that we have a pair of widely spaced acoustic sensors, each consisting of a small array of microphones and producing both an observation of a source and an estimate its of bearing. This in itself is insufficient to localize a source; however, triangulation of bearing measurements from both sensors can be used to estimate the target location. When there is only one target, this is a relatively simple problem.

Complications arise when there are multiple targets within each sensors' field of view. For two targets, each sensor determines two bearings, yielding four possible locations for the two targets, as depicted in Fig. 3. Using bearing information alone, one cannot determine which pair of locations is correct and which is not. However, under the assumption that the sources are statistically independent, this can be cast as a test between factorizations of the source estimates. This interpretation allows us to test an association even in the case that the source statistics and/or transmission medium are poorly specified or completely unknown.

This problem is further complicated by the fact that the sensors' observations may have long temporal dependency or be of high dimension (for example video images), either of which can render density estimation infeasible. However, the hypothesis test may not require that these distributions be estimated directly since [as evidenced by (6), (7), (9), and (10)] what we really wish to estimate is a KL-divergence value. We avoid the difficulties of density estimation in high dimension by instead estimating a lower bound on divergence via statistics whose dimension is controlled by design.

A. Mutual Information as a Special Case

When we are interested only in associations between pairs of variables, the terms related to statistical dependence within a given hypothesis simplify to the sum of the mutual information between each pair. In other words, each set in H_i's factorization is $S_j^i = \{x_{j_1}, x_{j_2}\}$, and the divergence from the intersection factorization is always a divergence from marginal distributions (leaving out any associations on which the two hypotheses agree):

$$D\left(\prod_j p\left(S_j^i\right)\middle\|\prod_{j,k} p\left(S_{jk}^{12}\right)\right) = \sum_j I\left(S_j^i\right) = \sum_j I(x_{j_1}; x_{j_2}) \quad (18)$$

where $I(x;y)$ is the mutual information (MI) between x and y. As we have already observed, if each variable x_j is high-di-

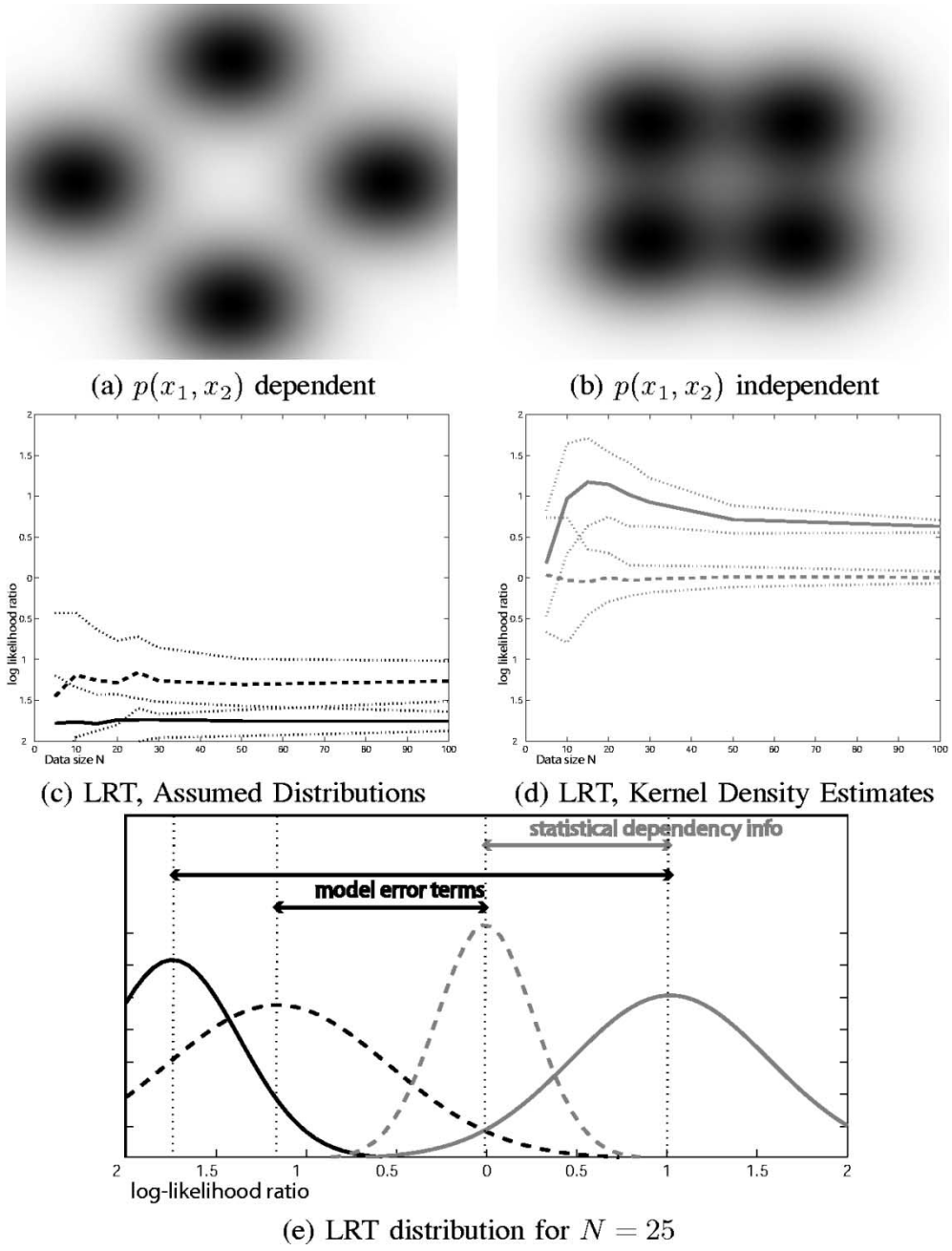


Fig. 2. Performance as a function of parameter assumptions II. If we have only factorization information but attempt to use the same parametric assumptions from Fig. 1, we may severely degrade performance (c). Here, because of the similar marginal distributions, the correct factorization is actually less likely under our correlated Gaussian model. In contrast, (d) nonparametric methods can still estimate the statistical dependence information. (e) The cross-section shows the relative influences.

mensional, direct estimation of even the pairwise mutual information terms becomes difficult. However, a tractable method of estimating mutual information for high-dimensional variables follows from application of the data processing inequality [6]. Specifically, note that

$$D\left(p(x_{j_1}, x_{j_2}) \| p(x_{j_1})p(x_{j_2})\right) = I(x_{j_1}; x_{j_2}) \geq \max_{f_j, g_j} I\left(f_j(x_{j_1}); g_j(x_{j_2})\right) \tag{19}$$

where $f_j(\cdot)$ and $g_j(\cdot)$ are differentiable functions (which we allow to be different for each pair of variables S_j^i). Gradient ascent can be used to maximize the bound; if the functions are scalar (a design choice), this is performed in a two-dimensional (2-D) space. We have chosen to apply kernel (Parzen window) density estimates [7], which can be used to obtain differentiable estimates of the required information theoretic quantities; this is discussed further in Appendix I. However, the focus of this paper is not on the specifics of this optimization, but rather on the utility of the optimized estimate. In fact, it is reasonable to assume that any estimate of mutual information for which we can optimize the functions f, g may be employed.

The left and right sides of (19) achieve equality when the f_j, g_j are sufficient statistics for the data. Thus, if we knew

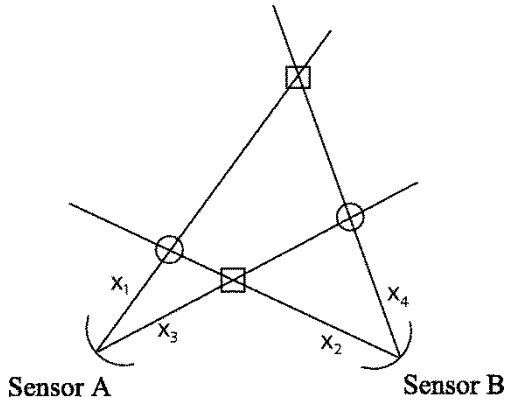


Fig. 3. Data association problem: Two pairs of measurements results in estimated sources at either the circles or the squares, but which one remains ambiguous.

sufficient statistics, we could replace the original log-likelihood ratio (5) with an alternate estimate of its limit (11), requiring only the pairwise distributions of the statistics.

It may be difficult to find low-dimensional sufficient statistics; in fact, in general, they will not exist. However, for *any* set of features, it can be shown (see Appendix II-B) that the following limit holds:

$$\lim_{N\to\infty}\frac{1}{N}\log\hat{L}=\sum_j I_j^1-\sum_j I_j^2+\sum_j D_j^1-\sum_j D_j^2 \quad (20)$$

where, for brevity, we have used the notation

$$\begin{aligned} I_j^i &= I\left(f_j(x_{j_1});g_j(x_{j_2})\right), \quad x_{j_1},x_{j_2}\in S_j^i \\ D_j^i &= D\big(p\left(x_{j_1},x_{j_2}\right)\|p\left(f_j(x_{j_1}),g_j(x_{j_2})\right) \\ &\qquad \cdot p\left(x_{j_1}|f_j(x_{j_1})\right)p\left(x_{j_2}|g_j(x_{j_2})\right)\big). \end{aligned}$$

The divergence terms D_j^i become negligible in direct proportion to the degree to which the functions f, g summarize the statistical dependency between x_{j_1} and x_{j_2}. For sufficient statistics, these divergence terms are exactly zero.

Notice that in (20), only the divergence terms involve high-dimensional measurements; 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 data x_j. We would like to minimize the effect of ignoring these terms on our estimate of the likelihood ratio (20) but cannot estimate the terms directly without evaluating high-dimensional densities. However, by non-negativity of the KL-divergence, we can bound the difference by the sum of the divergences:

$$\left|\sum_j D_j^1-\sum_j D_j^2\right|\le\sum_j D_j^1+\sum_j D_j^2. \quad (21)$$

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 optimizations are decoupled from each other and, thus, may be performed independently. An outline of the hypothesis testing procedure for pairwise interactions is given in Fig. 4.

Given two hypotheses
$H_1 \Rightarrow p(x_1,x_2,x_3,x_4)=p(x_1,x_2)p(x_3,x_4)$
$H_2 \Rightarrow p(x_1,x_2,x_3,x_4)=p(x_1,x_4)p(x_3,x_2)$
with factorization sets
$S_1^1=\{x_1,x_2\}, S_2^1=\{x_3,x_4\}, S_1^2=\{x_1,x_4\}$,
$S_2^2=\{x_3,x_2\}$
and intersection set factorization
$H_0 \Rightarrow p(x_1,x_2,x_3,x_4)=p(x_1)p(x_2)p(x_3)p(x_4)$
1) Estimate each set's divergence (see Appendix I), e.g. for S_1^1,
   a) Estimate $D(p(x_1,x_2)\|p(x_1)p(x_2))=I(x_1;x_2)$ via:
   b) Optimize to find $\hat{I}(S_1^1)=\max_{f,g} I(f(x_1);g(x_2))$
2) Compute the divergence estimates
   a) $\hat{D}(p_{H_1}\|p_{H_0})=\hat{I}(S_1^1)+\hat{I}(S_2^1)$
   b) $\hat{D}(p_{H_2}\|p_{H_0})=\hat{I}(S_1^2)+\hat{I}(S_2^2)$
3) Test

$$\hat{L}=\hat{D}(p_{H_1}\|p_{H_0})-\hat{D}(p_{H_2}\|p_{H_0})\underset{H_2}{\overset{H_1}{\gtrless}}0$$

Fig. 4. Example: Nonparametric factorization tests using mutual information estimated via low-dimensional features.

Finally, it should be pointed out that our estimate of the average log-likelihood is a *difference* of estimated lower bounds on the statistical dependence in the data supporting each hypothesis (11). When either hypothesis is correct, one of these terms is asymptotically negligible.

B. Example: Associating Data Between Two Sensors

We return to the example of associating observations of two sources, each received at two sensors, as depicted in Fig. 3. Specifically, let us assume that each sensor observes nonoverlapping portions of the Fourier spectrum (highpass versus lowpass). We would like to determine the proper association between low- and high-frequency observations. Note that for Gaussian sources, nonoverlapping portions of the spectrum are independent. However, in many cases—e.g., those involving rotating machinery or engines—nonlinearities lead to the presence of harmonics of underlying fundamental frequencies in the source, implying dependencies of variations in different parts of the spectrum. We simulate this situation by creating two independent frequency-modulated signals, passing them through a cubic nonlinearity and relating observations that have been lowpass filtered (x_1,x_3) and highpass filtered (x_2,x_4). We test between the two possible associations:

$$\begin{aligned} H_1&: p(x_1,x_2,x_3,x_4)=p(x_1,x_2)p(x_3,x_4) \\ H_2&: p(x_1,x_2,x_3,x_4)=p(x_1,x_4)p(x_2,x_3). \end{aligned} \quad (22)$$

Synthetic data illustrating this situation can be seen in Fig. 5. Here, we represent the signals $x_1,\ldots,x_4$ by their spectrograms (sequence of windowed Fourier spectra). Sensor A measures x_1 and x_3 (both low frequency), whereas sensor B measures x_2 and x_4 (both high frequency), and the issue is to determine the correct pairing of measurements. In the resulting filtered spectra

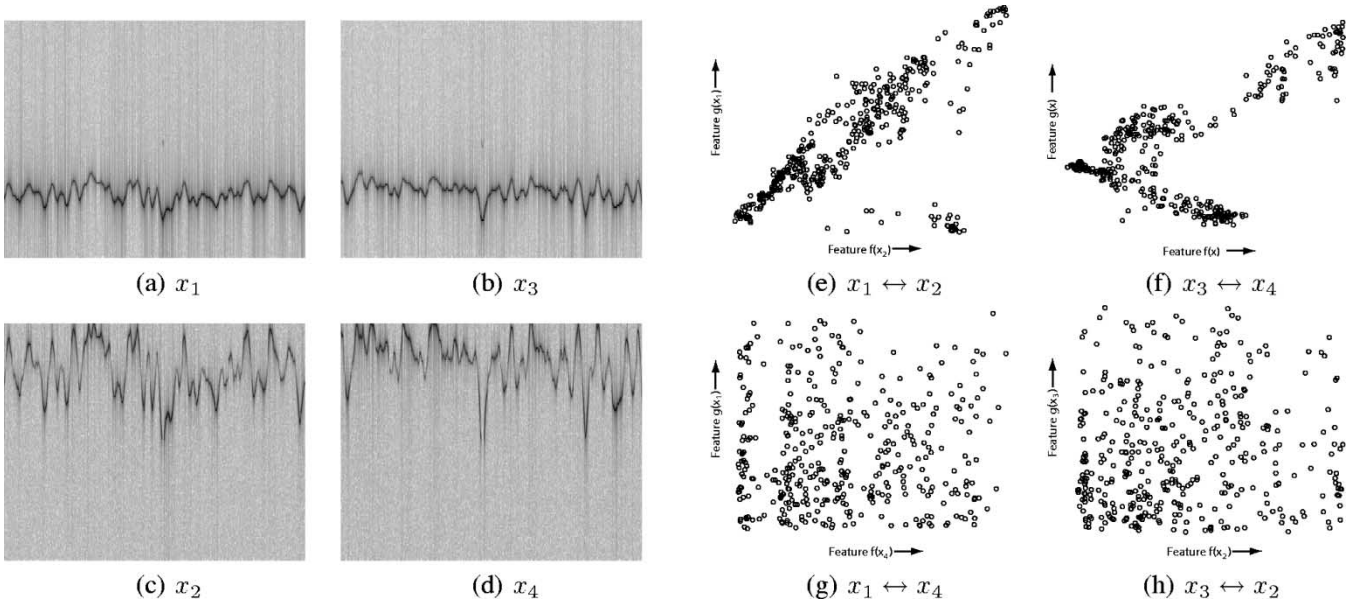


Fig. 5. Associating nonoverlapping harmonic spectra. The correct pairings of data sets (a)–(d) can be seen by inspection; the learned features yield estimates of mutual information that are high for correct pairings (e), (f) and low for incorrect pairings (g), (h).

[the images shown in Fig. 5(a)–(d)], the correct pairings are $x_1 \leftrightarrow x_2$, $x_3 \leftrightarrow x_4$ (which might be ascertained by close inspection).

Using the inequality of (19), let f and g be linear statistics of the Fourier spectra for each time window. Using the procedure described in Appendix I, we optimize these statistics with respect to their mutual information. Note that this is done for each potential pairing of data, as specified by the hypotheses of (22). Scatterplots of the trained features [see Fig. 5(e)–(h)] show that indeed, features of the correct pairings have noticeably higher statistical dependence than the incorrect pairings, the degree of which is quantified by an estimate of mutual information in the feature space.

IV. Testing General Factorizations

For general factorization tests, the KL divergence terms become more complex. In addition to the difficulty associated with high-dimensional measurements, we also have the potential for large *numbers* of variables. Large numbers of variables pose a two-fold problem: both an increase in the number of hypotheses to be tested (a difficulty which we do not attempt to address in this paper) and an increased difficulty in testing any given pair of hypotheses. A straightforward extension of Section III's approach—learning one feature for each variable—may not suffice in that as the number of variables grows, so does the dimensionality of the required density estimate. This motivates an alternate approach that (nominally) decouples the number of variables from the dimensionality of the density estimates. The advantage of this method is that we can control the number of dimensions over which we are required to estimate a density, independent of the number of variables.

A. Feature-Based Divergence Estimation

In contrast to learning a statistic for each of the x_i, we can exploit a superior bound based on a single statistic of all the variables. Specifically, let f be a differentiable function of $\mathbf{x} = [x_1, \ldots, x_d]$, and let us examine the distribution of f under the two hypotheses. For any deterministic function $f(\mathbf{x})$, we can formulate a lower bound on the divergence of H_1 and H_2 as follows:

$$\begin{aligned} D(p_{H_1}(\mathbf{x})\|p_{H_2}(\mathbf{x})) &= D(p_{H_1}(f(\mathbf{x}),\mathbf{x})\|p_{H_2}(f(\mathbf{x}),\mathbf{x})) \\ &= D(p_{H_1}(f(\mathbf{x}))\|p_{H_2}(f(\mathbf{x}))) \\ &\quad + D(p_{H_1}(\mathbf{x}|f(\mathbf{x}))\|p_{H_2}(\mathbf{x}|f(\mathbf{x}))) \\ &\geq D(p_{H_1}(f(\mathbf{x}))\|p_{H_2}(f(\mathbf{x}))). \end{aligned} \tag{23}$$

Consequently, the challenge is to optimize $f(\mathbf{x})$ to maximize the right-hand side of (23). In this paper, we describe a gradient ascent method over parameterized functions f and use an estimate of the KL-divergence, as discussed in Appendix I.

As was the case in Section II, if we have N independent observations $\{\mathbf{x}_t\}$ under *each* hypothesis, we could estimate the marginal distributions $p_{H_1}(f(\mathbf{x}))$, $p_{H_2}(f(\mathbf{x}))$ and, thus, the KL-divergence on the right of (23). However, as in Section II-B, it is more interesting to consider the scenario in which we have only *one* set of samples with which to perform a test between the two hypotheses, and again, these samples are necessarily drawn under a single hypothesis.

Through the use of bootstrap sampling [8], we can obtain samples of f according to the *factorizations* associated with each hypothesis. In essence, this has the same meaning as (8); but rather than estimating the full joint distributions, we merely need to draw samples from them, which can be used to create samples of and estimate the marginal over the feature f.

Explicitly, we obtain a sample of f that adheres to the factorization of hypothesis H_i by independently drawing joint samples of the variables in the set S_j^i for each j and evaluating the function f at this value. This process is illustrated in Fig. 6. Alternately, it may be convenient to draw samples from the sets *without* replacement; this leads to an estimate $\hat{p}_{H_i}(f)$ related to permutation statistics [8]. Either method provides samples from $\hat{p}_{H_i}(\mathbf{x})$ (and, thus, $\hat{p}_{H_i}(f(\mathbf{x}))$), which can be used to estimate

Given two hypotheses

$H_1 \Rightarrow p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2, x_3)p(x_4, x_5)$

$H_2 \Rightarrow p(x_1, x_2, x_3, x_4, x_5) = p(x_1, x_2)p(x_3, x_4, x_5)$

1) Given a set of samples $\{\mathbf{x}_t = [x_{1,t}\ x_{2,t}\ x_{3,t}\ x_{4,t}\ x_{5,t}]\}$, for $t \in \{1 \ldots N\}$
2) Generate samples of $f_1(\mathbf{x})$ under the factorization of H_1 by
   a) Draw $a, b, c \in \{1 \ldots N\}$
   b) Evaluate $f_1(\cdot)$ at $\mathbf{x} = [x_{1,a}\ x_{2,b}\ x_{3,b}\ x_{4,c}\ x_{5,c}]$
3) Generate samples of $f_2(\mathbf{x})$ under the factorization of H_2 by
   a) Draw $d, e \in \{1 \ldots N\}$
   b) Evaluate $f_2(\cdot)$ at $\mathbf{x} = [x_{1,d}\ x_{2,d}\ x_{3,e}\ x_{4,e}\ x_{5,e}]$
4) Similarly to steps $(2-3)$, obtain samples of $f_1(\mathbf{x}), f_2(\mathbf{x})$ under the intersection factorization H_0
5) Optimize the feature space portion of Equation (11) over the functions f_1, f_2:

$$D\big(p_{H_1}(f_1)\big\|p_{H_0}(f_1)\big) - D\big(p_{H_2}(f_2)\big\|p_{H_0}(f_2)\big)$$

using the above samples under each hypothesis; see Appendix I.

Fig. 6. Application of bootstrap sampling to estimate high-dimensional divergences via a low-dimensional feature f.

the two marginals and, thus, an estimate of (a lower bound on) the KL-divergence between our high-dimensional joint models.

Bootstrap sampling and permutation statistics are most commonly applied to finding confidence intervals for likelihood-based tests [9]; for example, a nonparametric test for independence based on this is proposed in [10]. Traditionally, these tests assume some prespecified statistic of the complete sample set $F(\{\mathbf{x}_t\})$ (where t indexes our N i.i.d. samples), use permutation statistics to determine the distribution of F under the null hypothesis, and construct a confidence interval or critical value against which to test F's observed value. In contrast to this, we use a statistic f of each joint observation $\mathbf{x}_t$ and employ randomization methods to draw samples that are forced to obey *both* of the assumed factorizations. This yields an estimate of the *distribution* of f under *both* hypotheses, enabling KL-divergence to be used as our distance metric for the test. It should also be possible to employ randomization statistics to estimate similar confidence intervals and critical values for our tests.

Another set of closely related methods are so-called *rank-like* tests [11], in which rank-based (and thus distribution-free) hypothesis tests are applied to some prespecified functional $f(\mathbf{x}_t)$. However, extensions of such rank-based tests beyond scalar dimension are nontrivial [12]. Additionally, using kernel density methods allows us to compute gradients with respect to the function f. Thus, we may easily *optimize* f to maximize this distance, removing the need to preselect a "good" statistic. Although not the subject of this paper, the methodology for optimizing the choice of f presented here may be extensible to rank-like tests as well.

B. Comparison with Optimization of MI

Mutual information is a common metric for learning and feature selection [13]–[16]. However, as previously noted, MI is sometimes insufficient to capture the kinds of interdependencies in, for example, a model selection problem involving many variables.

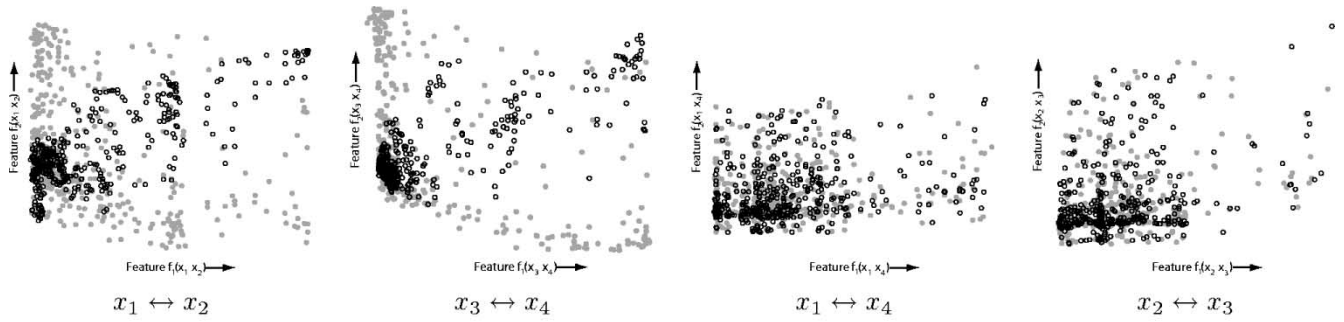
The estimation and optimization of KL-divergence presented here may be considered an improvement on Section III's mutual information-based method for a number of reasons. First, it is possible to estimate divergence over sets of many variables directly. This means that if desired, each sum of mutual information terms in (20) may be estimated as a whole rather than individually. Second, it does so using a single, possibly scalar, statistic. Estimates can thus be made in a lower dimensional space, which reduces the difficulty of density (or divergence) estimation. Finally, we note that the global maximum of (23) is always greater than or equal to the global maximum of (19) when the two are performed in equivalent dimensions, since the construction $f(x_{j_1}, x_{j_2}) = [f_j(x_{j_1}), g_j(x_{j_2})]$ achieves equality between the two.

Section IV introduces an application that will make use of some of these advantages. However, we first show the similar applicability of the two approaches by relearning statistics for the example problem of Section III-B using divergence-based estimates (see Fig. 7). To make the comparison to Fig. 5 more straightforward, we learn a 2-D feature for each pair of variables previously tested [as opposed to one-dimensional (1-D) features of each variable in the MI case]. In each plot, we show features of the two variables listed drawn from their observed joint distribution, drawn in black, and optimized with respect to the same feature sampled while enforcing the opposing hypothesis' factorization, drawn in gray (which in this case is equivalent to independence). As with the MI estimates, the pairings associated with the incorrect hypothesis (H_2) show only minor divergence from their distribution under a model based on the factorization in H_1, whereas the converse is not the case. The computed values tend to be slightly higher under the KL-optimized estimates, due to the relaxation of their statistics' forms; this is exhibited as a more tightly clustered joint distribution (black) in the figure.

Reiterating some possible benefits of using a KL-based approach, note that this same experiment could have been performed with only two learning operations (one accounting for *all* independence constraints of H_1 and optimizing with respect to H_0, the other doing the same for H_2). The feature could also be taken to be a scalar function, potentially reducing the amount of data required to adequately represent the distributions and, thus, reducing both data collection and computational costs.

V. Associating Image Sequences

To illustrate another high-dimensional data association problem, we consider the task of determining which of a set of cameras have overlapping fields of view. Such tasks are commonplace in video surveillance, for instance, in performing initial calibration (determining the camera locations and fields of view). The problem is also similar in nature to wide-baseline stereo correspondence tasks in computer vision [17]. In essence, it is similar to the association problem of Section III, except that every combination of sensors comprises a possible association.



(a) H_1 (black) vs. H_0 (gray), high divergence (b) H_2 (black) vs. H_0 (gray), low divergence

Fig. 7. Revisiting the pairwise association problem: Here, we have optimized a single (2-D) statistic for each divergence term in H_1, H_2 by comparison with its distribution under the intersection factorization. The estimated divergences are similar to the MI estimates of Fig. 5.

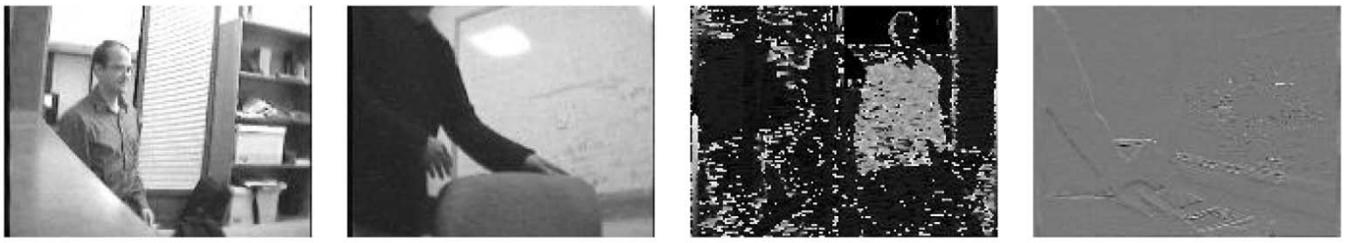


Fig. 8. Four cameras with different views of a room. To simulate multiple observation modalities, each image is transformed. In (a) and (b), we observe gray-scale image sequences, in (c) color (hue) information only, and in (d) instantaneous difference (edge) images.

We perform a test to determine the dependence between several loosely synchronized image sequences taken from a set of uncalibrated cameras that view the same region from substantially different locations. Each camera acquires an image only every few seconds (making tracking from frame to frame difficult), and as we possess relatively few data samples (several hundred frames), each of which is of high dimension (thousands of pixels), we are precluded from direct estimation and turn to information-preserving projections of the image data to estimate its interdependence. Note that for the purposes of this experiment, our conjecture is that statistical dependency, as measured by our approach, will be largely due to scene changes caused by the same object or objects.

Specifically, given image sequences $I_{1,t} \ldots I_{n,t}$ (where the first index represents the camera, and t indicates the time index), we test whether they are observations of independent scenes. Placed in the hypothesis testing framework discussed previously, this is

$$H_0: p(I_1, \ldots, I_n) = p(I_1) \cdot \ldots \cdot p(I_n) \tag{24}$$

$$H_1: p(I_1, \ldots, I_n) \neq p(I_1) \cdot \ldots \cdot p(I_n). \tag{25}$$

We can construct a test between the H_i by evaluating the divergence between the observed distribution $p(f(I_{1,t}, \ldots, I_{n,t}))$ and an independently resampled version, specifically $p(f(I_{1,\pi_1(t)}, \ldots, I_{n,\pi_n(t)}))$ for some permutations $\{\pi_j\}$.

The methodology proposed here makes no strong assumptions about the signal type, making it applicable for testing dependency across multimodal observations. To demonstrate this, we apply a postprocessing step to the images from two of the cameras as a proxy for different modalities. The observed values from the first two cameras are taken to be their gray-scale image intensities [Fig. 8(a) and (b)], whereas a third (c) retains only the color (hue) of its image. The fourth (d) observes instantaneous difference images, which we create by subtracting two images taken in quick succession. A notional camera geometry is shown in Fig. 9, where the camera pairs (a,c) and (b,d) have overlapping views. Note that cameras with overlapping fields of view observe *different* modalities.

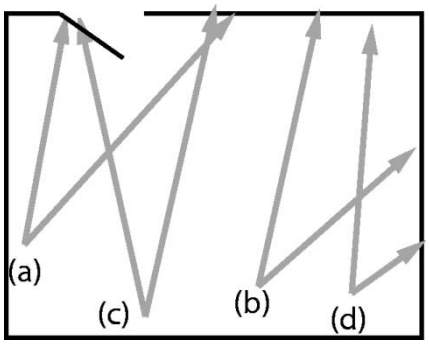


Fig. 9. Approximate locations and viewing angles of the four cameras described in Fig. 8.

We begin by examining only the pairwise association tests. Again, we do not address the combinatorial nature of the task of structure discovery (only the issue of high dimensionality). In the case of four cameras, there are six pairs to consider (as opposed to only four in Section III); therefore, enumeration of all possible pair-wise associations remains tractable. For each pair, we learn a single projection f_{ij}, which maximizes the KL-divergence between $f_{ij}(I_{i,t}, I_{j,t})$ and $f_{ij}(I_{i,\pi_i(t)}, I_{j,\pi_j(t)})$. The evaluated KL-divergences of these distributions are shown by the arrows in Fig. 10.

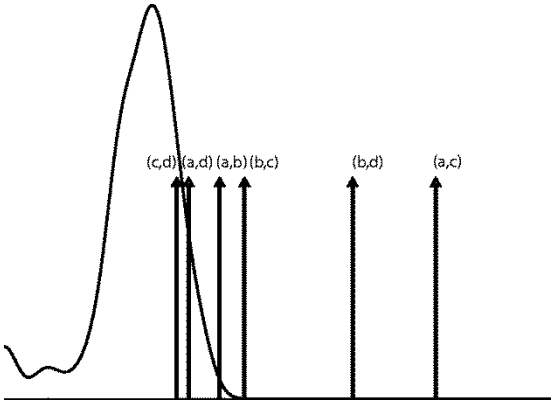


Fig. 10. KL-divergence from independence for the image sequences from each pair of cameras in Fig. 8. An estimate of the distribution of KL-divergence values under the null hypothesis (that the image sequences are mutually independent). The correct associations (a, c) and (b, d) are the rightmost two arrows, whereas the incorrect associations form the arrows to their left, indicating that the correct two pairs are the most likely to be dependent.

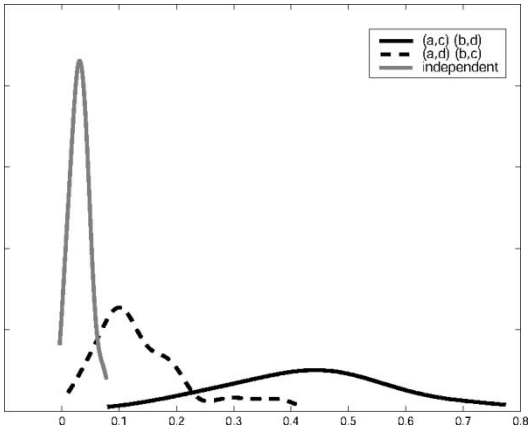


Fig. 11. Distributions of the proposed KL-divergence estimate over 100 Monte Carlo trials. An estimate of its distribution when the data are independent is given by the gray curve. The correct association (a, c) and (b, d) is shown by the solid black curve, whereas the incorrect association (a, d) and (b, c) is shown as dashed, indicating that the correct pairs are significantly more dependent. However, the differences observed from true independence (gray) indicate that there *is* some small but measurable coupling between nonoverlapping cameras.

However, we would *also* like to know how significant these values are. In order to test the statistical significance of the information content between each pair of image sequences, we also need to determine the distribution of our estimate under the null hypothesis (i.e., that the pair of image sequences are independent). Again, we employ randomization statistics. As discussed previously, we can easily generate data under the null hypothesis by permuting the time indices of both image sequences. We optimize a statistic to maximize the estimated KL-divergence between two data sets, *both* of which are generated under the null hypothesis. This yields a sample of the divergence estimate when the data is truly independent; many such samples can be used to estimate a distribution and enables us to evaluate the significance level of the observed divergence values.

The resulting comparison is shown in Fig. 10; the black curve indicates the null hypothesis distribution. As can be seen, the two rightmost arrows correspond to the correct associations between image sequences; these arrows have very little likelihood of coming from the null hypothesis.

Notably, the other arrows are all somewhat higher than expected under the null hypothesis as well, possibly indicating a slight interdependence even between nonoverlapping cameras. This coupling could result from second-order dependencies, such as joint lighting changes or the limited number of people moving in and between each scenes. In fact, the following similar experiment also indicates this dependence.

Let us next consider a somewhat more structured problem—to choose one (or neither) of two possible associations. Specifically, we assume (correctly) that the two gray-scale image sequences are *not* associated and test only between the three possibilities

$$\begin{aligned} H_0&: p(I_a, I_b, I_c, I_d) = p(I_a)p(I_b)p(I_c)p(I_d) \\ H_1&: p(I_a, I_b, I_c, I_d) = p(I_a, I_c)p(I_b, I_d) \\ H_2&: p(I_a, I_b, I_c, I_d) = p(I_a, I_d)p(I_b, I_c) \end{aligned} \tag{26}$$

where H_0 is the intersection set factorization of H_1 and H_2. We can do this by learning one statistic for each. To evaluate the variation in this procedure, we perform 100 Monte Carlo trials, using only half of the available images (chosen at random) for each trial. As in the previous example, we also perform the same procedure for data that have been permuted so as to obey H_0. The distributions of the (three) KL-divergence estimates so obtained are shown in Fig. 11. Notably, *neither* the divergence values assuming H_1 (solid black) *nor* the divergence values assuming H_2 (dashed) appear the same as the data assuming H_0 (gray). If the nonoverlapping camera pairs were truly independent, we would expect the data sampled under the incorrect factorization (H_2) to be the same as the data sampled under H_0. The fact that it is not reinforces the earlier observation that there exists some residual dependency between all four cameras. However, we correctly determine that the desired association (H_1) is significantly larger than its alternatives.

VI. Detecting Object Interaction

Another common association problem is that of determining which subsets of a group of objects move together and which move independently. For example, this application appears in computer vision for determining structure from motion [18], [19]. However, object motion is rarely characterized by independence between time samples, and thus we will require some mechanism to account for these dynamics and dependencies.

A. Incorporating Observation Dependency

One way to model temporal dynamics is via the conditional distribution $p(x_t|x_{t-1}, x_{t-2}, \ldots)$. Let us suppose that for each set of variables $S^i_{j,t}$ at time t, our process satisfies a Markov property—that for some T, we have $p\left(S^i_{j,t}|S^i_{j,t-1}, S^i_{j,t-2}, \ldots\right) = p\left(S^i_{j,t}|S^i_{j,t-1}, \ldots, S^i_{j,t-T}\right)$ and, additionally, that the process is conditionally stationary—that $p\left(S^i_{j,t}|S^i_{j,t-1}, \ldots\right)$ is the same for all t. For brevity, we present equations assuming $T = 1$ (i.e., first-order Markov), but it is straightforward to extend to the general case.

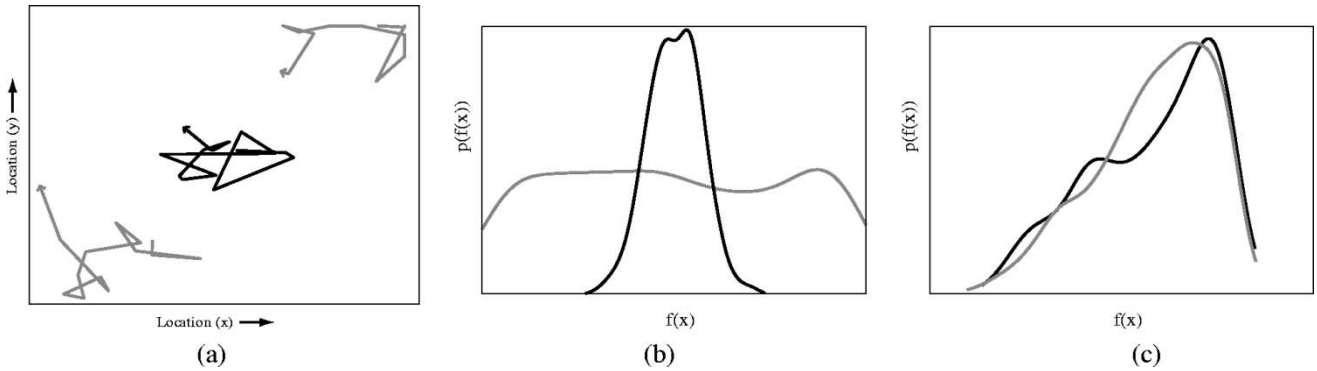


Fig. 12. (a) Two objects' motions are coupled by a third object that moves between them. We test for their interaction as given in (31) and (32); the density estimates of the resulting features are (b) H_1 (black curve) versus H_0 (gray), high divergence (indicating that the three objects move in a coupled fashion), and (c) H_2 (black) versus H_0 (gray), low divergence (indicating that, without x_3, x_1, and x_2 move independently).

Under these assumptions, the average log-likelihood ratio of (9) is instead a sum over each conditionally i.i.d. observation:

$$\frac{1}{N}\log L = \frac{1}{N}\sum_t \log \frac{\prod_j p_{H_1}\left(S^1_{j,t}|S^1_{j,t-1}\right)}{\prod_j p_{H_2}\left(S^2_{j,t}|S^2_{j,t-1}\right)} \tag{27}$$

and an identical analysis to that presented previously leads to a decomposition into statistical dependency and model divergence terms. If we again learn our models nonparametrically under different factorization assumptions, (11) becomes

$$D\left(\prod_j p\left(S^1_{j,t}|S^1_{j,t-1}\right)\middle\|\prod_{j,k} p\left(S^{12}_{jk,t}|S^{12}_{jk,t-1}\right)\right) - D\left(\prod_j p\left(S^2_{j,t}|S^2_{j,t-1}\right)\middle\|\prod_{j,k} p\left(S^{12}_{jk,t}|S^{12}_{jk,t-1}\right)\right) \tag{28}$$

and we can again estimate each of these terms separately via a lower bound and use their difference to evaluate a test between hypotheses.

This raises the question of how to apply the feature-based estimate of Section IV to determine the conditional divergences required in (28). It is easy to show that the conditional divergence is equivalent to a difference of two divergences between factorizations of all variables (see Appendix II-C). For example, the divergence term due to H_1 takes the form

$$\begin{aligned} &D\left(\prod_j p\left(S^1_{j,t}|S^1_{j,t-1}\right)\middle\|\prod_{j,k} p\left(S^{12}_{jk,t}|S^{12}_{jk,t-1}\right)\right) \\ &\quad = D\left(\prod_j p\left(S^1_{j,t}, S^1_{j,t-1}\right)\middle\|\prod_{j,k} p\left(S^{12}_{jk,t}, S^{12}_{jk,t-1}\right)\right) \\ &\qquad - D\left(\prod_j p\left(S^1_{j,t}\right) p\left(S^1_{j,t-1}\right)\middle\| \prod_j p\left(S^1_{j,t}\right)\prod_{j,k} p\left(S^{12}_{jk,t-1}\right)\right). \end{aligned} \tag{29}$$

Consequently, there are two KL-divergence terms, the arguments of which are in a form such that we can apply the sampling technique of Section IV-A. Specifically, for a given statistic f, we sample from the distribution of f under each of the four factorizations (one for each argument of the two divergence terms) on the right-hand side of (29), then estimate the difference of these two divergences. In order to maximize a bound on divergence, the function f is optimized (separately) for each term in the difference (29).

B. Example: Moving Objects

Here, we give an example of testing for dependency between sets of moving objects. In this problem, we compare sets of many variables, for which it would be difficult to estimate the mutual information in the manner of Section III. Thus, here, we apply (only) the KL-divergence estimation method described in Section IV. Fig. 12(a) shows two objects x_1, x_2 that move in independent, bounded random walks (gray paths), and a third (x_3) that attempts to interpose itself between them (black). Thus, the first two paths are coupled by the third, and the *correct* factorization is given by

$$\begin{aligned} &p(x_{1,t}, x_{2,t}, x_{3,t}|x_{1,t-1}, x_{2,t-1}, x_{3,t-1}) \\ &\quad = p(x_{3,t}|x_{3,t-1}x_{1,t}, x_{2,t})p(x_{1,t}|x_{1,t-1})p(x_{2,t}|x_{2,t-1}). \end{aligned} \tag{30}$$

We can test between possible factorizations of this distribution in a pairwise manner. It is not our goal to address the combinatoric nature of such a test; thus, we only compute values for two illustrative pairs of hypotheses: a full joint relationship to independent motions

$$\begin{aligned} H_1&: \hat{p}(x_{1,t}, x_{2,t}, x_{3,t}|x_{1,t-1}, x_{2,t-1}, x_{3,t-1}) \\ H_0&: \hat{p}(x_{1,t}|x_{1,t-1})\hat{p}(x_{2,t}|x_{2,t-1})\hat{p}(x_{3,t}|x_{3,t-1}) \end{aligned} \tag{31}$$

and a test between models that both assume x_3 to be independent

$$\begin{aligned} H_2&: \hat{p}(x_{1,t}, x_{2,t}|x_{1,t-1}, x_{2,t-1})\hat{p}(x_{3,t}|x_{3,t-1}) \\ H_0&: \hat{p}(x_{1,t}|x_{1,t-1})\hat{p}(x_{2,t}|x_{2,t-1})\hat{p}(x_{3,t}|x_{3,t-1}). \end{aligned} \tag{32}$$

The first test asks the question of whether there is *any* interaction between the three objects, whereas the second asks whether there is any direct interaction between x_1 and x_2. Furthermore, these two tests have a strong relationship to the true distribution, which we make precise shortly. In addiiton, note that each object position variable $x_{i,t}$ is 2-D, naively requiring H_1 (for example) to estimate a 12-D density. Using learned features, we may instead perform this test in a 1-D space.

The distributions of a scalar statistic maximizing the first likelihood ratio (31) is displayed in Fig. 12(b). As expected,

this demonstrates the large KL-divergence between these two models, i.e., that H_1 represents the data considerably better than H_0. However, the likelihood ratio between H_2, H_0 should be near-zero since x_1 and x_2 do, in fact, move independently of each other. Fig. 12(c) shows distributions for a statistic maximizing this divergence and that do, in fact, predict a small value.

The two example tests above are particularly illustrative in that they form the basis for several other tests of interest as well. For instance, the test between H_0 (full independence) and the true underlying distribution [which we denote H_T, (30)] can be expressed (see Appendix II-D) in terms of the computed divergence values. Specifically

$$D(p_{H_T}\|p_{H_0}) = D(p_{H_1}\|p_{H_0}) - D(p_{H_2}\|p_{H_0}) \tag{33}$$

which gives a large value, indicating that H_T is to be strongly preferred to H_0. Similarly, the test between H_1 (full joint) and H_T is given by

$$D(p_{H_1}\|p_{H_T}) = D(p_{H_2}\|p_{H_0}) \tag{34}$$

indicating that the two are nearly equivalent in a likelihood sense (note that H_1 will always have higher likelihood than any other hypothesis). Applying any reasonable penalty for the increased complexity of the full joint distribution, we may conclude that H_T is the preferred option.

VII. Conclusion

In the context of testing between alternative dependency structures for a set of random variables, we have cast the problem of determining this structure when the models are unknown or highly uncertain as a likelihood test between learned models with assumed factorizations and proposed the use of nonparametric density estimation techniques for evaluating this test. We then showed that the model-based likelihood ratio test may be decomposed into one set of terms measuring the statistical dependency structure supporting each hypothesis, and one set of terms that measure the fit of the models' parameterizations, allowing us to quantify the increased difficulty of tests in which the models must be learned directly from data.

We then addressed the difficulty of applying nonparametric methods to high-dimensional problems by proposing alternate estimates of the likelihood ratio based on the information-theoretic interpretation of its asymptotic limit. We showed how low-dimensional statistics of the data can be used to estimate lower bounds on mutual information and KL divergence and demonstrated that machine learning methods may be used to find statistics that optimize these bounds.

We have demonstrated the utility of this approach on three example problems. In the first, we showed each estimator's ability to perform association of data between multiple sensors, detecting harmonic relationships between two pairs of observed signals. The second example showed the ability of the proposed method to find which sets of image sequences have strong dependency (indicating overlap in their observed field) and, furthermore, to estimate the significance of this dependency. Finally, we applied nonparametric models to test between potential groupings of moving objects in order to determine whether a set of such objects moves coherently or independently.

Appendix I
Nonparametric Estimates of Entropy

There are a variety of nonparametric methods of estimating entropy (see [20] for an overview), several of which are based on kernel density estimates [7], [21]. Kernel methods model a distribution by assuming that the density is smooth around a set of observed samples; the kernel density estimate $\hat{p}(f)$, given a set of N i.i.d. samples $\{f(\mathbf{x}_i)\}$, is

$$\hat{p}(f) = \frac{1}{N}\sum_i K_\sigma(f - f(\mathbf{x}_i)) \tag{35}$$

where $K_\sigma(\mathbf{x})$ is a kernel function, and σ represents a smoothing parameter; we use the Gaussian density $K_\sigma(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, \sigma^2 I)$. There are a number of data-based methods for choosing the smoothing parameter σ; in practice, even simple methods such as the so-called *rule of thumb* [21] appear more than adequate.

The entropy estimate used in this work is the leave-one-out resubstitution estimate of [20]

$$\hat{H}(\{f(\mathbf{x}_i)\}) = -\frac{1}{N}\sum_i \log\left(\frac{1}{N-1}\sum_{j\neq i} K_\sigma(f(\mathbf{x}_j) - f(\mathbf{x}_i))\right) \tag{36}$$

where $K_\sigma(\mathbf{x})$ is again taken to be the the Gaussian distribution. It is then relatively straightforward to take the derivative of (36) with respect to any parameter α of the statistic $f(\cdot)$

$$\frac{\partial \hat{H}}{\partial \alpha} = -\frac{1}{N}\sum_i \frac{1}{\frac{1}{N-1}\sum_{j\neq i} K_\sigma(f(\mathbf{x}_j) - f(\mathbf{x}_i))} \cdot \left[\frac{1}{N-1}\sum_{j\neq i} K'_\sigma(f(\mathbf{x}_j) - f(\mathbf{x}_i))\,[f'(\mathbf{x}_j) - f'(\mathbf{x}_i)]\right] \tag{37}$$

where K' is the derivative of the kernel function (for the Gaussian kernel, $K'_\sigma(z) = -K_\sigma(z)(z/\sigma^2)$), and the statistic's derivative f' is with respect to the parameter α.

Additionally, the mutual information between two statistics $f(\mathbf{x})$, $g(\mathbf{y})$ may be estimated by

$$\hat{I}(f(\mathbf{x}); g(\mathbf{y})) = \hat{H}(f(\mathbf{x})) + \hat{H}(g(\mathbf{y})) - \hat{H}([f(\mathbf{x})g(\mathbf{y})]) \tag{38}$$

and its derivative is straightforward to compute by repeated application of (37).

We may use a resubstitution estimate similar to (36) for the KL-divergence given two sets of samples $\{f(\mathbf{x}_i)\}$, $\{f(\mathbf{y}_j)\}$:

$$\hat{D}(\{f(\mathbf{x}_i)\}\|\{f(\mathbf{y}_i)\}) = \frac{1}{N}\sum_i \log\left(\frac{1}{N-1}\sum_{j\neq i} K_\sigma\left(f(\mathbf{x}_j) - f(\mathbf{x}_i)\right)\right) - \frac{1}{N}\sum_i \log\left(\frac{1}{N}\sum_j K_\sigma\left(f(\mathbf{y}_j) - f(\mathbf{x}_i)\right)\right) \tag{39}$$

and take its derivative with respect to f in a similar fashion to (37), yielding a gradient-based learning rule for maximizing KL-divergence (or MI as a special case).

In the empirical sections of this paper, we have applied kernel-based methods, but any consistent density estimate whose gradient may be taken with respect to the function f

may be used. Extending the learning algorithm to alternate entropy estimates is a subject of ongoing research.

APPENDIX II
MISCELLANEOUS DERIVATIONS

Here, we present proofs and derivations that have been omitted from the main text.

A. Derivation of (6)

From (5), we have

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\log L &= \int\prod_l p_{H_1}\left(S_l^1\right)\log\frac{\prod_j p_{H_1}\left(S_j^1\right)}{\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)}\\ &\quad\cdot\frac{\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)}{\prod_j p_{H_2}\left(S_j^2\right)}\\ &= D\left(\prod_l p_{H_1}\left(S_l^1\right)\middle\|\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)\right)\\ &\quad+\int\prod_l p_{H_1}\left(S_l^1\right)\log\frac{\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)}{\prod_j p_{H_2}\left(S_j^2\right)}.\end{aligned}\tag{40}$$

The latter term can be rewritten as

$$\begin{aligned}&\int\prod_l p_{H_1}\left(S_l^1\right)\log\frac{\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)}{\prod_k p_{H_2}\left(S_j^2\right)}\\ &\quad=\sum_{j,k}\int\prod_l p_{H_1}\left(S_l^1\right)\log p_{H_1}\left(S_{jk}^{12}\right)\\ &\qquad-\sum_k\int\prod_l p_{H_1}\left(S_l^1\right)\log p_{H_2}\left(S_k^2\right)\end{aligned}$$

and, marginalizing over all variables which do not appear in the integral, we have

$$\begin{aligned}&=\sum_{j,k}\int\prod_l p_{H_1}\left(S_l^1\cap S_{jk}^{12}\right)\log p_{H_1}\left(S_{jk}^{12}\right)\\ &\quad-\sum_k\int\prod_l p_{H_1}\left(S_l^1\cap S_k^2\right)\log p_{H_2}\left(S_k^2\right)\\ &=\sum_{j,k}\int p_{H_1}\left(S_{jk}^{12}\right)\log p_{H_1}\left(S_{jk}^{12}\right)\\ &\quad-\sum_k\int\prod_l p_{H_1}\left(S_{lk}^{12}\right)\log p_{H_2}\left(S_k^2\right)\\ &=D\left(\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)\middle\|\prod_j p_{H_2}\left(S_k^2\right)\right)\end{aligned}$$

yielding (6):

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\log L &= D\left(\prod_j p_{H_1}\left(S_j^1\right)\middle\|\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)\right)\\ &\quad+D\left(\prod_{j,k}p_{H_1}\left(S_{jk}^{12}\right)\middle\|\prod_j p_{H_2}\left(S_j^2\right)\right).\end{aligned}$$

Thus, the divergence between two hypotheses H_1 and H_2 may be decomposed into one term corresponding to factorization differences between S^1 and the intersection sets S^{12}, and one term accounting for differences between the distribution of the intersection sets under H_1 and the distribution under H_2.

B. Derivation of (20)

Each of the divergences in (11) contributes one divergence for each associated pair of variables j:

$$D\left(p(x_{j_1},x_{j_2})\|p(x_{j_1})p(x_{j_2})\right).\tag{41}$$

To each term, we repeatedly apply the identity

$$\begin{aligned}&D\left(p(x_1,x_2)\|p(x_1)p(x_2)\right)\\ &\quad=\int p(x_1,x_2)\log\frac{p(x_1,x_2)}{p(x_1)p(x_2)}\\ &\quad=\int p(x_1,x_2)\log\frac{p(x_1,x_2)}{p(x_1,f(x_1))p(x_2,g(x_2))}\\ &\quad=\int p(x_1,x_2)\\ &\qquad\cdot\log\left[\frac{p(f(x_1),g(x_2))}{p(f(x_1))p(g(x_2))}\right.\\ &\qquad\qquad\left.\cdot\frac{p(x_1,x_2)}{p(f(x_1),g(x_2))p(x_1|f(x_1))p(x_2|g(x_2))}\right]\\ &\quad=I\left(f(x_1);g(x_2)\right)\\ &\qquad+D\left(p(x_1,x_2)\|p(f(x_1),g(x_2))\right.\\ &\qquad\qquad\left.\cdot p(x_1|f(x_1))p(x_2|g(x_2))\right)\end{aligned}$$

which match the mutual information and divergence terms summed in (20).

C. Derivation of (29)

Due to the similar form of both conditional divergences in (28), we simply show the identity stated in the text, that the divergence due to H_1's departure from the null hypothesis' factorization is given by the equation at the top of the next page, yielding the right-hand side of (29).

D. Derivation of (33) and (34)

For simplicity, here we leave as implied the dependence on past measurements. Recall that the correct distribution is $p_{H_T}=p(x_3|x_1,x_2)p(x_1)p(x_2)$ and that $p_{H_0}=p(x_1)p(x_2)p(x_3)$, $p_{H_1}=p(x_1,x_2,x_3)$, and $p_{H_2}=p(x_1,x_2)p(x_3)$. We then have

$$\begin{aligned}D(p_{H_T}||p_{H_0}) &= E\left[\log\frac{p(x_3|x_1,x_2)p(x_1)p(x_2)}{p(x_1)p(x_2)p(x_3)}\right]\\ &= E\left[\log\frac{p(x_3|x_1,x_2)}{p(x_3)}\right]\\ &= E\left[\log\frac{p(x_3|x_1,x_2)p(x_1,x_2)}{p(x_1,x_2)p(x_3)}\right]\\ &= E\left[\log\frac{p(x_1,x_2,x_3)}{p(x_1,x_2)p(x_3)}\right]\\ &= E\left[\log\frac{p(x_1,x_2,x_3)}{p(x_1)p(x_2)p(x_3)}\cdot\frac{p(x_1)p(x_2)p(x_3)}{p(x_1,x_2)p(x_3)}\right]\\ &= D(p_{H_1}||p_{H_0})-D(p_{H_2}||p_{H_0})\end{aligned}$$

$$
\begin{aligned}
&D\left(\prod_j p\left(S^1_{j,t}|S^1_{j,t-1}\right)\middle\|\prod_{j,k} p\left(S^{12}_{jk,t}|S^{12}_{jk,t-1}\right)\right)\\
&\quad=\int\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)\log\left[\frac{\prod_j p\left(S^1_{j,t}|S^1_{j,t-1}\right)}{\prod_{j,k} p\left(S^{12}_{jk,t}|S^{12}_{jk,t-1}\right)}\right]\\
&\quad=\int\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)\log\left[\frac{\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)}{\prod_j p\left(S^1_{j,t-1}\right)}\frac{\prod_{j,k} p\left(S^{12}_{jk,t-1}\right)}{\prod_{j,k} p\left(S^{12}_{jk,t},S^{12}_{jk,t-1}\right)}\right]\\
&\quad=\int\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)\log\left[\frac{\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)}{\prod_j p\left(S^1_{j,t}\right)\prod_j p\left(S^1_{j,t-1}\right)}\frac{\prod_j p\left(S^1_{j,t}\right)\prod_{j,k} p\left(S^{12}_{jk,t-1}\right)}{\prod_{j,k} p\left(S^{12}_{jk,t},S^{12}_{jk,t-1}\right)}\right]\\
&\quad=\int\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)\log\left[\frac{\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)}{\prod_{j,k} p\left(S^{12}_{jk,t},S^{12}_{jk,t-1}\right)}\frac{\prod_j p\left(S^1_{j,t}\right)\prod_{j,k} p\left(S^{12}_{jk,t-1}\right)}{\prod_j p\left(S^1_{j,t}\right)\prod_j p\left(S^1_{j,t-1}\right)}\right]\\
&\quad=D\left(\prod_j p\left(S^1_{j,t},S^1_{j,t-1}\right)\middle\|\prod_{j,k} p\left(S^{12}_{jk,t},S^{12}_{jk,t-1}\right)\right)\\
&\qquad-D\left(\prod_j p\left(S^1_{j,t}\right)\prod_j p\left(S^1_{j,t-1}\right)\middle\|\prod_j p\left(S^1_{j,t}\right)\prod_{j,k} p\left(S^{12}_{jk,t-1}\right)\right)
\end{aligned}
$$

which is (33); similarly

$$
\begin{aligned}
D(p_{H_1}\|p_{H_T}) &= E\left[\log\frac{p(x_1,x_2,x_3)}{p(x_3|x_1,x_2)p(x_1)p(x_2)}\right]\\
&= E\left[\log\frac{p(x_1,x_2,x_3)}{p(x_1)p(x_2)p(x_3)}\right.\\
&\qquad\left.\cdot\frac{p(x_1)p(x_2)p(x_3)}{p(x_3|x_1,x_2)p(x_1)p(x_2)}\right]
\end{aligned}
$$

which, by an argument similar to that of Appendix II-A

$$
\begin{aligned}
&= D(p_{H_1}\|p_{H_0}) - D(p_{H_T}\|p_{H_0})\\
&= D(p_{H_2}\|p_{H_0})
\end{aligned}
$$

giving (34).

REFERENCES

[1] A. T. Ihler, J. W. Fisher III, and A. S. Willsky, "Hypothesis testing over factorizations for data association," in *Proc. IPSN*, Apr. 2003.

[2] J. W. Fisher III and T. Darrell, "Probabilistic models and informative subspaces for audiovisual correspondence," in *Proc. ECCV*, vol. III, June 2002, pp. 592–603.

[3] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[4] P. J. Bickel and K. A. Doksum, *Mathematical Statistics:Basic Ideas and Selected Topics*. San Francisco: Holden-Day, 1977.

[5] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[6] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Stat.*, vol. 33, pp. 1065–1076, 1962.

[8] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*. New York: Wiley, 1973.

[9] P. Hall, "On the bootstrap and confidence intervals," *Ann. Stat.*, vol. 14, no. 4, pp. 1431–1452, Dec. 1986.

[10] J. P. Romano, "Bootstrap and randomization tests of some nonparametric hypotheses," *Ann. Stat.*, vol. 17, no. 1, pp. 141–159, Mar. 1989.

[11] R. H. Randles and D. A. Wolfe, *Introduction to the Theory of Nonparametric Statistics*. New York: Wiley, 1979.

[12] G. Fasano and A. Franceschini, "A multidimensional version of the Kolmogorov-Smirnov test," *Monthly Notices Roy. Astron. Soc.*, vol. 225, pp. 155–170, 1987.

[13] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[14] P. Viola and W. Wells III, "Alignment by maximization of mutual information," *IJCV*, vol. 24, no. 2, pp. 137–54, 1997.

[15] J. Fisher III and J. Principe, "A methodology for information theoretic feature extraction," in *IJC Neural Networks*, A. Stuberud, Ed., 1998.

[16] A. T. Ihler, J. W. Fisher III, and A. S. Willsky, "Nonparametric estimators for online signature authentication," in *Proc. ICASSP*, May 2001.

[17] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," in *Proc. ICCV*, 1998, pp. 754–760.

[18] Y. Song, L. Goncalves, and P. Perona, "Learning probabilistic structure for human motion detection," in *Proc. CVPR*, 2001, pp. 771–777.

[19] L. Taycher, J. W. Fisher III, and T. Darrell, "Recovering articulated model topology from observed rigid motion," in *Proc. NIPS*, 2002.

[20] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, June 1997.

[21] B. Silverman, *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall, 1986.

Alexander T. Ihler (S'01) received the B.S. degree from the California Institute of Technology, Pasadena, in 1998 and the S.M. degree from the Massachusetts Institute of Technology (MIT), Cambridge, in 2000. He is currently pursuing the doctoral degree at MIT with the Stochastic Systems Group.

His research interests are in statistical signal processing, machine learning, nonparametric statistics, distributed systems, and sensor networks.

John W. Fisher (M'98) received the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 1997.

He is currently a Principal Research Scientist with the Computer Science and Artificial Intelligence Laboratory and affiliated with the Laboratory for Information and Decision Systems, both at the Massachusetts Institute of Technology (MIT), Cambridge. Prior to joining MIT, he was with the University of Florida, as both a faculty member and graduate student since 1987, during which time he conducted research in the areas of ultra-wideband radar for ground penetration and foliage penetration applications, radar signal processing, and automatic target recognition algorithms. His current area of research focus includes information theoretic approaches to signal processing, multimodal data fusion, machine learning, and computer vision.

Alan S. Willsky (S'70–M'73–SM'82–F'86) joined the faculty of the Massachusetts Institute of Technology (MIT) in 1973 and is currently the Edwin Sibley Webster Professor of Electrical Engineering. He is a founder, member of the Board of Directors, and Chief Scientific Consultant of Alphatech, Inc. From 1998 to 2002 he served as a member of the US Air Force Scientific Advisory Board. He has held visiting positions in England and France. He has delivered numerous keynote addresses and is co-author of the undergraduate text *Signals and Systems* (Englewood Cliffs, NJ: Prentice-Hall, 1996, Second ed.). His research interests are in the development and application of advanced methods of estimation and statistical signal and image processing. Methods he has developed have been successfully applied in a variety of applications including failure detection, surveillance systems, biomedical signal and image processing, and remote sensing.

Dr. Willsky has received several awards, including the 1975 American Automatic Control Council Donald P. Eckman Award, the 1979 ASCE Alfred Noble Prize, and the 1980 IEEE Browder J. Thompson Memorial Award. He has held various leadership positions in the IEEE Control Systems Society (which made him a Distinguished Member in 1988).