ESTIMATING DEPENDENCY AND SIGNIFICANCE FOR HIGH-DIMENSIONAL DATA

Michael R. Siracusa[†] *Kinh Tieu*[†] *Alexander T. Ihler*[*] *John W. Fisher*[†,*] *Alan S. Willsky*[*,†]

[†]Computer Science and Artificial Intelligence Laboratory [*]Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
32 Vassar St., Cambridge MA 02139

ABSTRACT

Understanding the dependency structure of a set of variables is a key component in various signal processing applications which involve data association. The simple task of detecting whether any dependency exists is particularly difficult when models of the data are unknown or difficult to characterize because of high-dimensional measurements. We review the use of nonparametric tests for characterizing dependency and how to carry out these tests with high-dimensional observations. In addition we present a method to assess the significance of the tests.

1. INTRODUCTION

Determining the presence and structure of dependency between a set of variables is an important task in many sensor network, image processing, and computer vision applications. For example, the problem of data association arises frequently in sensor networks. When multiple sensors and sources are present, it may be necessary to determine which observations from different sensors correspond to the same target. This is particularly challenging when a prior model is either unavailable or only weakly specified. Such issues naturally arise when the goal is to estimate the dependency structure between sets of high-dimensional measurements; for example, when analyzing video streams for a multi-camera sensor network. When applicable, a joint learning/estimation approach is appealing, but the online case requires learning the signal distributions while simultaneously performing a test for dependency.

By formulating the online problem as a hypothesis test between different factorizations of the variables and treating the distributions involved as nuisance parameters, one finds that the relevant quantities can be interpreted as estimates of mutual information (MI) and more generally, Kullback-Leibler (KL) divergence. Direct estimation of these quantities is typically infeasible for high-dimensional data. In such cases machine learning techniques and bootstrap or permutation statistics can be used to find information-preserving subspaces and estimate lower bounds on MI and KL divergence. These lower bounds can be used in place of the optimal likelihood ratio as a measure of statistical dependency.

It is also desirable to have not only have an estimate of the significance of any such information measure. This is difficult without a prior model. However, in similar fashion bootstrap and permutation tests may utilized to assess the level of confidence. Such tests are complicated by the presence of the optimization (learning) step in estimation. We describe how one may still obtain an accurate significance level.

The use of information measures and nonparametric statistics in hypothesis testing and machine learning have a long history. Information theoretic measures of dependence for parametric distributions are described in Kullback [1]. Patrick and Fischer[2] represent an early example of using kernel density estimators to project high-dimensional measurements to low-dimensional representations, albeit in a different context. Our approach combines the flexibility of nonparametric estimates with information-preserving subspaces in a dependency scenario. Bootstrap sampling and permutation statistics have been used to obtain confidence intervals [3] and to test for independence [4]. However in those cases, the subspace was prespecified instead of being learned.

2. FACTORIZATION TESTS

A class of hypothesis tests referred to as factorization tests were introduced in [5]. The goal of such tests is to choose among dependency hypotheses in the absence of a parameterized model. Each hypothesis describes a particular graphical representation over the set of variable. Partitions of the variables into disjoint mutually independent subsets is a special case.

In this paper we focus on testing between fully dependent and fully independent factorizations which leads to a straightforward definition of significance for our factorization test where the null hypothesis (H_0) is that the variables are independent. Given N independent and identically distributed observations $\mathbf{x} = \{\mathbf{x}_n\}$ for $n \in \{1...N\}$ each of which is the output of K sensors such that $\mathbf{x}_n = \left[\mathbf{x}_n^{1\,\mathsf{T}}\ \mathbf{x}_n^{2\,\mathsf{T}}\ \dots\ \mathbf{x}_n^{K\,\mathsf{T}}\right]^{\mathsf{T}}$ the (normalized) likelihood ratio for distinguishing between H_1 (dependent) and H_0 (independent) is:

$$\begin{aligned} L(\{\mathbf{x}_n\}) &= \frac{1}{N}\sum_n \log \frac{p(\mathbf{x}_n|H_1)}{p(\mathbf{x}_n|H_0)} \\ &= \frac{1}{N}\sum_n \log \frac{p(\mathbf{x}_n^1, \mathbf{x}_n^2, ..., \mathbf{x}_n^K|H_1)}{\prod_k p(\mathbf{x}_n^k|H_0)} \end{aligned} \tag{1}$$

which has the following asymptotic behavior as $N \to \infty$

$$L(\{\mathbf{x}_t\}) = \begin{cases} D\Big(p(\mathbf{x}|H_1)\Big|\Big|p(\mathbf{x}|H_0)\Big) & ; \quad H_1 \text{ is true} \\ -D\Big(p(\mathbf{x}|H_0)\Big|\Big|p(\mathbf{x}|H_1)\Big) & ; \quad H_0 \text{ is true} \end{cases} \tag{2}$$

That is, in the limit the optimal likelihood ratio test converges to KL divergence. When we have the model for our data under both hypotheses we simply plug our data into Equation 1 to calculate the likelihood ratio and choose H_1 if its value is positive or H_0 otherwise (assuming equal priors).

2.1. Nonparametric Online Factorization Test

If we replace each density in Equation 1 with densities, $\hat{p}()$, estimated from training data using consistent density estimators, then we have a nonparametric factorization test that, in the limit, converges to the performance of the optimal test. However, a more interesting case is when we estimate the models and perform the hypothesis test from the same sample draw. In this case we have access to only one dataset, which necessarily comes from either H_1 or H_0. This makes our test considerably more difficult.

If our observed data comes from H_1 and we estimate densities from this data, our estimates $\hat{p}_{H_i}()$ under the factorization consistent with hypothesis H_i converge to:

$$\begin{aligned} \hat{p}_{H_1}(\mathbf{x}) &\rightarrow p(\mathbf{x}|H_1) \\ \hat{p}_{H_0}(\mathbf{x}) &\rightarrow \prod_k p(\mathbf{x}^k|H_1) \end{aligned} \tag{3}$$

conversely if observed data comes from H_0:

$$\begin{aligned} \hat{p}_{H_1}(\mathbf{x}) &\rightarrow p(\mathbf{x}|H_0) \\ \hat{p}_{H_0}(\mathbf{x}) &\rightarrow p(\mathbf{x}|H_0) \end{aligned} \tag{4}$$

Plugging in these estimates into the likelihood ratio gives an estimate that in the limit under H_1 becomes

$$\hat{L}(\{\mathbf{x}_t\}) = \frac{1}{N}\sum_n \log \frac{\hat{p}_{H_1}(\mathbf{x}_n)}{\hat{p}_{H_0}(\mathbf{x}_n)} \tag{5}$$

$$\approx \begin{cases} D\Big(p(\mathbf{x}|H_1)\big|\big|\prod_k p(\mathbf{x}^k|H_1)\Big) & ; H_1 \text{ true} \\ 0 & ; H_0 \text{ true} \end{cases} \tag{6}$$

We see that by performing an online test which estimates distributions from the sample under test yields some loss in separation between the two hypotheses and that the test becomes an estimate of the KL divergence in Equation 6. This loss is similar to the issues that arise in generalized likelihood ratio (GLR) tests [6]. In the special case of two variables ($\mathbf{x}^1$ and $\mathbf{x}^2$) this quantity is equivalent to measuring mutual information (MI).

3. FEATURE-BASED DIVERGENCE ESTIMATION

Direct estimation of the divergence term in Equation 6 is hampered by the infeasibility of high-dimensional density estimation. However, for any deterministic function $T(\mathbf{x})$, the lower bound lower bound on the divergence of H_1 and H_0 is easily shown using the data processing inequality [7]:

$$D\Big(p(T(\mathbf{x})|H_1)\big|\big|p(T(\mathbf{x})|H_0)\Big) \leq D\Big(p(\mathbf{x}|H_1)\big|\big|p(\mathbf{x}|H_0)\Big) \tag{7}$$

The approach adopted here is to optimize the feature/function $T(\mathbf{x})$ so as to maximize the left side of Equation 7. Nonparametric density estimation over samples of $T(\mathbf{x})$ is feasible provided the dimensionality of the mapping is chosen to be low enough.

3.1. Sufficiency

To help gain some intuition about the optimization, suppose that the dependency in our observations is explained (at least approximately) by some lower dimensional latent variable $\boldsymbol{\theta}$. The corresponding generative model is a Markov chain: $H \rightarrow \boldsymbol{\theta} \rightarrow \mathbf{x}$. For a particular hypothesis H we generate a low dimensional variable $\boldsymbol{\theta} = [\theta^1\ \theta^2\ \ldots\ \theta^k]^\mathsf{T}$ that captures the dependency between a set of k sensors. These low dimensional variables θ^k generate a high-dimensional observations $\mathbf{x}^k$.

The optimization to maximize the left side of Equation 7 finds a feature $T(\mathbf{x})$ which extends the chain to $H \rightarrow \boldsymbol{\theta} \rightarrow \mathbf{x} \rightarrow T(\mathbf{x})$. If this feature satisfies Equation 7 with equality, it is said to be *sufficient* for θ. That is $p(\theta|\mathbf{x}) = p(\theta|T(\mathbf{x}))$ and we have the equivalent Markov chain $H \rightarrow \boldsymbol{\theta} \rightarrow T(\mathbf{x}) \rightarrow \mathbf{x}$. Note that it can also be shown that if this feature is sufficient for $\boldsymbol{\theta}$ is also sufficient for H, that is $p(H|\mathbf{x}) = p(H|T(\mathbf{x}))$.

By way of example, if we assumed Gaussian distributions and restrict T to linear 1-d projections a closed form solution exists to maixmize the left side of Equation 7 [1]. This solution extends to m-d projections by iteratively finding optimal 1-d projections that are orthogonal to the previous ones. Alternatively with kernel based methods for density estimation we can use a simple gradient ascent method over parameterized functions T. In this paper we restrict T to be linear subject to an L_1 regularization penalty.

3.2. Obtaining Samples from H_0

Notice that our optimization requires estimates of the distribution of the feature $T(\mathbf{x})$, under both H_1 and H_0. However, during our online test we only receive a single set of observations $\mathbf{x} = \{\mathbf{x}_n\}$ that were drawn under a single hypothesis. That is, they are either dependent (under H_1) or independent (under H_0). Fortunately, through the use of bootstrap sampling, we can always obtain samples of $T(\mathbf{x})$ under H_0. For example, given a set of observations $\mathbf{x} = \{\mathbf{x}_n\}$ for $n \in \{1...N\}$ and 3 sensors ($K = 3$) we can generate a sample of $T(\mathbf{x})$ under H_0 by:

1. Draw $\pi_0 = a, b, c \in \{1...N\}$.
2. Evaluate $T(\mathbf{x})$ at $\mathbf{x} = \pi_0(\{\mathbf{x_n}\}) = [\mathbf{x_a^1}^\mathsf{T}\mathbf{x_b^1}^\mathsf{T}\mathbf{x_c^1}^\mathsf{T}]^\mathsf{T}$.

Bootstrap sampling draws these samples with replacement and can be used to estimate $\hat{p}(T(\mathbf{x})|H_0)$. Alternatively by sampling without replacement we can obtain an estimate of $\hat{p}(T(\mathbf{x})|H_0)$ related to permutation statistics. Note that estimating $\hat{p}(T(\mathbf{x})|H_1)$ is done by directly using the observed samples $\{\mathbf{x}_n\}$. That is, there is no need for bootstrap sampling to obtain our density estimate under H_1. If we are given dependent data the direct estimate of $\hat{p}(T(\mathbf{x})|H_1)$ will differ from the bootstrap estimate of $\hat{p}(T(\mathbf{x})|H_0)$ giving us a positive KL value. However if the data we are given is truly independent then both the direct and the bootstrap distribution estimates should yield KL value of zero.

So far we have shown that measuring KL divergence is the key component in our factorization tests and that we can avoid the difficulty associated with estimating high dimensional distributions by maximizing a lower bound on KL via a low dimensional feature. Calculating this lower bound requires estimates of both $\hat{p}(T(\mathbf{x})|H_1)$ and $\hat{p}(T(\mathbf{x})|H_0)$. We directly estimate $\hat{p}(T(\mathbf{x})|H_1)$ from the observed data, but use bootstrap sampling techniques to obtain samples $T(\mathbf{x})$ under the independent hypothesis. These samples are then used to estimate $\hat{p}(T(\mathbf{x})|H_0)$.

4. MEASURING SIGNIFICANCE

We have seen that the basic factorization test for dependence involves taking the observation $\{\mathbf{x}_n\}$ and estimating the likelihood ratio $L(\{\mathbf{x}_n\})$. If $L(\{\mathbf{x}_n\})$ is the optimal test where the distributions are given we know to choose H_1 when $L(\{\mathbf{x}_n\})$ is positive

and H_0 otherwise. However, here we do not know the distribution under H_1. In such cases it is common to estimate a p-value or the probability of observing a value greater than $\hat{L}(\{\mathbf{x}_n\})$ under H_0. To do so we need the distribution of $\hat{L}(\{\mathbf{x}_n\})$ under H_0. We can obtain this through bootstrap sampling:

1. Draw a set $\mathbf{x}_{0,n} = \{\pi_0(\{\mathbf{x}_n\})\}$.
2. Apply factorization test to obtain $\hat{L}(\{\mathbf{x}_{0,n}\})$.
3. Repeat N_{null} times to estimate $p(\hat{L}(\{\mathbf{x}_n\})|H_0)$.
4. Measure significance as $\int_{\hat{L}(\{\mathbf{x}_n\})}^{\infty} p(\hat{L}(\{\mathbf{x}_n\})|H_0)dL$.

with the last step simply being a count of the number of times $\hat{L}(\{\mathbf{x}_{0,n}\})$ was greater than $L(\{\mathbf{x}_n\})$ divided by N_{null}. This bootstrap sampling procedure is separate from the one described previously in Section 3.2. In that section we used permutations to obtain samples of $T(\mathbf{x})$ under H_0 in order to evaluate a lower bound on KL. Here we use permutations to obtain samples of $\hat{L}(\{\mathbf{x}_n\})$ under H_0 to assess significance.

It is important to note that using the feature based divergence estimate described in Section 3 involves an optimization in step 2 in the procedure above. This is a key difference between our procedure and traditional approaches that use bootstrap sampling or permutation statistics to assess significance for likelihood based tests that use a pre-specified statistic/function of the data [8]. We now can calculate a measure of dependency for high-dimensional data and asses the significance of our measurement.

5. EXPERIMENTS

We experiment with a simple model for generating high dimensional data. The model is chosen specifically with regard to properties as a function dimensionality. We start by defining a low dimensional distribution $p(\boldsymbol{\theta}|H_1)$ that defines the dependency structure of a K dimensional variable $\boldsymbol{\theta}$. This yields a $p(\boldsymbol{\theta}|H_0)$ as the product of the K marginals. These distributions can have any form and will capture all the information we have about dependency. We then generate K high dimensional observations from the following linear model:

$$\mathbf{x}^k = \mathbf{a}^k\theta^k + \mathbf{b}^k z^k + \mathbf{n}^k \tag{8}$$

with

$$\mathbf{a}^k = \alpha \begin{bmatrix} \mathbf{1}_M \\ \mathbf{0}_{D/3-M} \end{bmatrix} \tag{9}$$

This creates a $D/3 \gg K$ dimensional observation with information about $\boldsymbol{\theta}$ evenly distributed in an M dimensional subspace. We add distractors $\mathbf{b}^k z^k$ were $\mathbf{b}^k$ is orthogonal to a^k and z^k is Gaussian and uncorrelated with z^j for all $j \neq k$. Additionally, $\mathbf{n}^k$ adds Gaussian noise in the $D/3$ dimensional space.

With this model there exists a K dimensional sufficient statistic of the form $T(\mathbf{x}) = [T(\mathbf{x}^1)\ T(\mathbf{x}^2)\ \ldots\ T(\mathbf{x}^k)]^{\mathsf{T}}$ with

$$T(\mathbf{x}^k) = {\mathbf{a}^k}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \mathbf{x}^k \tag{10}$$

where $\boldsymbol{\Sigma}_{\mathbf{n}}^{-1}$ is the inverse covariance of the noise $\mathbf{n}^k$. *Imporatantly* it can also be shown that by setting $\alpha = M^{-\frac{1}{2}}$ in Equation 9 the posterior $p(\theta|T(\mathbf{x}))$ is invariant to both M and D as well as to the marginal distribution $p(\boldsymbol{\theta})$ (excepting some degenerate cases).

Thus the low dimensional variable $\boldsymbol{\theta}$ describes the dependency across measurements. This information is distributed into a higher dimensional space via Equation 8. We can analytically calculate the sufficient statistic $T(\mathbf{x})$ for this model. We can control the dimensionality of our observations and how this information is distributed by setting D and M without affecting the posterior $p(\theta|T(\mathbf{x}))$. Our high dimensional observations contain distractors $\mathbf{b}^k z^k$ add high variance irrelevant information and present problems for generic dimensionality reduction techniques such as principle component analysis (PCA) or independent component analysis (ICA).

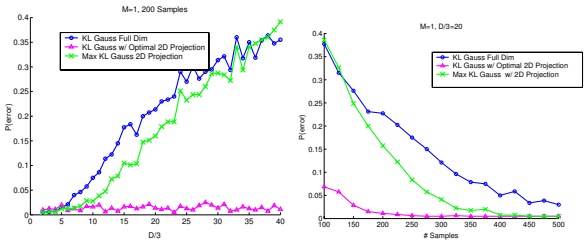

Fig. 1. Trends with Gaussian data. The y axis of each plot is $p(error)$. From left to right : varying D (fixed M, and N), varying N (fixed M and D).

5.1. Gaussian Data

We begin with a simple case where $p(\theta|H_1)$ is a 2-d correlated Gaussian ($\rho = .75$). Figure 1 compares the performance of three different dependency tests for various settings of M and D in Equation 8 and the number of samples used N. The tests were: 1) KL directly from the data, 2) estimating KL using the sufficient statistic projection $T(\mathbf{x})$, and 3) optimizing for the projection that maximizes KL. All three techniques assumed a Gaussian model. Multiple trials were performed with both dependent and independent data. For each trial and each method $L(\mathbf{x})$ was calculated as well as its significance. Using the results for multiple trials we found the threshold on significance that minimized the probability of error when choosing a hypothesis and reordered this $p(error)$.

We see clear trends in Figure 1. First, as explained, varying M or D does not affect the method using the sufficient statistic (a baseline - as we do not know the statistic for the general case). We note that the performance from best to worst is knowing the sufficient statistic, optimizing for maximum KL, and finally estimating in the full high-dimensional space.

The performance gap between direct estimation and optimization changes as function of D and N. For small D all the methods do equally well. For large D both the optimization and direct estimation perform poorly. For small number of samples N and large N both techniques have similar performance, while in between we see a significant advantage with optimization method. It also is interesting to note that varying M has little affect on performance for a fixed D.

These experiments illustrate the value of estimating a lower-dimensional statistic, even if only approximate, versus direct estimation in the high dimensional space.

5.2. 3D Ball example

Next we study the case where $p(\theta|H_1)$ to a mixture of four Gaussian:

$$p(\boldsymbol{\theta}) = \frac{1}{J}\sum_{j=1}^{J} \mathcal{N}(\boldsymbol{\theta}; \mu_j, \sigma_{\boldsymbol{\theta}}^2 \mathbf{I}_3) \tag{11}$$

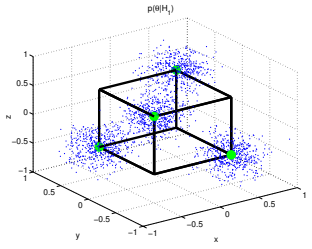

Fig. 2. 3D Ball Data with $\sigma_\theta^2 = .01$.

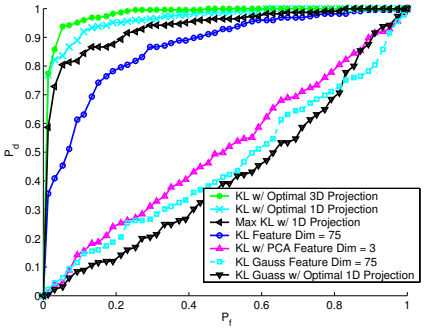

Fig. 3. ROC Performance Comparison. KL values computed using $T(\mathbf{x})$ give the best performance.

with $J = 4$ means, μ_j, $(.5, .5, .5)$, $(.5, -.5, -.5)$, $(-.5, .5, -.5)$, $(-.5, -.5, .5)$ respectively. A sampling of this data is shown in Figure 2. Note that in this example all the marginal and pairwise marginal densities are identical and have the same covariance. Consequently, Gaussian assumptions are not sufficient for detecting dependency. For our experiments we set $\sigma_\theta^2 = .15$ and set $\mathbf{\Sigma_n}^{-1}$ so that each of the non-distractor dimensions have equal variance. The distractor $\mathbf{z}$ is set so that has the highest variance. For these experiments $M = 12$ and $D = 75$.

Estimating densities in high dimensions is difficult and may lead to inaccurate KL values. A common approach is to apply standard dimensionality reduction techniques such as PCA. However in our synthetic case, choosing the top principal component will result in zero KL divergence because the maximum variance dimensions correspond to the distractor variables. In practice we are not given the projection corresponding to $T(\mathbf{x})$ and must find it by maximizing KL as previously described. The projections found are not necessarily identical to $T(\mathbf{x})$ but do lead to more accurate estimates of KL and hence give better performance than either PCA or working in the high dimensional observation space.

Figure 3 shows that the KL values computed using the projection corresponding to $T(\mathbf{x})$ lead to better performance than computing KL in high dimensions or in the maximum variance subspace in deciding between dependence and independence. We see that the performance of the maximizing KL optimization falls in between. We also see that all the techniques which made Gaussian assumptions or used PCA failed.

5.3. Significance using permutation samples

To test how well our permutation samples approximate independent samples, we compared significance values computed using

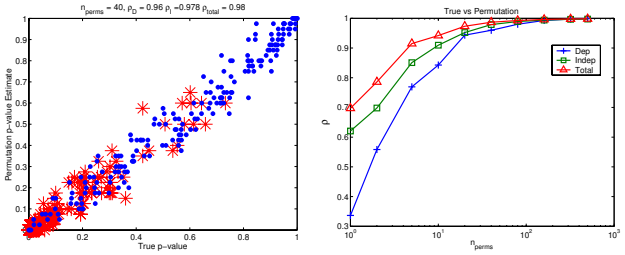

Fig. 4. (**Left**) Strong linear relationship between significance computed from independent and permuted data. (**Right**) Correlation increases with the number of permutations.

both types of samples. They are highly correlated as shown in Figure 4. In fact the correlation increases with sample size. Recall that this allows us to compute the distribution of KL under the null hypothesis by effectively treating the permutation samples as independent samples.

In this figure the "true" p-value is calculated using a distribution created using 2000 trials with independent data.

6. DISCUSSION

We have discussed a method for estimating dependency across high-dimensional measurments. Implicit in our approach is the assumption that such dependency can be approximately explained by a low-dimensional, but unknown latent variable and that additionally, the difference between dependency hypotheses can be characterized by permutations. We presented empirical results demonstrating the efficacy of the approach on a model which lends itself to analysis. Importantly, we also showed that the method points to a practical method for estimating significance via a similar permutation approach. Our experiments demonstrated that for the model we considered, estimates of dependency obtained from a single sample draw show strong correlation to significance estiamated from multiple sample draws.

7. REFERENCES

[1] S. Kullback, *Information Theory and Statistics*, Dover, 1968.

[2] E. Patrick and F.P. Fischer II, "Nonparametric feature selection," *Information Theory, IEEE Transactions on*, pp. 577–584, Sep 1969.

[3] P. Hall, "On the bootstrap and confidence intervals," *Annals of Statistcs*, vol. 14, no. 4, pp. 1431–1452, 1986.

[4] J. P. Romano, "Bootstrap and randomization tests of some nonparametric hypotheses," *Annals of Statistcs*, vol. 17, no. 1, pp. 141–159, 1989.

[5] A.T. Ihler, J.W. Fisher, and A.S. Willsky, "Nonparameteric hypothesis tests for statistical dependency," in *Trans. on signal processing, special issue on machine learning*, 2004.

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

[7] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley and Sons, Inc., 1991.

[8] R.H. Randles and D.A. Wolfe, *Introduction to the Theory of Nonparametric Statistics*, John Wiley and Sons, Inc., 1979.