INFORMATION THEORETIC STRUCTURE LEARNING WITH CONFIDENCE

Kevin R. Moon
Yale University
Department of Genetics
New Haven, Connecticut, U.S.A

Morteza Noshad, Salimeh Yasaei Sekeh, Alfred O. Hero III*
University of Michigan
Electrical Engineering and Computer Science
Ann Arbor, Michigan, U.S.A

ABSTRACT

Information theoretic measures (e.g. the Kullback Liebler divergence and Shannon mutual information) have been used for exploring possibly nonlinear multivariate dependencies in high dimension. If these dependencies are assumed to follow a Markov factor graph model, this exploration process is called structure discovery. For discrete-valued samples, estimates of the information divergence over the parametric class of multinomial models lead to structure discovery methods whose mean squared error achieves parametric convergence rates as the sample size grows. However, a naive application of this method to continuous nonparametric multivariate models converges much more slowly. In this paper we introduce a new method for nonparametric structure discovery that uses weighted ensemble divergence estimators that achieve parametric convergence rates and obey an asymptotic central limit theorem that facilitates hypothesis testing and other types of statistical validation.

Index Terms— mutual information, structure learning, ensemble estimation, hypothesis testing

1. INTRODUCTION

Information theoretic measures such as mutual information (MI) can be applied to measure the strength of multivariate dependencies between random variables (RV). Such measures are used in many applications including determining channel capacity [1], image registration [2], independent subspace analysis [3], and independent component analysis [4]. MI has also been used for structure learning in graphical models (GM) [5, 6], which are factorizable multivariate distributions that are Markovian according to a graph, called a factor graph, where edges between pairs of vertices represent pairwise dependencies [7]. GMs have been used in fields such as bioinformatics, image processing, control theory, social science, and marketing analysis. However, structure learning for GMs remains an open challenge since the most general case requires a combinatorial search over the space of all possible structures [8,9] and nonparametric approaches have poor convergence rates as the number of samples increases. This prevents reliable application of nonparametric structure learning except for impractically large sample sizes. This paper proposes a nonparametric MI-based ensemble estimator for structure learning that achieves the optimal parametric mean squared error (MSE) rate of $O(1/N)$ (where $N$ is the sample size) when the densities are sufficiently smooth and admits a central limit theorem (CLT), which enables us to perform hypothesis testing. We demonstrate this estimator in multiple structure learning experiments.

*The research in this paper was partially supported by grant W911NF-15-1-0479 from the US Army Research Office.
of optimally weighted ensemble estimation to obtain MI estimators that achieve the parametric MSE rate.

2. FACTOR GRAPH LEARNING

This paper focuses on learning a second-order product approximation (i.e. a dependence tree) of the joint probability distribution of the data. Let \( X^{(i)} \) denote the \( i \)-th component of a \( d \)-dimensional random vector \( X \). We approximate the joint probability density \( p(X) \) as a product of marginal (first-order) and conditional (second-order) probability densities denoted as \( p' (X) \). The CLT algorithm [5] provides an information theoretic method for selecting the second-order terms in \( p' (X) \). It chooses the second-order terms that minimize the Kullback-Leibler (KL) divergence between the joint density \( p(X) \) and the approximation \( p'(X) \). This reduces to constructing the maximal spanning tree where the edge weights correspond to the MI between the RVs at the vertices of the factor graph [5].

While the sum of the pairwise MI gives a measure of the quality of the approximation, it does not indicate if the approximation is a sufficiently good fit or whether a different model should be used. This problem can be framed as testing the hypothesis that \( p'(X) = p(X) \) at a prescribed false positive level. This test can be performed using MI estimation. We also propose that \( p'(X) \) can be learned by performing hypothesis testing on the MI of all pairs of RVs and assigning an edge between two vertices (RVs) if the MI is statistically different from zero. To account for the multiple comparisons bias, we control the false discovery rate (FDR) [27].

3. MUTUAL INFORMATION ESTIMATION

Information theoretic methods for learning nonlinear structures require accurate estimation of MI and estimates of its sample distribution for hypothesis testing. In this section, we extend the ensemble divergence estimators given in [26] to obtain an accurate MI estimator and use the CLT to justify a large sample Gaussian approximation to the sampling distribution. We consider general MI functionals. Let \( g : (0, \infty) \to \mathbb{R} \) be a smooth functional, e.g. \( g(u) = \ln u \) for Shannon MI or \( g(u) = u^\alpha \), with \( \alpha \in [0, 1] \), for Rényi MI. Then the pairwise MI between \( X^{(i)} \) and \( X^{(j)} \) can be defined as

\[
G_{ij} = \int g \left( \frac{p(x^{(i)}) p(x^{(j)})}{p(x^{(i)}, x^{(j)})} \right) p(x^{(i)}, x^{(j)}) \, dx^{(i)} \, dx^{(j)}. \tag{1}
\]

For hypothesis testing, we are interested in the following

\[
G(p; p') = \int g \left( \frac{p'(x)}{p(x)} \right) p(x) \, dx. \tag{2}
\]

In this paper we focus only on the case where the RVs are continuous with smooth densities. To extend the method of ensemble estimation in [26] to MI, we 1) define simple KDE-based plug-in estimators, 2) derive expressions for the bias and variance of these base estimators, and 3) then take a weighted average of an ensemble of these simple plug-in estimators to decrease the bias based on the expressions derived in step 2). To perform hypothesis testing on the estimator of (2), we invoke the CLT to specify the likelihood ratio and decision threshold. Note that we cannot simply extend the divergence estimation results in [26] to MI as [26] assumes that the random variables from different densities are independent, which may not be the case for (1) or (2).

We first define the plug-in estimators. The conditional probability density is defined as the ratio of the joint and marginal densities. Thus the ratio within the \( g \) functional in (2) can be represented as the ratio of the product of some joint densities with two random variables and the product of marginal densities in addition to \( p \). For example, if \( d = 3 \) and \( p'(X) = p\left(X^{(1)} X^{(2)}|X^{(3)}\right) p\left(X^{(3)}\right), \) then

\[
p' (X) = \frac{p\left(X^{(1)} X^{(2)}\right) p\left(X^{(2)}, X^{(3)}\right)}{p\left(X^{(2)}\right) p\left(X^{(1)}, X^{(2)}, X^{(3)}\right)}. \tag{3}
\]

For the KDEs, assume that we have \( N \) i.i.d. samples \( \{X_1, \ldots, X_N\} \) available from the joint density \( p(X) \). The KDE of \( p(X_j) \) is

\[
\hat{p}_{X,h}(X_j) = \frac{1}{N h^d} \sum_{i=1}^{N} K\left( \frac{X_j - X_i}{h} \right),
\]

where \( K \) is a symmetric product kernel function, \( h \) is the bandwidth, and \( M = N - 1 \). Define the KDEs \( \hat{p}_{12,h}(X_j^{(1)}, X_j^{(2)}) \) and \( \hat{p}_{23,h}(X_j^{(2)}, X_j^{(3)}) \) for \( p\left(X_j^{(1)}, X_j^{(2)}\right) \) and \( p\left(X_j^{(2)}, X_j^{(3)}\right) \), respectively.

Let \( \hat{p}_{X,h}(X_j) \) be defined using the KDEs for the marginal densities and the joint densities with two random variables. For example, in the example given in (3), we have

\[
\hat{p}_{X,h}(X_j) = \frac{\hat{p}_{12,h}(X_j^{(1)}, X_j^{(2)}) \hat{p}_{23,h}(X_j^{(2)}, X_j^{(3)})}{\hat{p}_{2,h}(X_j^{(2)})}.
\]

For brevity, we use the same bandwidth and product kernel for each of the KDEs although our method generalizes to differing bandwidths and kernels. The plug-in MI estimator for (2) is then

\[
\hat{G}_h = \frac{1}{N} \sum_{j=1}^{N} g \left( \frac{\hat{p}'(X_j)}{\hat{p}(X_j)} \right).
\]

The plug-in estimator \( \hat{G}_{h,ij} \) for (1) is defined similarly.

To apply bias-reducing ensemble methods similar to [26] to the plug-in estimators \( \hat{G}_h \) and \( \hat{G}_{h,ij} \), we need to derive their MSE convergence rates. As in [26], we assume that 1) the density \( p(X) \) and all other marginal densities and pairwise joint densities are \( s \geq 2 \) times differentiable and the functional \( g \) is infinitely differentiable; 2) \( p(X) \) has bounded support set \( \mathcal{S} \); 3) all densities are strictly lower bounded on their support sets. Additionally, we make the same assumption on the boundary of the support as in [26]: 4) the support is smooth wrt the kernel \( K(u) \) in the sense that the expectation of the area outside of \( \mathcal{S} \) for any RV \( u \) with smooth distribution is a smooth function of the bandwidth \( h \). This assumption is satisfied, for example, when \( \mathcal{S} \) is the unit cube and \( K(x) \) is the uniform rectangular kernel. See [28, 29] for details on the assumptions.

**Theorem 1.** If \( g \) is infinitely differentiable, then the biases are

\[
\mathbb{E}\left[ \hat{G}_{h,ij} \right] = \sum_{m=1}^{s} c_{5,i,j,m} h^m + O\left( \frac{1}{Nh^2} + h^s \right),
\]

\[
\mathbb{E}\left[ \hat{G}_h \right] = \sum_{m=1}^{s} c_{6,m} h^m + O\left( \frac{1}{Nh^d} + h^s \right). \tag{4}
\]

If \( g(t_1/t_2) \) also has \( k \), \( l \)-th order mixed derivatives \( \frac{\partial^{k+l} g(t_1/t_2)}{\partial t_1^k \partial t_2^l} \) that depend on \( t_1 \), \( t_2 \) only through \( t_1^\alpha t_2^\beta \) for some \( \alpha, \beta \in \mathbb{R} \) for each
$1 \leq k, l \leq \lambda$ then the bias of $\tilde{G}_h$ is

$$
\mathbb{E}\left[ \tilde{G}_h \right] = \sum_{m=1}^{s} c_{6,m} h^m + \sum_{m=0}^{s} \sum_{q=1}^{\lfloor \lambda/2 \rfloor} (c_{7,1,q,m} (Nh^d)^q + \ldots) \sum_{m=0}^{s} \left( \frac{1}{(Nh^d)^{q/2}} + h^s \right).
$$

(5)

The constants in (4) and (5) are independent of $h$ and $N$. The expression in (5) allows us to achieve the parametric MSE rate of $O(1/N)$ under less restrictive assumptions on the smoothness of the densities ($s > d/2$ for (5) compared to $s \geq d$ for (4)). The extra condition required on the mixed derivatives of $g$ to obtain the expression in (5) is satisfied, for example, for Shannon and Rényi information measures. Note that a similar expression could be derived for the bias of $G_{h,ij}$. However, since $s \geq 2$ is required and the largest dimension of the densities estimated in $G_{h,ij}$ is 2, we would not achieve any theoretical improvement in the convergence rate.

**Theorem 2.** If the functional $g(t_1, t_2)$ is Lipschitz continuous in both of its arguments with Lipschitz constant $C_g$, then the variance of both $G_h$ and $G_{h,ij}$ is $O(1/N)$.

The Lipschitz assumption on $g$ is comparable to assumptions required by other nonparametric distributional functional estimators [20–22, 26] and is ensured for functionals such as Shannon and Rényi informations by our assumption that the densities are bounded away from zero. The proofs of Theorems 1 and 2 share some similarities with the bias and variance proofs for the divergence functional estimators in [26]. The primary differences deal with the product of KDEs. See the appendices for the full proofs.

From Theorems 1 and 2, letting $h \rightarrow 0$ and $Nh^2 \rightarrow \infty$ or $Nh^d \rightarrow \infty$ is required for the respective MSE of $G_{h,ij}$ and $G_h$ to go to zero. Without bias correction, the optimal MSE rate is, respectively, $O\left(N^{-2/3}\right)$ and $O\left(N^{-2/(d+1)}\right)$. Using an optimally weighted ensemble of estimators enables us to perform bias correction and achieve much better (parametric) convergence rates [23, 26].

The ensemble of estimators is created by varying the bandwidth $h$. Choose $\tilde{l} = \{l_1, \ldots, l_L\}$ to be a set of positive real numbers and let $h(l)$ be a function of the parameter $l \in \tilde{l}$. Define $w = \{w(l_1), \ldots, w(l_L)\}$ and $G_w = \sum_{l \in \tilde{l}} w(l) G_h(l)$. Theorem 4 in [26] indicates that if enough of the terms in the bias expression of an estimator within an ensemble of estimators are known and the variance is $O(1/N)$, then the weight $w_0$ can be chosen so that the MSE rate of $G_{w_0}$ is $O(1/N)$, i.e., the parametric rate. The theorem can be applied as follows. For general $g$, let $h(l) = lN^{-1/(2d)}$ for $G_{h}(l)$. Denote $\psi_m(l) = l^m$ with $m \in \{1, \ldots, [s]\}$. The optimal weight $w_0$ is obtained by solving

$$
\min_{w_0} \left( \frac{1}{|\tilde{l}|^2} \sum_{l \in \tilde{l}} w(l) \right) \text{ subject to } \left( \sum_{l \in \tilde{l}} w(l) \right)^2 = 1, \sum_{l \in \tilde{l}} w(l) \psi_m(l) = 0, m \in J.
$$

(6)

It can be shown by using the last line in (6) to cancel the lower-order terms in the bias that the MSE of $G_{w_0}$ is $O(1/N)$ as long as $s \geq d$. Similarly, by using the same optimization problem we can define a weighted ensemble estimator $G_{w_{0,j}}$ of $G_{ij}$ that achieves the parametric rate when $h(l) = lN^{-1/4}$ which results in $\psi_{m}(l) = l^m$ for $m \in J = \{1, 2\}$. These estimators are similar (due to bandwidth choice) to the ODin1 divergence estimators defined in [26].

Another estimator of $G(p; p')$, similar to the ODin2 divergence estimator (due to bandwidth choice) in [26], can be derived using (5). Let $\delta > 0$, assume that $s \geq (d + \delta)/2$, and let $h(l) = lN^{-1/(d+\delta)}$. This result in the function $\psi_{1,m,q}(l) = l^{m-q}$ for $m \in \{0, \ldots, (d + \delta)/2\}$ and $q \in \{0, \ldots, (d + \delta)/\delta\}$ with the restriction that $m + q \neq 0$. Additionally we have $\psi_{2,m,q}(l) = l^{m-q}$ for $m \in \{0, \ldots, (d + \delta)/2\}$ and $q \in \{1, \ldots, (d + \delta)/(2(d + \delta - 2))\}$. These functions correspond to the lower order terms in the bias. Then using (6) with these functions results in a weight vector $w_0$ such that $G_{w_0}$ achieves the parametric rate as long as $s \geq (d + \delta)/2$.

Thus we can achieve the parametric rate for $s > d/2$.

We conclude this section by giving a CLT. This theorem provides justification for performing structural hypothesis testing with the estimators $G_{w_0}$ and $G_{w_{0,j}}$. The proof uses an application of Slutsky’s Theorem preceded by the Efros-Stein inequality that is similar to the proof of the CLT of the divergence ensemble estimators in [26]. The extension of the CLT in [26] to $G_w$ is analogous to the extension required in the proof of the variance results in Theorem 2.

**Theorem 3.** Assume that $h = o(1)$ and $Nh^d \rightarrow \infty$. If $S$ is a standard normal random variable, $L = |\tilde{l}|$ is fixed, and $g$ is Lipschitz in both arguments, then

$$
\Pr\left( \frac{G_w - \mathbb{E}[G_w]}{\sqrt{\mathbb{V}[G_w]}} \leq t \right) \rightarrow \Pr(S \leq t).
$$

4. EXPERIMENTS

We perform multiple experiments to demonstrate the utility of our proposed methods for structure learning of a GM with $d = 3$ nodes whose structure is a nonlinear Markov chain from nodes $i = 1$ to $i = 2$ to $i = 3$. That is, out of a possible 6 edges in a complete graph, only the node pairs $(1, 2)$ and $(2, 3)$ are connected by edges.

In all experiments, $X^{(1)} \sim \text{Unif}(-0.5, 0.5)$, $N^{(1)} \sim N(0, 0.5)$, and $N^{(1)}$ and $N^{(2)}$ are independent. We have $N = 500$ i.i.d. samples from $X^{(1)}$ and choose an ensemble of bandwidth parameters with $L = 50$ based on the guidelines in [26]. To better control the variance, we calculate the weight $w_0$ using the relaxed version of (6) given in [26]. We compare the results of the MI ODin1 and ODin2 ensemble estimators ($\delta = 1$ in the latter) to the simple plug-in KDE estimator. All $p$-values are constructed by applying Theorem 3 where the mean and variance of the estimators are estimated via bootstrapping. In addition, we studentize the data at each node by dividing by the sample standard deviation as is commonly done in entropic machine learning. This introduces some dependency between the nodes that decreases as $O(1/N)$. This studentization has the effect of reducing the dependence of the MI on the marginal distributions and stabilizing the MI estimates. We estimate the Rényi-$\alpha$ integral for Rényi MI with $\alpha = 0.5$; i.e., $g(u) = u^\alpha$. Thus if the ratio of densities within $(1)$ or $(2)$ is 1, the Rényi-$\alpha$ integral is also 1.

In the first type of experiments, we vary the signal-to-noise ratio (SNR) of a Markov chain by varying the parameter $\alpha$ and setting

$$
X^{(2)} = \left( X^{(1)} \right)^2 + aN^{(1)},
X^{(3)} = \left( X^{(2)} \right)^2 + bX^{(1)} + aN^{(2)}.
$$

(7)

In the second type of experiments, we create a cycle within the graph by fixing $b$ and varying $a$ or vice versa:

$$
X^{(2)} = \left( X^{(1)} \right)^2 + aN^{(1)},
X^{(3)} = \left( X^{(2)} \right)^2 + bX^{(1)} + aN^{(2)}.
$$

(8)
We first use hypothesis testing on the estimated pairwise MI to learn the structure in (7). We do this by testing the null hypotheses that each pairwise Rényi-$\alpha$ integral is equal to 1. We do not use the ODin2 estimator in this experiment as there is no theoretical gain in MSE over ODin1 for pairwise MI estimation. Figure 1 plots the mean FDR from 100 trials as a function of $a$ under this setting with significance level $\gamma = 0.1$. The dependence between $X^{(1)}$ and $X^{(3)}$ decreases as the noise increases resulting in lower mean FDR.

In the next experiment set, the CL algorithm estimates the tree structure in (7) and we test the hypothesis that $G(p; p') = 1$ to determine if the CL algorithm output gives the correct structure. Figure 2 gives the resulting mean $p$-value with error bars at the 20th and 80th percentiles from 90 trials. High $p$-values indicate that both the CL algorithm performs well and that $G(p; p')$ is not statistically different from 1. The ODin1 estimator generally has higher values than the ODin2 and KDE estimators which indicates better performance.

The final experiment set focuses on (8) where the CL tree does not give the correct structure. Figure 3 shows that the mean FDR decreases slowly for the KDE estimator. The ODin1 estimator tracks this approach better as the corresponding FDR decreases at a faster rate compared to the KDE estimator.

In general, the ODin1 estimator outperforms the other estimators in these experiments. Future work includes investigating higher dimension (larger number of vertices) and larger sample sizes. Based on the experiments in [26, 28], it is possible that the ODin2 estimator will perform comparably to the ODin1 estimator and that both ODin estimators will outperform the KDE estimator in higher dimensions.

5. CONCLUSION

We derived the convergence rates for a kernel density plug-in estimator of MI functionals and proposed nonparametric ensemble estimators with a CLT that achieve the parametric rate when the densities are sufficiently smooth. We proposed two approaches for hypothesis testing based on the CLT to learn the factor graph structure of the joint distribution. The experiments demonstrated the utility of these approaches in structure learning and the improvement of ensemble methods over the plug-in method for a low dimensional example.

A principal direction for future work is adapting the MI estimation approaches to higher dimensions. One approach is to explore alternative density estimation methods that behave better than KDEs for high feature dimension, e.g., methods incorporating information preserving dimensionality reduction methods [30, 31]. Another direction is to investigate fast, parallelizable methods for reliably computing the pairwise MI measures over large factor graphs with many nodes, e.g., in analogy to high dimensional paranormal GMs [32].
6. REFERENCES


