Estimating Information-Theoretic Quantities with Random Forests

Anonymous Author(s)
Affiliation
Address
email

Abstract

Information-theoretic quantities, such as mutual information and conditional entropy, are useful statistics for measuring the dependence between two random variables. However, estimating these quantities in a non-parametric fashion is difficult, especially when the variables are high-dimensional, a mixture of continuous and discrete values, or both. In this paper, we propose a decision forest method to estimate conditional entropy when one of the variables is categorical. We demonstrate through simulations that the decision forest estimate performs well in low and high dimensional settings. We then extend our method to estimate mutual information and show, under high-dimensional mixture settings, better performance over existing estimators.

1 Introduction

In data science investigations, it is often crucial to ask whether a relationship exists between a pair of disparate data modalities. Only when statistically significant relationships are discovered is further investigation warranted. For example, deciphering relationships is fundamental in high-throughput screening for drug discovery, precision medicine, and causal analyses [1, 2, 3].

From an information theoretic perspective, this question can be answered through two closely related quantities, conditional entropy and mutual information. Suppose we are given a pair of random variables (X, Y), where X is a d-dimensional vector and Y is a 1-dimensional, categorical variable of interest. Conditional entropy $H(Y|X)$ measures the remaining uncertainty in Y given the outcome of X. On the other hand, mutual information quantifies the shared information between X and Y.

Although both statistics are readily estimated when X and Y are low-dimensional and "nicely" distributed, an important problem arises in measuring these quantities from higher-dimensional data in a nonparametric fashion [4]. Additional issues emerge when dealing with mixtures of continuous and discrete random variables [5].

We present an algorithm for estimating these information-theoretic quantities using decision forests. Simulation demonstrate the our conditional forests performs well in low and high-dimensional settings for estimating conditional distributions and conditional entropy. We extend our algorithm to estimate mutual information which compares favorably to state-of-the-art methods. Finally, we provide a real-world application of our estimator by measuring information contained in *Drosophila* neuron labels, achieving a significantly larger estimate of mutual information between graph topology and vertex type.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

2 Problem Formulation and Related Works

Suppose we are given two random variables X and Y with support sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let x, y denote specific values that the random variables take on and $p(x), p(y)$ be the probabilities of $X = x$ and $Y = y$. The unconditioned Shannon entropy of Y is calculated as follows:

$$H(Y) = \sum_{y \in \mathcal{Y}} p(y) \log p(y) \tag{1}$$

Analogously, conditional entropy can be calculated with the following equations:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x), \tag{2}$$

where $p(y|x)$ is the conditional probability that $Y = y$ given $X = x$ and $H(Y|X = x)$ is the entropy of Y conditioned on X equaling x. In the case of a continuous random variable, the sum over the corresponding support is replaced with an integral.

Mutual information, $I(X, Y)$ can be computed from conditional entropy. Namely,

$$I(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{3}$$

However, a much more common approach to computing mutual information is the *3-H* principle [5].

$$I(X, Y) = H(Y) + H(X) - H(X, Y) \tag{4}$$

In many cases, mutual information is more informative as a notion of dependence than conditional entropy. For example, if the support of Y is large, $H(Y|X)$ can still be large, which suggests a weak relationship between Y and X, even if the two variables are highly correlated. Furthermore, mutual information has many appealing properties, such as symmetry, and is widely used in data science applications [5]. Because of this, estimating mutual information remains a more active problem than estimating conditional entropy.

In particular, the most popular approaches for estimating mutual information rely on the *3-H* principle, where each individual entropy term is computed separately. Different family of entropy estimators include kernel density estimates and ensembles of *k*-NN estimators [6, 7, 8, 9]. One method, the KSG estimator, popularized by excellent empirical performance, improves *k*-NN estimates via heuristics [10]. Other approaches include binning, von Mises estimators, etc. [11, 12].

However, many modern datasets contain a mixture of discrete and continuous variables. In these general mixture spaces, few of the above methods work well. This is mainly because individual entropies ($H(X)$, $H(Y)$, $H(Y, X)$) are not well defined or easily estimated; thus invalidating the *3-H* approach [5]. A recent approach, referred to as Mixed KSG, focuses on this issue by modifying the KSG estimator for mixed data [5].

Furthermore, computing both mutual information and conditional entropy becomes difficult in higher dimensional data. Numerical summations or integration becomes computationally intractable and nonparametric methods (k-NN, kernel density estimates, binning, etc.) typically do not scale well with increasing dimensions [13]. A neural network approach addresses this issue and scales better in high-dimensions [4]. However, existing open implementations require X and Y to be the same dimension, which does not fit the scope of this paper.

We address both problems of mixed spaces and high-dimensionality by proposing a decision forest method for estimating conditional entropy under the framework that X is any large d-dimensional random variable and Y is discrete or categorical. Because we restrict Y to be categorical, we can easily use our estimator to compute mutual information using Equation 3. At the conclusion of this paper, we discuss future work on extending our estimator for other Y settings.

3 Random Forest Estimate of Conditional Entropy

3.1 Background

CART Random Forest is a robust, powerful algorithm that leverages ensembles of decision trees for classification and regression tasks [14]. In a study of over 100 classification problems, Férnandez-Delgado et al. [15] showed that random forests have the best performance over 178 other classifiers.

Furthermore, random forests are highly scalable. Efficient implementations can build a forest of 100 trees from 110 Gigabyte data (n = 10,000,000, d = 1000) in little more than an hour [16].

A brief summary of the algorithm is given as follows: Given a labeled set of data $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ as an input, individual decision trees are grown by recursively splitting a randomly selected subsample of the input data based on an impurity measure [14]. The randomly subsampled points used in tree construction are called the 'in-bag' samples, while those that are left out (usually for evaluation) are called 'out-of-bag' samples. Additionally, only a random subset of features in X are considered for each tree. The trees are grown until nodes reach a certain minimum number of samples. The bottom-most nodes are called leaf nodes. For regression tasks, an individual decision tree predicts the response value for a given x by averaging the y values in the leaf node that x "falls" into. The random forests algorithm then outputs the average of the response values from all decision trees in the ensemble. For classification tasks, averaging is replaced by a plurality vote.

3.2 Conditional Forests

We now present **conditional forests** (CF). Unlike CART forests, conditional forests employ techniques that provide a consistent estimate of the conditional entropy in practice.

Suppose we have a decision forest trained on data drawn from random variables (X, Y). Given a point $x \in X$, the samples in each appropriate leaf node can be viewed as the remaining uncertainty about y after knowing x. In other words, leaf estimators in each decision tree represent the posterior distribution of Y given the value of X [17]. By aggregating the samples in leaf nodes across multiple decision trees in the random forest ensemble, we arrive at an estimate for the conditional distribution of Y given $X = x$, $\hat{P}(Y|X = x)$. Plugging this result into the entropy equation yields $\hat{H}(Y|X = x)$. Finally, in order to estimate conditional entropy, we note that conditional entropy can be written as an expected value (See Equation (2)):

$$H(Y|X) = \mathbb{E}_X[H(Y|X = x)] \tag{5}$$

Thus, given a dataset $\boldsymbol{X}$ of size n, the conditional entropy estimate is just the sample mean of $\hat{H}(Y|X = x)$ values.

$$\hat{H}(Y|X) = \frac{1}{n} \sum_{x \in \boldsymbol{X}} \hat{H}(Y|X = x) \tag{6}$$

There are two main differences between CART random forest and conditional forests. First, we employ honesty subsampling [18, 19]. Honest subsampling partitions the data into structure and estimation points. estimators in each leaf node but are not allowed to affect construction. Honest subsampling empirically results in less biased estimates. We add an additional partition for evaluating conditional entropy, which does not impact honesty.

Additionally, conditional forests address issues that arise when building probability distributions from finite samples. When Y is categorical, all samples in a leaf estimator may belong to one class even though the probabilities for other classes are nonzero. As a result, the empirical distribution function is biased and does not accurately capture the population class probabilities. To remedy this, we adapt a robust finite sampling technique described in Vogelstein et al. [20]. We replace all zero probabilities with $1/k\eta$ where η is the number of samples in a leaf node and k is the number of unique Y values. Similarly, we replace all one probabilities with $1 - (k - 1)/k\eta$. The conditional forest estimator is described in detail in Algorithm 1.

3.3 Training and Hyperparameter Tuning

3.3.1 Tree Construction

In constructing random forests, the two main considerations are 1) how to split the leaves and 2) how to introduce randomness between trees. Since we focus on Y being categorical, we split leaves by minimizing gini impurity, a measure popularized by its great practical performance in classification [17].

To introduce randomness, we randomly partition our data into structure, estimation, and evaluation data points during the honest sampling step. This technique is used in Denil, Matheson, and Freitas both for its good performance and as a requirement for theoretical consistency [19]. Additionally, when our data is multi-dimensional, we select a random combination of features.

Algorithm 1 Conditional Forests Estimator for Conditional Entropy

Input: data $(x_i, y_i) \in (X, Y)$, size n, number of features d,
Hyperparameters: n_trees, $min\_samples\_leaf$, max_depth, $max\_features$
for i **in** *range(n_trees)* **do**
    Partition data into `STRUCT`, `EST`, `EVAL` sets
    Train decision tree on `STRUCT`
    for (x_i, y_i) **in** `EST` **do**
        Get leaf that x_i "falls" into
        Add y_i to leaf
    end for
end for
Initialize empty array $estimates$
for (x_i, y_i) **in** `EVAL` subsamples **do**
    Initialize empty array $posterior$
    for tree **in** random forest **do**
        Get leaf that x_i "falls" into
        Add y samples in leaf to $posterior$
    end for
    Construct $\hat{P}(Y|X = x_i)$ from $posterior$ counts and robust finite sampling
    Compute $\hat{H}(Y|X = x_i)$
    Add $\hat{H}(Y|X = x_i)$ to $estimates$
end for
return $mean(estimates)$

3.3.2 Hyperparameters

Several hyperparameters can be set when constructing random forests. These include minimum samples in a leaf in order for it to undergo a split ($min\_samples\_leaf$), maximum tree depth (max_depth), number of features to subsample ($max\_features$), and number of trees (n_trees). Fortunately, however, random forest has been shown in practice to be very robust to hyperparameters [21]. For conditional forests, we allow our trees to grow relatively deep ($max\_depth = range(30, 40)$) and use general rule-of-thumbs for the other choices ($min\_samples\_leaf = 6$, $max\_features = \sqrt{d}$, $n\_trees = 300$).

3.4 Consistency

We do not prove consistency in this paper. However, there exist several important theoretical results on consistency for random forests worth discussing. In particular, Meinshausen [22] shows that **quantile regression forests** can consistently estimate the full conditional distribution of a response variable. Recent work by Athey, Tibshirani, and Wager [23] extends this further by showing that **generalized random forests** are consistent estimators for any quantity identified by local moment conditions (which includes quantile and conditional probability estimation). Although conditional forests contain key similarities with generalized random forests, namely honesty, more work is needed to extend the arguments for gini impurity as a splitting criterion. Under the assumption that random forests are consistent for conditional distributions, it is straightforward to prove consistency for conditional entropy by plugin-in and Strong Law of Large Numbers.

Finally, a significant tool conditional forests use is robust finite sampling, which leads to better estimates empirically. It is important to note that robust finite sampling does not affect consistency since as n increases, $1/k\eta$ goes to 0.

3.5 Simulated Experiments on Condition Distribution and Entropy Estimation

In this section, we perform simulations to demonstrate that the conditional forests provide good estimates of conditional entropy in low and high-dimensional settings.

We begin by first comparing the effect honesty and robust finite sampling has on the posterior distributions. Consider the following setting: let Y be Bernoulli with 50% probability to be either $+1$ or

-1; let X be normally distributed with mean $y \times \mu$ and variance one, where μ is a hyperparameter controlling effect size. A CART random forest, random forest with honest subsampling, and conditional forest (using both honest sampling and finite sample correction) are trained on data drawn from the above distribution with $\mu = 1$. We construct posterior distributions as described in Algorithm 1 and plot the posterior for $Y = 1$ in Figure 1.

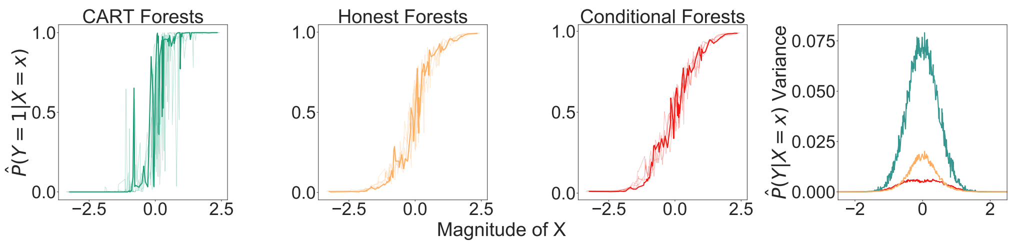

Figure 1: Comparison of estimated posterior distributions from random forest algorithms. Left plots show posterior distribution of $Y = 1$ given X from CART, honest, and CF algorithms. Five trials are plotted for each algorithm. Four trials are plotted with high transparency to show variance. Right-most plot shows variance of posterior estimates vs x. Variance was estimated from 500 trials. $\mu = 1, n =$ 10,000 for all plots.

As μ increases, the sign of our x values becomes more likely to be equal to the value of the corresponding y value. Thus, unsurprisingly, all random forest algorithms have $P(Y = 1)$ decrease to 0 as X becomes more negative, and increase to 1 as X becomes more positive. However, the posterior estimated from conditional forests has significantly lower variance than both normal CART forests and honest forests (Figure 1 right). Thus, combining honest subsampling and robust finite sampling obtains better posterior estimation in practice.

These better posterior estimates from conditional forests carry over to better estimates of conditional entropy. In the top two plots for Figure 2, we see that conditional forest estimates converge to truth as sample size increases, while honest forest estimates and CART forest estimates are biased. We also see conditional forest estimates behave correctly as μ increases. The conditional entropies drop to 0 as the two Gaussians grow farther apart.

For the high-dimensional experiment, we change X to be multivariate Gaussians, where the mean of the first dimension is still $y \times \mu$ but each additional dimension has mean 0.

$$X \sim \mathcal{N}(\underbrace{(y\mu, 0, \dots, 0)}_{\text{d}}, \underbrace{\mathbb{I}}_{\text{Identity matrix}}) \tag{7}$$

The covariance is the identity matrix[1]. Bottom plots for Figure 2 show that when $d = 40$, our conditional forest estimate still converges to truth as sample size increases. Interestingly, the bias in honest forests is also improved in this multi-dimensional setting.

4 Mutual Information Estimation

In this section, we proceed to extend our conditional forest estimate to measure the more popularized mutual information statistic. This is easily done because Y is restricted to being categorical. We can estimate $H(Y)$ by plugging in sample frequencies for each class divided by total number of samples. We can then use Equation 3 to return mutual information.

We estimate mutual information using conditional forests for a variety of different settings and compare our values to the KSG and mixed KSG estimators [10, 5]. The simulated settings we use are modified versions of the mixture of Gaussians example in Section 3. The first setting is just the standard Gaussians presented in Section 3. The second setting focuses on nonlinear discriminant boundaries by setting the variances of one of the classes to > 1:

$$\begin{aligned} X|Y = 1 &\sim \mathcal{N}(\mu, 3) \\ X|Y = -1 &\sim \mathcal{N}(-\mu, 1) \end{aligned} \tag{8}$$

[1]Because each added dimension is noise, the conditional entropy does not change. This allows us to compare behavior of our forest estimates to truth [24].

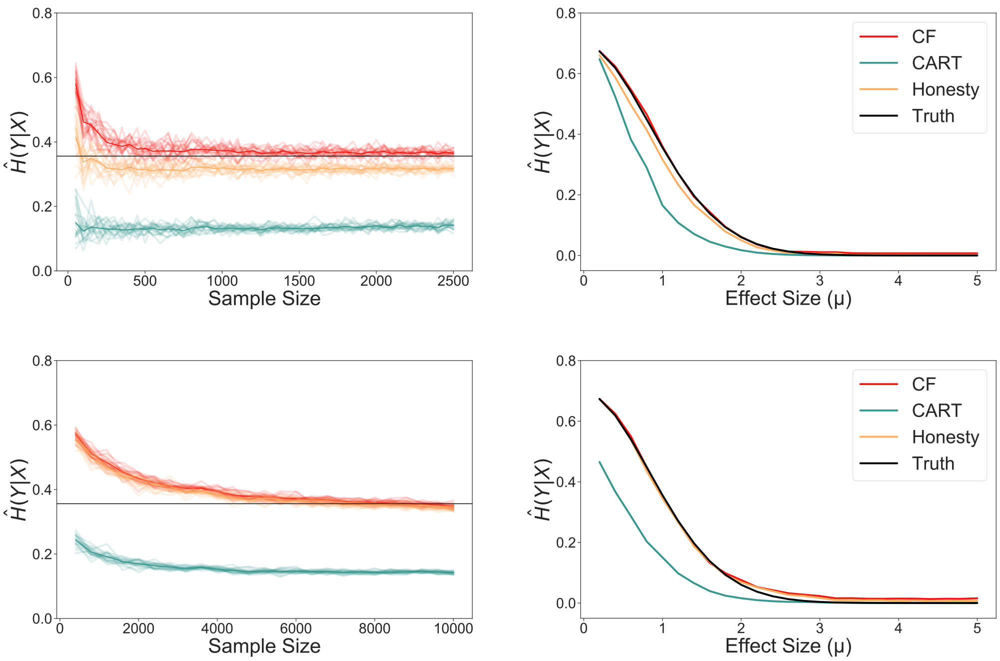

Figure 2: Behavior of random forest estimates for conditional entropy. Top plots are for $d = 1$; bottom plots are for $d = 40$. The left plot shows estimates vs. increasing sample size (n) ($\mu = 1$). Twenty trials are plotted with high transparency to show variance. Right plot shows estimates vs. increasing μ (n = 6000 for $d = 1$ and 10,000 for $d = 40$).

Next, we study the effect on mutual information estimators when classes are imbalanced. We fix $\mu = 1$ but vary a different hyperparameter p, where $P(Y = 1) = p$ and $P(Y = -1) = 1 - p$. The fourth setting emphasizes a discontinuous separating boundary between the two classes. To do this, we truncate the Gaussians so $X|Y = -1 < 0$ and $X|Y = 1 > 0$. Because there is always a separating boundary between the two classes, mutual information is maximal and does not change as μ varies. Finally, the last setting is three classes of multivariate Gaussians. To make the class means equidistant from each other, the minimal number of dimensions is 2. More specifically,

$$\begin{aligned} X|Y = 0 &\sim \mathcal{N}((0, \mu), \mathbb{I}) \\ X|Y = 1 &\sim \mathcal{N}((\mu, 0), \mathbb{I}) \\ X|Y = 2 &\sim \mathcal{N}((-\mu, 0), \mathbb{I}) \end{aligned} \tag{9}$$

We compute normalized mutual information, $I(X, Y)/min(H(X), H(Y))$, for each setting when $d = 6$ and $d = 40$; each added dimension is an independent, standard Gaussian. Again, because every additional dimension is noise, mutual information does not change. [24].

Figure 3 shows the performance of each estimator. When $d = 6$, conditional forests, KSG, and mixed KSG all do reasonably well. However, when the Gaussians are truncated, the KSG estimator is not able to discern the separating boundary. The mixed KSG does better but still suffers a bias when the two classes are close (μ is small). Only the conditional forest estimator converges to truth. When $d = 40$, the KSG estimator suffers a significant bias while the mixed KSG drops to 0 completely.

On the other hand, the conditional forest estimator correctly returns estimates equal or almost equal to the lower dimensional setting. Finally, when we extend the Gaussian setting to three classes, the KSG suffers a worse bias in both low and high-dimensional settings.

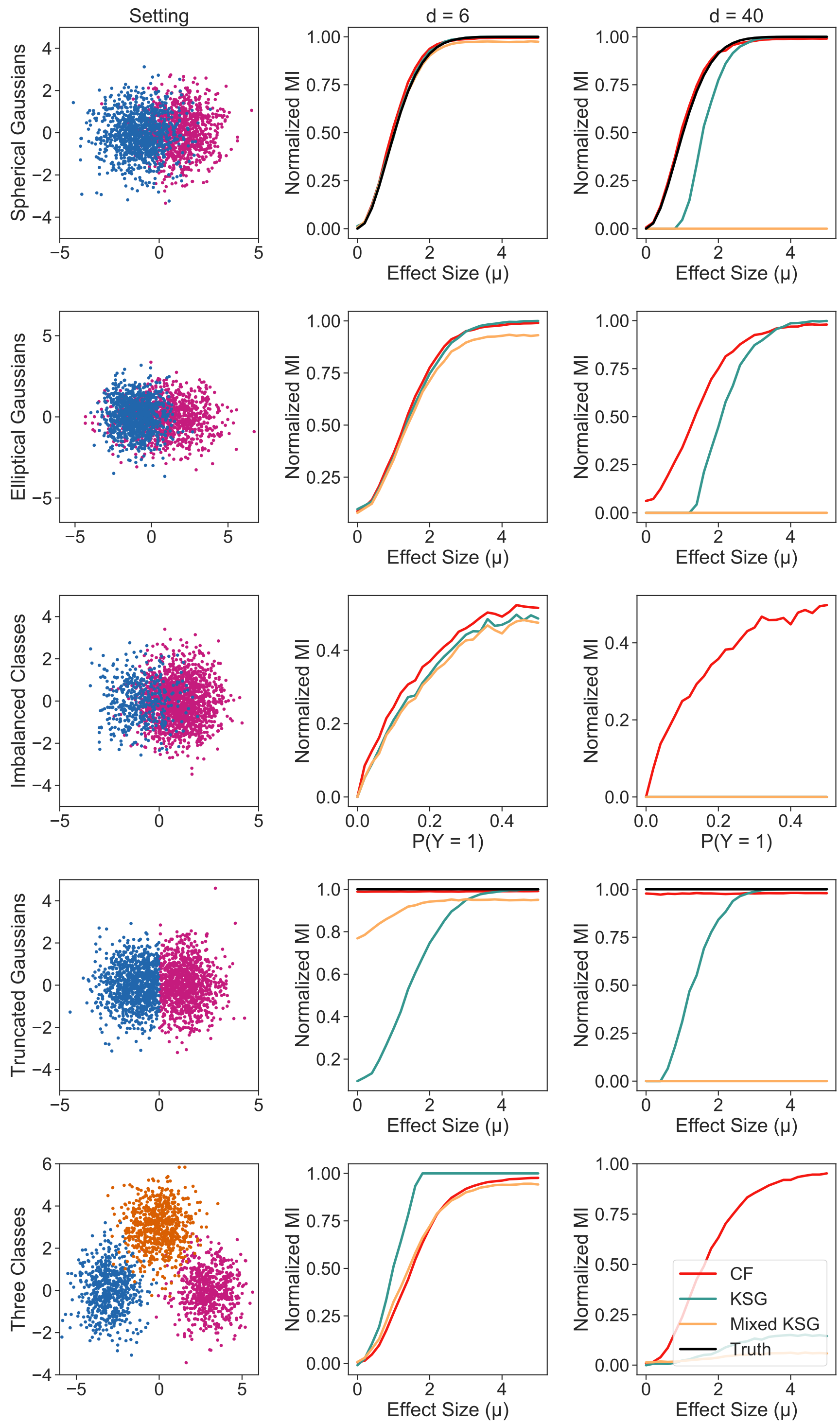

Figure 3: Mutual information estimates in different Gaussian settings. Left-most column shows from top-down: mixed Gaussian setting with spherical covariance, elliptical covariance, imbalanced classes ($Y = 1$ with probability p), truncation, and three classes. Middle column shows CF, KSG, and Mixed KSG estimates vs. increasing μ when $d = 6$. Note, for imbalanced classes, $\mu = 1$ and p increases. Right column is identical to middle except $d = 40$.

5 Mutual Information in Drosophila Neural Data

An immediate application of our random forest estimate of conditional entropy is measuring information contained in neuron labels for the larval *Drosophila* mushroom body (MB) connectome. This dataset obtained via serial section transmission electron microscopy provides a real and important example for investigation into synapse-level structural connectome modeling [25].

The connectome consists of 213 different neurons ($n = 213$) with four distinct types: Kenyon Cells (KC), Input Neurons (MBIN), Output Neurons (MBON), and Projection Neurons (PN). Each neuron comes with a mixture of categorical and continuous features (claw, age, dist, connectome cluster label) ($d = 4$). An important initial scientific question may be whether or not different neuron types correspond to different structural differences in the neuron features. We can compute normalized mutual information with Y as the neuron type and X as the other features. We expect mutual information to be high (Figure 4). However, from our estimates, only conditional forest is able to detect significant shared information between neuron type and neuron features (Table 1).

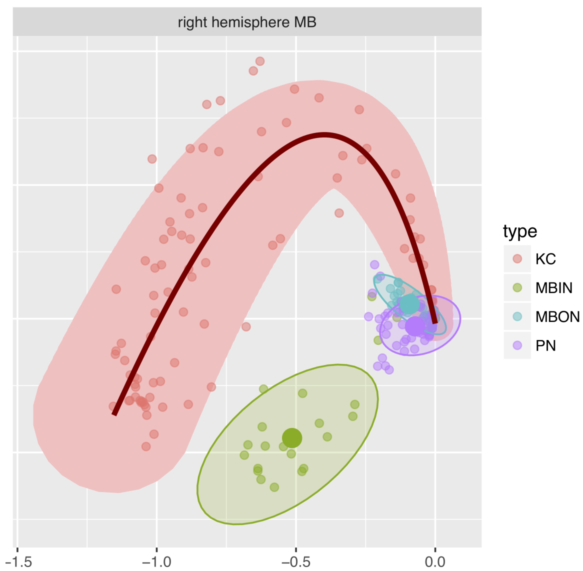

Figure 4: Adjacency spectral embedding applied to the MB connectome shows clear cluster groups for each neuron type. This suggests a strong dependency between neuron type and neural features.

Algorithm	Mutual Information
CF	.6485
KSG	.2434
Mixed KSG	.0468

Table 1: Normalized mutual information estimates for neuron type and neural features.

6 Conclusion

We presented conditional forests, a nonparametric method of estimating conditional entropy through randomized decision trees. Empirically, conditional forests performs well in low and high dimensional settings. Furthermore, when extending our estimator to estimate mutual information, conditional forest performs better than the mixed KSG and KSG estimators in a variety settings.

Although in this paper focuses on categorical Y, it is easy to modify the conditional forest algorithm for continuous Y as well. Regression trees can be used in place of classification trees. Computing the posterior distribution $\hat{P}(Y|X = x)$ can be accomplished with a kernel density estimate instead of simply binning the probabilities. When Y is multivariate, a heuristic approach such as subsampling Y dimensions or using multivariate random forests (cite paper) can be explored.

On the theoretical side, important next steps include proofs for consistency and convergence rates. Studying the behavior of conditional forest estimates in more complicated nonlinear, high-dimensional settings should be explored as well. Practical applications such as dependence testing and k-sample testing for high-dimensional, nonlinear data will be natural applications for these information theoretic estimates.

References

[1] J. H. Zhang, T. D. Chung, and K. R. Oldenburg, "A simple statistical parameter for use in evaluation and validation of high throughput screening assays," *J Biomol Screen*, vol. 4, no. 2, pp. 67–73, 1999.

[2] J. W. Prescott, "Auantitative imaging biomarkers: the application of advanced image processing and analysis to clinical and preclinical decision making," *J Digit Imaging*, vol. 26, pp. 97–108, Feb. 2013.

[3] J. Pearl, *Causality: Models, Reasoning, and Inference*. New York, NY, USA: Cambridge University Press, 2000.

[4] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, (Stockholmsmässan, Stockholm Sweden), pp. 531–540, PMLR, 10–15 Jul 2018.

[5] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 5986–5997, Curran Associates, Inc., 2017.

[6] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van Der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[7] N. Leonenko, L. Pronzato, and V. Savani, "Estimation of entropies and divergences via nearest neighbors," 2008.

[8] T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *Ann. Statist.*, vol. 47, pp. 288–318, 02 2019.

[9] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE Trans Inf Theory*, vol. 59, pp. 4374–4388, Jul 2013.

[10] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004.

[11] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A*, vol. 33, pp. 1134–1140, Feb 1986.

[12] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and j. m. robins, "Nonparametric von mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 397–405, Curran Associates, Inc., 2015.

[13] S. Gao, G. V. Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," *CoRR*, vol. abs/1411.2003, 2014.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[15] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?," *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.

[16] B. Li, X. Chen, M. J. Li, J. Z. Huang, and S. Feng, "Scalable random forests for massive data," in *Proceedings of the 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, PAKDD'12, (Berlin, Heidelberg), pp. 135–146, Springer-Verlag, 2012.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

[18] L. Breiman, *Classification and Regression Trees*. Routledge, 1984.

[19] M. Denil, D. Matheson, and N. D. Freitas, “Narrowing the gap: Random forests in theory and in practice,” in *Proceedings of the 31st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.), vol. 32 of *Proceedings of Machine Learning Research*, pp. 665–673, Jun 2014.

[20] J. T. Vogelstein, W. G. Roncal, R. J. Vogelstein, and C. E. Priebe, “Graph classification using signal-subgraphs: Applications in statistical connectomics,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1539–1551, July 2013.

[21] P. Probst, A.-L. Boulesteix, and B. Bischl, “Tunability: Importance of hyperparameters of machine learning algorithms,” *Journal of Machine Learning Research*, vol. 20, no. 53, pp. 1–32, 2019.

[22] N. Meinshausen, “Quantile regression forests,” *Journal of Machine Learning Research*, vol. 7, pp. 983–999, 2006.

[23] S. Athey, J. Tibshirani, and S. Wager, “Generalized random forests,” *Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019.

[24] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman, “On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions,” AAAI, pp. 3571–3577, AAAI Press, 2015.

[25] K. Eichler, F. Li, A. Litwin-Kumar, Y. Park, I. Andrade, C. M. Schneider-Mizell, T. Saumweber, A. Huser, C. Eschbach, B. Gerber, R. D. Fetter, J. W. Truman, C. E. Priebe, L. F. Abbott, A. S. Thum, M. Zlatic, and A. Cardona, “The complete connectome of a learning and memory centre in an insect brain,” *Nature*, vol. 548, pp. 175–182, 08 2017.