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.

Several structure learning algorithms have been proposed for parametric GMs including discrete Markov random fields [10], Gaussian GMs [11], and Bayesian networks [12]. Recently, the authors of [13] proposed learning latent variable models from observed samples by estimating dependencies between observed and hidden variables. Numerous other works have demonstrated that latent tree models can be learned efficiently in high dimensions (e.g. [14, 15]).

We focus on two methods of nonparametric structure learning based on ensemble MI estimation. The first method is the Chow-Liu (CL) algorithm which constructs a first order tree from the MI of all pairs of RVs to approximate the joint pdf [5]. Since structure learning approaches can suffer from performance degradation when the model does not match the true distribution, we propose hypothesis testing via MI estimation to determine how well the tree structure imposed by the CL algorithm approximates the joint distribution. The second method learns the structure by performing hypothesis testing on the MI of all pairs of RVs. An edge is assigned between two vertices (RVs) if the MI is statistically different from zero.

Accurate MI estimation is necessary for both methods. Estimating MI is often difficult, especially in high dimensions when there is no parametric model for the data. Nonparametric methods of estimating MI have been proposed including k-nearest neighbor based methods [16, 17] and minimal spanning trees [18]. However, the MSE convergence rates of the latter estimator are currently unknown, while the k-nn based methods achieve the parametric rate only when the dimension of each of the RVs is less than 3 [19].

Recent work has focused on the more general problem of nonparametric divergence estimation including approaches based on optimal kernel density estimators (KDE) [20–22] and ensemble methods [23–26]. While the optimal KDE-based approaches can achieve the parametric MSE rate for smooth densities (i.e. the densities are at least d [21] or $d/2$ [20,22] times differentiable where d is the dimension of the data), they can be difficult to construct near the density support boundary and they require knowledge of the boundary. Also, for some types of divergence functionals, these approaches require numerical integration which is computationally difficult.

In contrast, the divergence and entropy ensemble estimators in [23–26] vary the neighborhood size of nonparametric density estimators to construct an ensemble of simple plug-in divergence or entropy estimators. The final estimator is a weighted average of the ensemble of estimators where the weights are chosen to decrease the bias with only a small increase in the variance. Specifically, the ensemble estimator in [26] achieves the parametric MSE rate without any knowledge of the densities' support set when the densities are $(d+1)/2$ times differentiable. In this paper, we extend these ensemble estimation approaches to MI estimation for structure learning. We do this by deriving expressions for the bias and variance of simple plug-in MI estimators and then apply the theory

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 $\mathbf{X}^{(i)}$ denote the ith component of a d-dimensional random vector $\mathbf{X}$. We approximate the joint probability density $p(\mathbf{X})$ as a product of marginal (first-order) and conditional (second-order) probability densities denoted as $p'(\mathbf{X})$. The CL algorithm [5] provides an information theoretic method for selecting the second-order terms in $p'(\mathbf{X})$. It chooses the second-order terms that minimize the Kullback-Leibler (KL) divergence between the joint density $p(\mathbf{X})$ and the approximation $p'(\mathbf{X})$. 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 $p'(\mathbf{X}) = p(\mathbf{X})$ at a prescribed false positive level. This test can be performed using MI estimation. We also propose that $p'(\mathbf{X})$ 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 $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ can be defined as

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

For hypothesis testing, we are interested in the following

$$G\left(p; p'\right) = \int g\left(\frac{p'(x)}{p(x)}\right) p(x)dx. \quad (2)$$

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 $p'(\mathbf{X}) = p\left(\mathbf{X}^{(1)}|\mathbf{X}^{(2)}\right) p\left(\mathbf{X}^{(2)}|\mathbf{X}^{(3)}\right) p\left(\mathbf{X}^{(3)}\right)$, then

$$\frac{p'(\mathbf{X})}{p(\mathbf{X})} = \frac{p\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) p\left(\mathbf{X}^{(2)}, \mathbf{X}^{(3)}\right)}{p\left(\mathbf{X}^{(2)}\right) p\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}\right)}. \quad (3)$$

For the KDEs, assume that we have N i.i.d. samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ available from the joint density $p(\mathbf{X})$. The KDE of $p(\mathbf{X}_j)$ is

$$\tilde{\mathbf{p}}_{X,h}(\mathbf{X}_j) = \frac{1}{Mh^d} \sum_{\substack{i=1 \\ i \neq j}} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right),$$

where K is a symmetric product kernel function, h is the bandwidth, and $M = N - 1$. Define the KDEs $\tilde{\mathbf{p}}_{ik,h}\left(\mathbf{X}_j^{(i)}, \mathbf{X}_j^{(k)}\right)$ and $\tilde{\mathbf{p}}_{i,h}\left(\mathbf{X}_j^{(i)}\right)$ (for $p\left(\mathbf{X}_j^{(i)}, \mathbf{X}_j^{(k)}\right)$ and $p\left(\mathbf{X}_j^{(i)}\right)$, respectively) similarly. Let $\tilde{\mathbf{p}}'_{X,h}(\mathbf{X}_j)$ 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

$$\tilde{\mathbf{p}}'_{X,h}(\mathbf{X}_j) = \frac{\tilde{\mathbf{p}}_{12,h}\left(\mathbf{X}_j^{(1)}, \mathbf{X}_j^{(2)}\right) \tilde{\mathbf{p}}_{23,h}\left(\mathbf{X}_j^{(2)}, \mathbf{X}_j^{(3)}\right)}{\tilde{\mathbf{p}}_{2,h}\left(\mathbf{X}_j^{(2)}\right)}.$$

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

$$\tilde{\mathbf{G}}_h = \frac{1}{N} \sum_{j=1}^{N} g\left(\frac{\tilde{\mathbf{p}}'_{X,h}(\mathbf{X}_j)}{\tilde{\mathbf{p}}_{X,h}(\mathbf{X}_j)}\right).$$

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

To apply bias-reducing ensemble methods similar to [26] to the plug-in estimators $\tilde{\mathbf{G}}_h$ and $\tilde{\mathbf{G}}_{h,ij}$, we need to derive their MSE convergence rates. As in [26], we assume that 1) the density $p(\mathbf{X})$ and all other marginal densities and pairwise joint densities are $s \geq 2$ times differentiable and the functional g is infinitely differentiable; 2) $p(\mathbf{X})$ 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 $\mathcal{S}$ wrt 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*

$$\mathbb{B}\left[\tilde{\mathbf{G}}_{h,ij}\right] = \sum_{m=1}^{\lfloor s \rfloor} c_{5,i,j,m} h^m + O\left(\frac{1}{Nh^2} + h^s\right)$$

$$\mathbb{B}\left[\tilde{\mathbf{G}}_h\right] = \sum_{m=1}^{\lfloor s \rfloor} c_{6,m} h^m + O\left(\frac{1}{Nh^d} + h^s\right). \quad (4)$$

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 α, $\beta \in \mathbb{R}$ for each

$1 \leq k, l \leq \lambda$ *then the bias of* $\tilde{\mathbf{G}}_h$ *is*

$$\mathbb{B}\left[\tilde{\mathbf{G}}_h\right] = \sum_{m=1}^{\lfloor s \rfloor} c_{6,m} h^m + \sum_{m=0}^{\lfloor s \rfloor} \sum_{q=1}^{\lfloor \lambda/2 \rfloor} \left(\frac{c_{7,1,q,m}}{(Nh^d)^q} + \frac{c_{7,2,q,m}}{(Nh^2)^q} \right) h^m + O\left(\frac{1}{(Nh^d)^{\lambda/2}} + h^s \right). \quad (5)$$

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 $\tilde{\mathbf{G}}_{h,ij}$. However, since $s \geq 2$ is required and the largest dimension of the densities estimated in $\tilde{\mathbf{G}}_{h,ij}$ 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 $\tilde{\mathbf{G}}_h$ and $\tilde{\mathbf{G}}_{h,ij}$ 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 $h \to 0$ and $Nh^2 \to \infty$ or $Nh^d \to \infty$ is required for the respective MSE of $\tilde{\mathbf{G}}_{h,ij}$ and $\tilde{\mathbf{G}}_h$ 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 $\bar{l} = \{l_1, \ldots, l_L\}$ to be a set of positive real numbers and let $h(l)$ be a function of the parameter $l \in \bar{l}$. Define $w = \{w(l_1), \ldots, w(l_L)\}$ and $\tilde{\mathbf{G}}_w = \sum_{l \in \bar{l}} w(l) \tilde{\mathbf{G}}_{h(l)}$. 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 $\tilde{\mathbf{G}}_{w_0}$ 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 $\tilde{\mathbf{G}}_{h(l)}$. Denote $\psi_m(l) = l^m$ with $m \in J = \{1, \ldots, \lfloor s \rfloor\}$. The optimal weight w_0 is obtained by solving

$$\begin{array}{rl} \min_w & ||w||_2 \\ subject\ to & \sum_{l \in \bar{l}} w(l) = 1, \\ & \left| \sum_{l \in \bar{l}} w(l) \psi_m(l) \right| = 0, \ m \in J, \end{array} \quad (6)$$

It can be shown by using the last line in (6) to cancel the lower-order terms in the bias that the MSE of $\tilde{\mathbf{G}}_{w_0}$ is $O(1/N)$ as long as $s \geq d$. Similarly, by using the same optimization problem we can define a weighted ensemble estimator $\tilde{\mathbf{G}}_{w_0,ij}$ 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 the 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 results in the function $\psi_{1,m,q}(l) = l^{m-dq}$ 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 $\psi_{2,m,q}(l) = l^{m-2q}$ 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 $\tilde{\mathbf{G}}_{w_0}$ 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 $\tilde{\mathbf{G}}_{w_0}$ and $\tilde{\mathbf{G}}_{w_0,ij}$. The proof uses an application of Slutsky's Theorem preceded by the Efron-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 $\tilde{\mathbf{G}}_w$ is analogous to the extension required in the proof of the variance results in Theorem 2.

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

$$\Pr\left(\left(\tilde{\mathbf{G}}_w - \mathbb{E}\left[\tilde{\mathbf{G}}_w\right] \right) / \sqrt{\mathbb{V}\left[\tilde{\mathbf{G}}_w\right]} \leq t \right) \to \Pr(\mathbf{S} \leq t).$$

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, $\mathbf{X}^{(1)} \sim \text{Unif}(-0.5, 0.5)$, $\mathbf{N}^{(j)} \sim \mathcal{N}(0, 0.5)$, and $\mathbf{N}^{(1)}$ and $\mathbf{N}^{(2)}$ are independent. We have $N = 500$ i.i.d. samples from $\mathbf{X}^{(1)}$ 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-α 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-α 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 a and setting

$$\mathbf{X}^{(2)} = \left(\mathbf{X}^{(1)}\right)^2 + a\mathbf{N}^{(1)},$$
$$\mathbf{X}^{(3)} = \left(\mathbf{X}^{(2)}\right)^2 + a\mathbf{N}^{(2)}. \quad (7)$$

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

$$\mathbf{X}^{(2)} = \left(\mathbf{X}^{(1)}\right)^2 + a\mathbf{N}^{(1)},$$
$$\mathbf{X}^{(3)} = \left(\mathbf{X}^{(2)}\right)^2 + b\mathbf{X}^{(1)} + a\mathbf{N}^{(2)}. \quad (8)$$

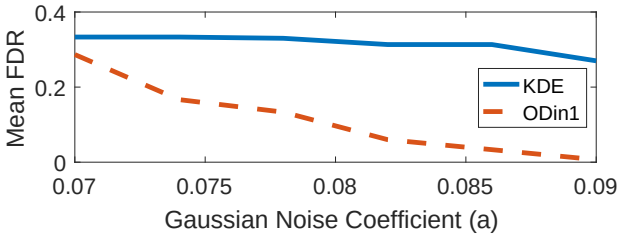

Fig. 1. The mean FDR from 100 trials as a function of a when estimating the MI between all pairs of RVs for (7) with significance level $\gamma = 0.1$. The dependence between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$ decreases as the noise increases resulting in lower mean FDR.

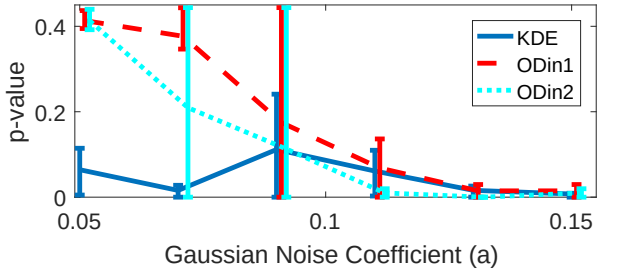

Fig. 2. The average p-value with error bars at the 20th and 80th percentiles from 90 trials for the hypothesis test that $G(p; p') = 1$ after running the CL algorithm for (7). The graphs are offset horizontally for better visualization. Higher noise levels lead to higher error rates in the CL algorithm and thus lower p-values.

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-α 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 $\gamma = 0.1$. In ths case, the FDR is either 0 (no false discoveries) or $1/3$ (one false discovery). Thus the mean FDR provides an indicator for the number of trials where a false discovery occurs. Figure 1 shows that the mean FDR decreases slowly for the KDE estimator and rapidly for the ODin1 estimator as the noise increases. Since $\mathbf{X}^{(3)}$ is a function of $\mathbf{X}^{(2)}$ which is a function of $\mathbf{X}^{(1)}$, then $G_{13} \neq 1$. However, as the noise increases, the relative dependence of $\mathbf{X}^{(3)}$ on $\mathbf{X}^{(1)}$ decreases and thus G_{13} approaches 1. The ODin1 estimator tracks this approach better as the corresponding FDR decreases at a faster rate compared to the KDE estimator.

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 include the edge between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(3)}$. We report the p-values for the hypothesis that $G(p; p') = 1$ when varying either a or b. The mean p-value with error bars at the 20th and 80th percentiles from 100 trials are given in Figure 3. In the top figure, we fix $b = 0.5$ and vary the noise parameter a while in the bottom figure we fix $a = 0.05$ and vary b. Thus the true structure does not match the CL tree and low p-values are desired. For low noise in the top figure (fixed dependency coefficient), the ODin estimators perform better than the KDE estimator and have less variability. In the bottom figure (fixed noise), the ODin1 estimator generally outperforms the others.

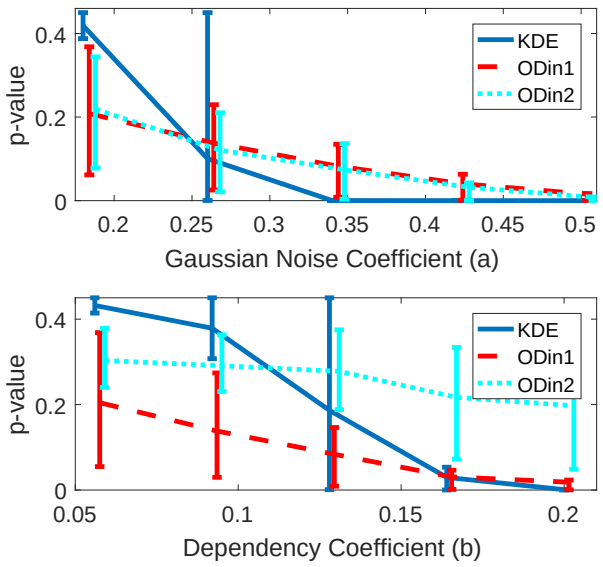

Fig. 3. The mean p-value with error bars at the 20th and 80th percentiles from 100 trials for the hypothesis test that $G(p; p') = 1$ for (8) when the CL tree does not give the correct structure. Top: $b = 0.5$ and a varies. Bottom: $a = 0.05$ and b varies. The graphs are offset horizontally for better visualization. Low p-values indicate better performance. The ODin1 estimator generally matches or outperforms the other estimators, especially in the lower noise cases.

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

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[2] Paul Viola and William M Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[3] Dávid Pál, Barnabás Póczos, and Csaba Szepesvári, "Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems*, 2010, pp. 1849–1857.

[4] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[5] C Chow and C Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[6] Alexander T Ihler, John W Fisher, and Alan S Willsky, "Nonparametric hypothesis tests for statistical dependency," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2234–2249, 2004.

[7] S.L. Lauritzen, *Graphical Models*, Clarendon Press, 1996.

[8] Karthik Mohan, Mike Chung, Seungyeop Han, Daniela Witten, Su-In Lee, and Maryam Fazel, "Structured learning of gaussian graphical models," in *Advances in Neural Information Processing Systems*, 2012, pp. 620–628.

[9] Ming Yuan and Yi Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[10] R. Kindermann and J.L. Snell, *Markov Random Fields and Their Applications*, American Mathematical Society, 1980.

[11] David Edwards, *Introduction to graphical modelling*, Springer Science & Business Media, 2012.

[12] Judea Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann, 2014.

[13] Animashree Anandkumar and Ragupathyraj Valluvan, "Learning loopy graphical models with latent variables: Efficient methods and guarantees," *The Annals of Statistics*, vol. 41, no. 2, pp. 401–435, 2013.

[14] Myung Jin Choi, Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky, "Learning latent tree graphical models," *Journal of Machine Learning Research*, vol. 12, no. May, pp. 1771–1812, 2011.

[15] Elchanan Mossel, "Distorted metrics on trees and phylogenetic forests," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 4, no. 1, pp. 108–116, 2007.

[16] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, pp. 066138, 2004.

[17] LF Kozachenko and Nikolai N Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[18] Shiraj Khan, Sharba Bandyopadhyay, Auroop R Ganguly, Sunil Saigal, David J Erickson III, Vladimir Protopopescu, and George Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, pp. 026209, 2007.

[19] Weihao Gao, Sewoong Oh, and Pramod Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *arXiv preprint arXiv:1604.03006*, 2016.

[20] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman, "Nonparametric estimation of Rényi divergence and friends," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 919–927.

[21] S. Singh and B. Póczos, "Exponential concentration of a density functional estimator," in *Advances in Neural Information Processing Systems*, 2014, pp. 3032–3040.

[22] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James Robins, "Nonparametric von Mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems*, 2015, pp. 397–405.

[23] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *Information Theory, IEEE Transactions on*, vol. 59, no. 7, pp. 4374–4388, 2013.

[24] K. R. Moon and A. O. Hero III, "Ensemble estimation of multivariate f-divergence," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 356–360.

[25] K. R. Moon and A. O. Hero III, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, 2014, pp. 2420–2428.

[26] Kevin R Moon, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero III, "Improving convergence of divergence functional ensemble estimators," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016.

[27] Dongxiao Zhu, Alfred O Hero, Zhaohui S Qin, and Anand Swaroop, "High throughput screening of co-expressed gene pairs with controlled false discovery rate (fdr) and minimum acceptable strength (mas)," *Journal of Computational Biology*, vol. 12, no. 7, pp. 1029–1045, 2005.

[28] Kevin R Moon, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero III, "Nonparametric ensemble estimation of distributional functionals," *arXiv preprint arXiv:1601.06884v2*, 2016.

[29] Kevin R Moon, Morteza Noshad, Salimeh Yasaei Sekeh, and Alfred O. Hero III, "Information theoretic structure learning with confidence," *arXiv preprint arXiv:1609.03912*, 2017.

[30] Kevin M Carter, Raviv Raich, William G Finn, and Alfred O Hero III, "Fine: Fisher information nonparametric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2093–2098, 2009.

[31] Kevin M Carter, Raviv Raich, William G Finn, and Alfred O HeroIII, "Information-geometric dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 89–99, 2011.

[32] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman, "High-dimensional semiparametric gaussian copula graphical models," *The Annals of Statistics*, pp. 2293–2326, 2012.