First-Order Dependence Trees with Cumulative Residual Entropy

Muhammed Sutcu and Ali E. Abbas

Industrial and Systems Engineering Department, University of Illinois Urbana-Champaign,
104 S Mathews Ave Urbana, IL 61801

Abstract. This paper presents a method to approximate discrete joint probability distributions using first-order dependence trees and the recent concept of cumulative residual entropy. A first-order dependence tree is one where each variable is conditioned on at most one variable. The cumulative residual entropy measure is the entropy functional applied to the survival function instead of the probability measure. We formulate the cumulative residual Kullback-Leibler (KL)-divergence and the cumulative residual mutual information measures in terms of the survival function. We then show that the optimal first-order dependence tree approximation of the joint distribution using the cumulative Kullback-Leibler divergence is the one with the largest sum of cumulative residual mutual information pairs. The results parallel Chow-Liu's approximation of joint probability distributions using the traditional Kullback-Leibler divergence and mutual information but applied to survival functions. We compare the approximation results with those of Chow-Liu using the traditional entropy measure. Using a Monte Carlo simulation, we show that the two approximations perform almost equally but they are not same.

Keywords: Dependence tree; Mutual information; Cumulative residual entropy; Joint probability distribution; Approximation of probability distributions; Chow Liu tree; Survival function
PACS: 89.70.Cf

INTRODUCTION

The problem of approximating joint probability distributions is fundamental in numerous fields including decision analysis and information systems [1, 2, 3, 4, 5]. It is often convenient to approximate a joint probability distribution for both the computational aspects of conducting Bayesian inference and for the elicitation of the conditional probabilities. Several methods have been proposed to approximate joint probability distributions in the literature. For example, Chow and Liu [6] approximate a joint probability distribution using the notion of a "first-order tree dependence" where each child has only one parent. Ku and Kullback [7] generalized Chow and Liu's algorithm [6], allowing any lower-order marginal distributions to be used in the approximation. In related work, Keefer [8] presented a model for approximating probability dependence among binary events and Abbas [9, 10, 11, 12, 13] explored the use of the maximum entropy principle to approximate joint distributions and utility functions using any number of lower order assessments and partial information.

In this paper we determine the best first order dependence tree approximation using the concept of cumulative residual entropy (CRE), which is an alternative measure of entropy that was introduced by Rao et.al. [14] using cumulative probability distributions. The cumulative residual entropy measure requires numeric variables for the construction of a cumulative distribution as opposed to the discrete entropy where probabilities can be assigned to non-numeric variables.

The contribution of this paper is as follows. We first formulate the concepts of Kullback-Leibler (KL)-divergence [15] and mutual information [16] in terms of cumulative residual entropy. These definitions are different from Baratpour and Rad's cumulative KL, and Wang et.al's cross cumulative residual entropy definitions [17, 18]. We then derive the optimal first-order dependence tree approximation of the joint distribution in terms of the cumulative residual KL-divergence. We show that the optimal tree approximation is the one with the highest sum of cumulative residual mutual information pairs. This result parallels the Chow-Liu dependence tree formulation that was based on Shannon's entropy [19] but uses the survival function instead of the probability.

The remainder of this paper is structured as follows: Section 2 presents the basic notation and definitions that will be used in the remaining sections of the paper. Section 3 discusses the optimal cumulative residual entropy-based dependence tree. Section 4 presents a Monte Carlo simulation to quantify and compare the accuracy of the cumulative residual entropy approximation with the Chow-Liu approximation.

BASIC NOTATIONS AND DEFINITIONS

This section presents the basic notation and definitions that will be used in the remaining sections of the paper. Let

$$F_x(x) = P(X \leq x) \tag{1}$$

be the marginal cumulative distribution function of the random variable X, and let

$$F(x,y) = P(X \leq x, Y \leq y) \tag{2}$$

be the bivariate cumulative distribution function of random variables X and Y.

Define a marginal survival function for variable X as

$$S_x(x) = 1 - F_x(x) = P(X > x) \tag{3}$$

and a bivariate survival function for random variables X and Y as

$$S(x,y) = P(X > x, Y > y) = 1 - F_x(x) - F_y(y) + F(x,y) \tag{4}$$

The conditional survival function between two variables (X given Y) is

$$S_{x|y}(x|y) \cong \frac{S(x,y)}{S_y(y)} \tag{5}$$

Shannon [19] defined the entropy measure using a probability mass function as

$$H(X) = -\sum_{i=1}^{n} p_i log p_i \tag{6}$$

where p_i is the probability of outcome i.

Kullback and Leibler [15] extended the entropy definition and introduced a new measure. The KL- divergence of Q from P is defined as

$$D_{KL}(P||Q) = \sum_{i=1}^{n} p(x_i)log\left(\frac{p(x_i)}{q(x_i)}\right) \tag{7}$$

Mutual information is a special case of KL-divergence between the joint distribution of two variables and the product of their marginals. The mutual information between variables X and Y is

$$MI(X;Y) = \sum_{x\varepsilon X}\sum_{y\varepsilon Y} p(x,y)log\left(\frac{p(x,y)}{p_x(x)p_y(y)}\right) \tag{8}$$

Rao et.al. [14] proposed an alternative entropy measure, $\varepsilon(S(x))$, using cumulative survival functions as

$$\varepsilon(S(x)) = -\int_0^\infty S(x)logS(x)dx \tag{9}$$

After Rao et.al.'s entropy definition, Baratpour and Rad [17] defined a measure similar to KL-divergence as

$$CKL(S_F : S_G) = \int_0^\infty S_F(x)ln\frac{S_F(x)}{S_G(x)}dx - [E(F) - E(G)] \tag{10}$$

where $E(F)$ and $E(G)$ are expected values of variables X and Y, respectively.

Wang et.al. [18] defined a quantity similar to mutual information which is called Cross Cumulative Residual Entropy (CCRE) as

$$CCRE(X:Y) = \varepsilon(X) - E[\varepsilon(Y|X)] \tag{11}$$

In this paper, we apply cumulative residual entropy to KL-divergence and mutual information. Our definitions are different from Baratpour and Rad's cumulative KL, and Wang et.al's cross cumulative residual entropy definitions. We simplify the cumulative KL divergence by removing the expected values of random variables and taking the absolute value of the expression.

Definition 1: ***Cumulative Residual Kullback-Leibler Divergence***

The cumulative residual KL-divergence between two distributions S_T and S_A is

$$KL_{CRE}(S_T||S_A) = \left|\sum S_T(x)log\frac{S_T(x)}{S_A(x)}\right| \tag{12}$$

where $KL_{CRE}(S_T||S_A) \geq 0$ and equality holds if and only if $S_T = S_A$.

We also define another quantity similar to mutual information and called it as Cumulative Residual Mutual Information (MI_{CRE}).

Definition 2: ***Cumulative Residual Mutual Information***

The cumulative residual mutual information MI_{CRE}, between variables X and Y is

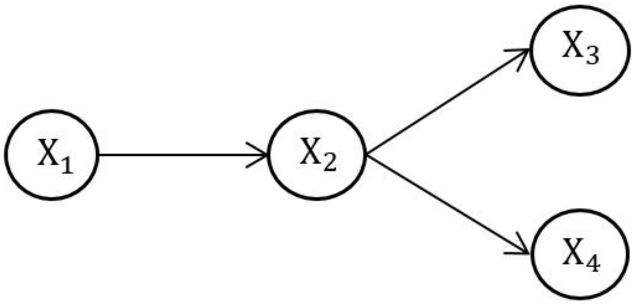


FIGURE 1. Example of a four-dimensional dependence tree

$$MI_{CRE} = \left| \sum_{x\epsilon X} \sum_{y\epsilon Y} S(x,y) \left[log \left(\frac{S(x,y)}{S_x(x)S_y(y)} \right) \right] \right| \tag{13}$$

The definition (13) is symmetric and expressed as a divergence of the product of the marginal survival functions of two random variables from the joint survival function of random variables.

FIRST-ORDER DEPENDENCE TREES USING CRE

In first order dependence trees, each variable is conditioned on at most one variable, and there cannot be a cycle between the variables. Figure-1 shows an example of a first-order dependence tree of four variables. A four-variate joint distribution $P(X_1,X_2,X_3,X_4)$ can be approximated as in Figure-1 using a first-order dependence tree as

$$P^t(X_1,X_2,X_3,X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_2) \tag{14}$$

In Chow-Liu's first-order dependence tree approach, the mutual information between each two variables is calculated. Chow and Liu show that a probability distribution of first order dependence tree structure is the best approximation to the true distribution with respect to the KL-divergence measure if its dependence tree has the maximum sum of mutual information pairs from all such first order dependence trees.

Chow and Liu provide a simple algorithm for constructing the optimal tree and determine which conditional probabilities are to be used in the product approximation. The method is based on evaluating the mutual information pairs of variables. So, the algorithm simply adds the maximum mutual information pairs to the tree.

We define the optimum first order dependence tree formulation with respect to the Cumulative Residual KL-divergence measure.

Theorem 1:

The first-order dependence tree approximation is an optimum first order tree approximation of the joint distribution with respect to the Cumulative Residual KL-divergence in (12) if its dependence tree has the maximum sum of cumulative residual mutual information pairs.

Proof: See appendix.

TABLE 1. 2x2x2x2 Joint Probability Distribution

X1	X2	X3	X4	$P(x_1, x_2, x_3, x_4)$
0	0	0	0	**0.10**
0	0	0	1	**0.10**
0	0	1	0	**0.05**
0	0	1	1	**0.05**
0	1	0	0	**0.00**
0	1	0	1	**0.00**
0	1	1	0	**0.10**
0	1	1	1	**0.05**
1	0	0	0	**0.05**
1	0	0	1	**0.10**
1	0	1	0	**0.00**
1	0	1	1	**0.00**
1	1	0	0	**0.05**
1	1	0	1	**0.05**
1	1	1	0	**0.15**
1	1	1	1	**0.15**
				$\sum p_i =$ **1.000**

To illustrate the implications of Theorem 1, we now apply the CRE approach to the same probability distribution used in [6] to compare the two approaches.

Example: Consider four binary variables where each variable takes on values "0" and "1". Table-1 shows the outcomes and corresponding probabilities of joint distribution.

To compare both methods we calculate the mutual information (MI) and cumulative residual mutual information between pairs of variables (MI_{CRE}) using equations (8) and (13) respectively. All combination of pairs of variables and mutual information and cumulative residual mutual information quantities are given at Table-2.

We construct the optimal first order dependence trees using mutual information and cumulative residual mutual information pairs at Table-2. Figure-2 shows the optimal dependence tree approximations. The first three diagrams in Figure-2 are same as what is found in Chow-Liu's paper. The fourth one is the dependence tree obtained using cumulative residual entropy.

MONTE CARLO SIMULATION FOR CRE FIRST ORDER DEPENDENCE TREE

For numeric illustration, we discuss the simulation steps of a four-variate distribution and each variable has three different values ($3 \times 3 \times 3 \times 3$ joint distribution).

Step 1: Generate a joint distribution;

1. Generate 80 independent samples from a uniform [0,1] distribution

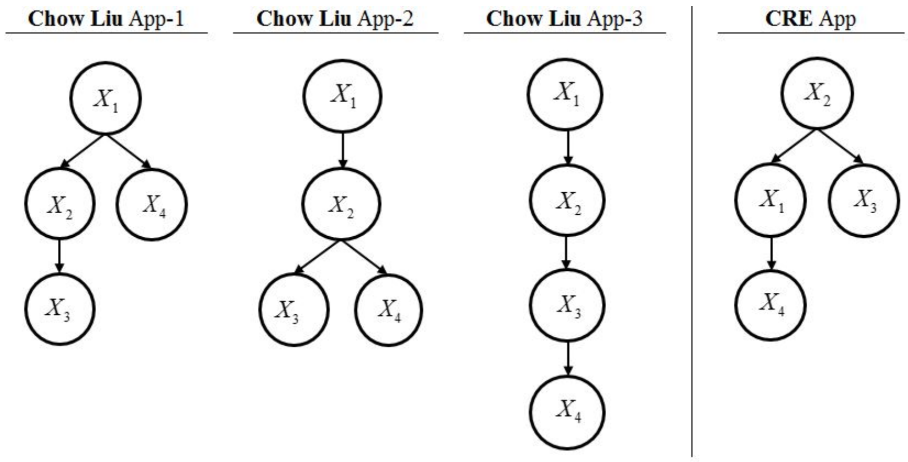
$$s_1 \varepsilon U[0,1], s_2 \varepsilon U[0,1], \cdots\cdots, s_{80} \varepsilon U[0,1] \tag{15}$$

2. Sort 80 independent samples from lowest to highest to form an ascending order

$$u_1 \leq u_2 \leq \cdots\cdots \leq u_{80} \tag{16}$$

TABLE 2. Mutual Information pairs of Example-1

Pair of variables	MI	MI_{CRE}
X1-X2	0.07900	0.11175
X1-X3	0.00005	0.00249
X1-X4	0.00510	0.02610
X2-X3	0.18900	0.17872
X2-X4	0.00510	0.02383
X3-X4	0.00510	0.02383



	Chow-Liu 1	Chow-Liu 2	Chow Liu 3	CRE
Absolute Deviation	**0.26006**	0.27521	0.27521	**0.26006**
Least Squares	**0.00598**	0.00599	0.00599	**0.00598**

FIGURE 2. Optimal Tree Approximations Approximated by Chow-Liu and CRE Method

3. Take the difference between each two consecutive samples.

$$u_1 - 0, u_2 - u_1, u_3 - u_2, \cdots\cdots, u_{80} - u_{79}, 1 - u_{80} \tag{17}$$

The increments form the $3 \times 3 \times 3 \times 3$ joint distribution sample at the end of step-1.

Step 2: In this step, we calculate mutual information and cumulative residual mutual information for all possible pairs of variables by using equation (8) and (13), respectively. We here have six different mutual information pairs: $(X1-X2); (X1-X3); (X1-X4); (X2-X3); (X2-X4); (X3-X4)$. Then, we assign mutual information pairs as branch weights for each pair of variables.

Step 3: Construct the first order dependence tree using mutual information and cumulative residual mutual information calculated in step-2 which has maximum total mutual information and cumulative residual mutual information.

TABLE 3. Simulation Results of Chow-Liu and CRE based Approximations

	Absolute Deviation				Least Squares			
	Chow-Liu		CRE		Chow-Liu		CRE	
	Mean	St.dev	Mean	St.dev	Mean	St.dev	Mean	St.dev
Three binary variables	0.1810	0.1256	0.1822	0.1264	0.0074	0.0094	0.0077	0.0099
Three variables each has three values	0.3886	0.0866	0.3934	0.0914	0.0095	0.0049	0.0098	0.0053
Three variables each has four values	0.4883	0.0596	0.4942	0.0620	0.0073	0.0017	0.0071	0.0019
Three variables each has five values	0.5505	0.0469	0.5484	0.0441	0.0045	0.009	0.0042	0.008
Four binary variables	0.3612	0.1121	0.3548	0.1118	0.0138	0.0084	0.0136	0.0083
Four variables each has three values	0.5888	0.0549	0.5879	0.0545	0.0077	0.0018	0.0074	0.0017
Four variables each has four values	0.6638	0.0319	0.6557	0.0311	0.0032	0.0004	0.0030	0.0004
Four variables each has five values	0.6980	0.0254	0.6915	0.0241	0.0015	0.0003	0.0015	0.0003

We generated 10 million discrete joint probability distribution samples to check performance and accuracy of Chow-Liu's method and cumulative residual entropy method. For convenience, we run the simulation with several different combination of joint distributions including; three binary variables, three three-outcome variables, three four-outcome variables, three five-outcome variables, four binary variables, four three-outcome variables, four four-outcome variables, and four five-outcome variables.Table 3 displays a summary of mean and variance of errors of second order joint probability distributions calculated by Chow-Liu and cumulative residual entropy methods.

From Table 3, for the case of $3 \times 3 \times 3 \times 3$ joint distributions, we have found that Chow-Liu and CRE methods' results are almost exactly same after 10 million runs. The mean of absolute deviation for Chow-Liu method is 0.5888, and for CRE method is 0.5879. The ratio of the means of absolute deviation of Cho-Liu's method to the CRE method is less than (0.5888/0.5879)=0.15

The second observation we can see from Table-3 is that the mean and variance of errors are very close in the long run for the two methods which means that the two approximation methods are very close but they are not same. Also, we found that Chow-Liu and CRE methods approximated the same first order dependence tree more than

Anoyher observation after simulation results we can see that the mean value of absolute deviation for our CRE approximation method and traditional Chow-Liu method are increasing when the number of outcomes of a variable increases. Also, the mean value of least squares error decreases when the outcomes of a variable increase. This implies that these two methods are sensitive to the number of variables and its outcomes.

CONCLUSION

In this paper, we discussed the problem of approximating multidimensional discrete probability distributions using first-order dependence trees. We showed that the optimal first-order dependence tree approximation in terms of the cumulative residual KL divergence is the one with the largest sum of cumulative residual mutual information pairs. We then ran a Monte Carlo simulation to illustrate the performance of the approximation. The results show that the cumulative residual approximation and the Chow-Liu approximations give almost identical accuracy results.

Cumulative residual entropy method can be used as an alternative method to Chow-Liu's method if cumulative functions, especially survival functions, are present. Therefore, we don't need to calculate the density functions to approximate first order dependence trees, and by using CRE method we can directly approximate first-order dependence trees from cumulative functions.

APPENDIX

In this proof, we follow the proof of Chow-Liu's first order dependence tree theorem but applied to the survival functions and the two proposed measures defined on page-3: cumulative residual KL-divergence (KL_{CRE}) and cumulative residual mutual information (MI_{CRE}). Let S_A be a second order product approximation (first order dependence tree). The optimal first-order dependence tree is determined by minimizing the cumulative residual KL-divergence between true distribution S_T and approximate distribution S_A as $S_{A^*} = argminKL_{CRE}[S_T||S_A]$. We first have the equation

$$KL_{CRE}(S_T||S_A) = -\left|\sum S_T(x) log \frac{S_T(x)}{S_A(x)}\right| = \\ -\left|\sum S_T(x) log S_T(x) + \sum S_T(x) log \prod_{i=1}^{n} S_A(x_i|x_{j(i)})\right| \quad (18)$$

The first term of the right hand side of equation (18) is cumulative residual entropy of true distribution S_T, $\varepsilon(S_T) = -\sum S_T(x) log S_T(x)$. So, re-arranging equation (18) gives

$$KL_{CRE}(S_T||S_A) = -\left|-\varepsilon(S_T) + \sum S_T(x) log \prod_{i=1}^{n} S_A(x_i|x_{j(i)})\right| \quad (19)$$

We can write conditional survival function $S_A(x_i|x_{j(i)})$ as $\frac{S_A(x_i,x_{j(i)})}{S_A(x_{j(i)})}$, then equation (19) can be written as

$$KL_{CRE}(S_T||S_A) = -\left|-\varepsilon(S_T) + \sum_{i=1}^{n} \sum_{x_i,x_{j(i)}} S_T(x_i,x_{j(i)}) log \frac{S_A(x_i,x_{j(i)})}{S_A(x_{j(i)})}\right| \quad (20)$$

Multiplying the numerator and the denominator of last term of equation (20) by marginal survival function, $S_A(x_i)$

$$KL_{CRE}(S_T||S_A) = -\left|-\varepsilon(S_T) + \sum_{i=1}^{n} \sum_{x_i,x_{j(i)}} S_T(x_i,x_{j(i)}) log \frac{S_A(x_i,x_{j(i)})}{S_A(x_{j(i)})} \times \frac{S_A(x_i)}{S_A(x_i)}\right| \quad (21)$$

Our aim to multiply by $S_A(x_i)$ is to re-arrange the equation (20) and obtain cumulative residual mutual information. By using the logarithm of a product is the sum of the logarithms of the factors rule, we can rewrite equation (21) as

$$KL_{CRE}(S_T||S_A) = \\ -\left|\varepsilon(S_T)+\sum_{i=1}^{n}\sum_{x_i}S_T(x_i,x_{j(i)})logS_A(x_i)+\sum_{i=1}^{n}\sum_{x_i,x_{j(i))}}S_T(x_i,x_{j(i)})log\frac{S_A(x_i,x_{j(i)})}{S_A(x_i)S_A(x_{j(i)})}\right| \quad (22)$$

In order to minimize the cumulative residual KL information, we expect that the true and approximate distribution satisfy the equality condition that achieves the maximal value with $S_A(x_i,x_{j(i)}) = S_T(x_i,x_{j(i)})$. We rewrite the equation (22) by substituting $S_A(x_i,x_{j(i)})$ with $S_T(x_i,x_{j(i)})$

$$KL_{CRE}(S_T||S_A) = \\ -\left|\varepsilon(S_T)+\sum_{i=1}^{n}\sum_{x_i}S_T(x_i)logS_T(x_i)+\sum_{i=1}^{n}\sum_{x_i,x_{j(i))}}S_T(x_i,x_{j(i)})log\frac{S_A(x_i,x_{j(i)})}{S_T(x_i)S_T(x_{j(i)})}\right| \quad (23)$$

Using the rule of subadditivity rule of absolute values, we can rewrite the equation (23) as

$$KL_{CRE}(S_T||S_A) \leq -|-\varepsilon(S_T)|+\left|\sum_{i=1}^{n}\sum_{x_i}S_T(x_i)logS_T(x_i)\right|+ \\ \left|\sum_{i=1}^{n}\sum_{x_i,x_{j(i))}}S_T(x_i,x_{j(i)})log\frac{S_A(x_i,x_{j(i)})}{S_T(x_i)S_T(x_{j(i)})}\right| \quad (24)$$

So, minimizing the cumulative residual KL divergence is same as minimizing the right hand side of the equation (21). First and second terms of right hand side of equation (21) are independent to the dependence tree, therefore minimizing the cumulative residual KL divergence is equivalent to maximizing the sum of cumulative residual mutual information in each branch.

ACKNOWLEDGMENTS

This research was partially supported by the NSF CMMI 12-58482, NSF CMMI 13-01150, and NSF SES 08-46417 (CAREER) grants.

REFERENCES

1. P. M. Lewis, *Information and control* **2**, 214–225 (1959).
2. D. T. Brown, *Information and Control* **2**, 386–392 (1959).
3. D. Brook, and D. Evans, *Biometrika* **59**, 539–549 (1972).
4. A. C. Miller III, and T. R. Rice, *Management Science* **29**, 352–362 (1983).

5. K. Høyland, and S. W. Wallace, *Management Science* **47**, 295–307 (2001).
6. C. Chow, and C. Liu, *Information Theory, IEEE Transactions on* **14**, 462–467 (1968).
7. H. H. Ku, and S. Kullback, *IEEE Transactions on Information Theory* **15**, 444–447 (1969).
8. D. L. Keefer, *Engineering Management, IEEE Transactions on* **51**, 173–182 (2004).
9. A. E. Abbas, *Engineering Management, IEEE Transactions on* **53**, 146–159 (2006).
10. A. E. Abbas, "Entropy methods for univariate distributions in decision analysis," in *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 22nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, AIP Publishing, 2003, vol. 659, pp. 339–349.
11. A. E. Abbas, "An entropy approach for utility assignment in decision analysis," in *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: 22nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, AIP Publishing, 2003, vol. 659, pp. 328–338.
12. A. E. Abbas, *Operations Research* **54**, 277–290 (2006).
13. A. E. Abbas, *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* **34**, 169–178 (2004).
14. M. Rao, Y. Chen, B. C. Vemuri, and F. Wang, *Information Theory, IEEE Transactions on* **50**, 1220–1228 (2004).
15. S. Kullback, and R. A. Leibler, *The Annals of Mathematical Statistics* pp. 79–86 (1951).
16. T. M. Cover, Ja thomas elements of information theory (1991).
17. S. Baratpour, and A. H. Rad, *Communications in Statistics-Theory and Methods* **41**, 1387–1396 (2012).
18. F. Wang, B. C. Vemuri, M. Rao, and Y. Chen, "A new & robust information theoretic measure and its application to image alignment," in *Information Processing in Medical Imaging*, Springer, 2003, pp. 388–400.
19. C. E. Shannon, *ACM SIGMOBILE Mobile Computing and Communications Review* **5**, 3–55 (2001).