A Study of Recently Discovered Equalities about Latent Tree Models Using Inverse Edges

Nevin L. Zhang[1], Xiaofei Wang[2], and Peixian Chen[1]

[1] Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong
{lzhang,pchenac}@cse.ust.hk
[2] School of Mathematics and Statistics, Northeast Normal University, China
wangxf341@nenu.edu.cn

Abstract. Interesting equalities have recently been discovered about latent tree models. They relate distributions of two or three observed variables with joint distributions of four or more observed variables, and with model parameters that depend on latent variables. The equations are derived by using matrix and tensor decompositions. This paper sheds new light on the equalities by offering an alternative derivation in terms of variable elimination and structure manipulations. The key technique is the introduction of inverse edges.

Keywords: Matrix decomposition, parameter estimation, latent tree models.

1 Introduction

The Expectation-Maximization (EM) algorithm, introduced by Dempster et al. [4], is commonly used to estimate parameters of latent variable models. A well-known drawback of EM is that it tends to get trapped in local optima. New estimation techniques have recently been developed that do not share the shortcoming [7,5,2]. Those techniques are suitable for a class of latent variable models called latent tree models [11,8]. They are based on matrix decompositions and tensor computations.

Anandkumar et al. [2] further developed the techniques and provided a unified framework in terms of low order moments and tensor decompositions. Their works cover a number of latent variable models, including latent tree models, Gaussian mixture models, hidden Markov models, and Latent Dirichlet allocation.

At the heart of the techniques are equalities that related model parameters to quantities that can be directly estimated from data. Equalities of similar flavor are discovered by Parikh et al. [10] that relate joint distributions of four or more observed variables in a latent tree model with distributions of two or three observed variables. Those equalities allows one to estimate the joint probability of a particular assignment of all observed variables without estimating the model parameters.

L.C. van der Gaag and A.J. Feelders (Eds.): PGM 2014, LNAI 8754, pp. 567–580, 2014.

In this paper, we study the equalities in the context of latent tree models. We augment latent tree models with what we call inverse edges. The equalities are then derived by eliminating variables from the augmented models according to different orders. The derivations are insightful and give intuitive explanations as why the equalities hold.

The rest of this paper is organized as follows. In Section 2 we introduce preliminary concepts and notations, and in Section 3 we introduce the key concept of inverse edges. Equalities for joint probability estimation are derived in Section 4, and equalities for parameter estimation are derived in Section 5. Conclusions are provided in Section 6.

2 Preliminaries

We start by introducing several technical concepts.

2.1 Markov Random Fields

A Markov random field (MRF) over a set of discrete variables is defined by a list of potentials. Each potential is a non-negative function of some of the variables. The product of the potentials is the joint distribution of all the variables.

An example MRF is shown in Figure 1. There are four potentials $\phi_1(A, B)$, $\phi_2(B, C, D)$, $\phi_3(D, E)$ and $\phi_4(E, F)$. The figure shows the structure of the MRF. The edge between A and B indicates that there is a potential for them; the hyperedge consisting of B, C, and D indicates the same for those three variables; and so on.

2.2 Latent Tree Model

A *latent tree model (LTM)* is a tree structured MRF where the leaf nodes represent observed variables, while the internal nodes represent latent variables [11]. An example is shown in Figure 2. The variables A, B, C and D are observed, while H and G are latent. There are multiple ways to specify parameters for the LTM. One way is to give: $P(A|H)$, $P(B|H)$, $P(H, G)$, $P(C|G)$ and $P(D|G)$.

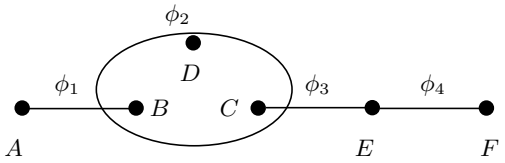


Fig. 1. An example on Markov random fields

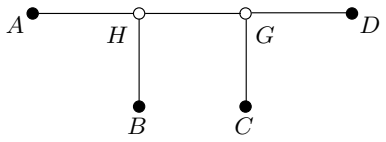


Fig. 2. An example on latent trees

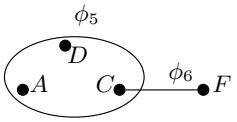


Fig. 3. Result of eliminating B and E from Figure 1

2.3 Variable Elimination in MRF

To *eliminate a variable* X from an MRF means to remove all the potentials that involve X, compute their product, marginalize out X from the product, and add the result to the MRF as a new potential [12].

Consider eliminating the variable B from the MRF shown in Figure 1. The potentials ϕ_1 and ϕ_2 are first removed and the following new potential is created:

$$\phi_5(A, C, D) = \sum_B \phi_1(A, B)\phi_2(B, C, D). \tag{1}$$

If we further eliminate E, then the potentials ϕ_3 and ϕ_4 are also removed and the following new potential is created:

$$\phi_6(D, F) = \sum_E \phi_3(D, E)\phi_4(E, F). \tag{2}$$

The new model is shown in Figure 3. The new structure is obtained from the old one by deleting the edges that involve B and E, and creating two new edges that consist of the neighbors of B and E respectively.

2.4 Matrix Representation of Potentials

For any variable X, use $|X|$ to denote its *cardinality*, i.e., the number of possible values. Denote the values of X as 1, 2, ..., $|X|$. A generic value of X will be referred to using the lower case letter x.

A potential of two variables can be represented using a matrix. Take $\phi_1(A, B)$ for example. The matrix representation of $\phi_2(A, B)$ is a $|A| \times |B|$ matrix. The value at the a-th row and b-th column is $\phi_2(A{=}a, B{=}b)$. We denote this matrix as ϕ_{AB}, and the value $\phi_2(A{=}a, B{=}b)$ as ϕ_{ab}.

Matrix representation allows us to write the results of variable elimination as matrix multiplications. Let ϕ_{DE}, ϕ_{EF}, ϕ_{DF} be the matrix representations of

$\phi_3(D, E)$, $\phi_4(E, F)$ and $\phi_6(D, F)$ respectively. Using matrix multiplication, we can write Equation (2) simply as

$$\phi_{DF} = \phi_{DE}\phi_{EF}. \tag{3}$$

For particular values a, b, c and d of the corresponding variables, denote the values $\phi_5(A{=}a, C{=}c, D{=}d)$ and $\phi_2(B{=}b, C{=}c, D{=}d)$ as ϕ_{acd} and ϕ_{bcd} respectively. Use ϕ_{Bcd} to denote the column vector where the value on the b-th row is ϕ_{bcd}, and use ϕ_{aB} to denote the column vector where the value on the b-th row is ϕ_{ab}. Those notations allow us to rewrite Equation (1) as:

$$\phi_{acd} = \phi_{aB}^T \phi_{Bcd}, \tag{4}$$

for any given value a, c and d of A, C and D,

For the probability distributions $P(A|H)$, $P(B|H)$, $P(H, G)$, $P(C|G)$ and $P(D|G)$ of the LTM shown in Figure 2, their matrix representations are denoted as $P_{A|H}$, $P_{B|H}$, P_{HG}, $P_{C|G}$ and $P_{D|G}$ respectively.

3 Identity and Inverse Edges

In an MRF, potentials are non-negative. In this paper, for introducing inverse matrices or inverse edges into the potential operations, we generalize the concept by allowing potentials to take negative values. This results in *generalized MRF*. In a generalized MRF, the product of all potentials is a *joint potential* on all the variables. It is not necessarily a probability distribution. *Marginal potentials* can be obtained from the joint through marginalization. Variable elimination and the corresponding manipulations with model structure are the same as in the case of MRF. LTMs are MRFs, and hence are generalized MRFs.

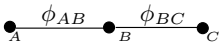
3.1 Identity Edges

For the rest of this section, we consider only generalized MRFs with tree structures. Let X and X' be two neighboring variables that have equal cardinality. Suppose the matrix representation $\phi_{XX'}$ of the potential $\phi(X, X')$ is an identity matrix I. Then we say that the edge (X, X') is an *identity edge*. The potential matrix is written as $I_{XX'}$.

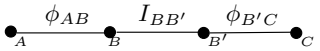
Let X be a variable with two or more neighbors. *Splitting X into an identity edge* means to: (1) create another variable X' such that $|X'| = |X|$, (2) divide the neighbors of X between X' and X, (3) connect X' and X and make it an identity edge.

In the model of Figure 4 (a), splitting the variable B results in the model of (b). The identity edge (B, B') is introduced. The potential matrix $\phi_{B'C}$ equals ϕ_{BC}. It is the same matrix, except the rows are indexed by values of B', not of B.

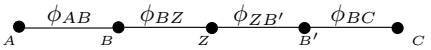
The following theorem is obvious.



(a) A simple MRF.



(b) Obtained from (a) by splitting variable B into an identity edge.



(c) Obtained from (b) by replacing the identity edge (B, B') with a pairs of edges (B, Z) and (Z, B') that are inverse of each others.

Fig. 4. Illustration of identity and inverse edge introduction

Theorem 1. *Let m be a generalized MRF with a tree structure, and X be a variable in m and $\mathbf{Y}$ be a subset of other variables. Suppose X has two or more neighbors. Let m' be a new model obtained from m by splitting X into an identity edge. Then the marginal potential of $\mathbf{Y}$ in the model m equals that in m'.*

Another way to state the theorem is that the elimination of X' from m' results in the model m.

3.2 Inverse Edges

Continue with the example shown in Figure 4 (b). Suppose there is another variable Z such that $|Z| \geq |B|$. Let ϕ_{BZ} be matrix representation for a potential of the two variables. Construct a new model by: (a) inserting Z between B and B', (2) setting the potential matrix of (B, Z) to be ϕ_{BZ}, and (3) setting the potential matrix $\phi_{ZB'}$ of (Z, B') to satisfy that $I_{BB'} = \phi_{BZ}\phi_{ZB'}$. This is well-defined because $|B'| = |B|$. The new model is shown in Figure 4 (c).

It is obvious that $\phi_{BZ} = \phi_{ZB'}^{-1}$ when $|Z| = |B|$. So, we say that (Z, B') is the *inverse edge* of (B, Z) and vice versa. Because of $I_{BB'} = \phi_{BZ}\phi_{ZB'}$, eliminating Z from (c) gives us (b).

In general, we have the following theorem.

Theorem 2. *Suppose (X, X') is an identity edge in a generalized MRF m whose model structure is a tree. Let $\mathbf{Y}$ be a subset of other variables. Constructing a new model m' by replacing the edge (X, X') with a pair of edges that are inverse of each other. Then the marginal potential of $\mathbf{Y}$ in the model m equals that in m'.*

Two remarks are in order. First, the matrix $\phi_{ZB'}$ might contain negative values even though the matrix ϕ_{BZ} does not. This is why we need to generalize the concept of MRF. Second, when $|Z| = |B|$, the notation $\phi_{ZB'}$ in Figure 4 (c)

can be replaced by ϕ_{BZ}^{-1}, with the understanding the columns of the matrix are indexed using values of B'.

4 Equalities for Joint Probability Estimation

Parikh *et al.* [10] have recently discovered equations about LTMs that relate distributions of four or more observed variables to distributions of two or three observed variables. Those equations enable the estimation of the joint probability of particular value assignments of all observed variables without having to estimate the model parameters. The equations were derived using matrix decomposition and tensor computation. In this section, we give an alternative derivation using inverse edges.

4.1 Quartet Trees

We start with model shown in Figure 2. It is called a *quartet model* and will be referred to as M1. We assume that all the variables have equal cardinality and all the probability matrices have full rank. In the following, we will derive an equation that relates $P(A, B, C, D)$ to $P(A, B, D)$, $P(B, D)$, and $P(B, C, D)$.

Starting from M1, we split the two latent nodes H and G into two identity edges (H', H) and (G, G'), resulting in the model M2 of Figure 5 (a). Next, we replace the edge (H', H) with a pairs of edges (H', D') and (D', H), where D' is new variable such that $|D'| = |D|$. The potential matrices for the two edges are set as: $\phi_{H'D'} = P_{HD}$ and $\phi_{D'H} = P_{HD}^{-1}$. Here P_{HD} is the matrix representation of $P(H, D)$. It is invertible because all the probability matrices in M1 have full rank. So, the two edges (H', D') and (D', H) are inverse of each other. Similarly, we replace the edge (G, G') with two edges (G, B') and (B', G') that are inverse of each other. The resulting model $M3$ is shown in Figure 5 (b). According to Theorems 1 and 2, the joint distribution $P(A, B, C, D)$ in M3 is the same as that in M1. In other words, eliminating D', B', H' and G' from $M3$ yields M1. Note that further eliminating H and G in M1 gives us $P(A, B, C, D)$.

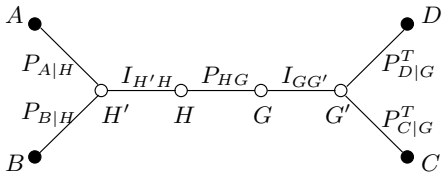
Another way to compute $P(A, B, C, D)$ in M3 is to eliminate the variables in the following order: H', G', H, G, D' and B'. The model M4 shown in Figure 5 (c) is that we obtain after eliminating the first four variables. In the following, we explain how the result is obtained.

The elimination of H' involves three potentials. In function form, they are $P(A|H')$, $P(B|H')$, and $P(H', D')$. The elimination of H' gives us the potential

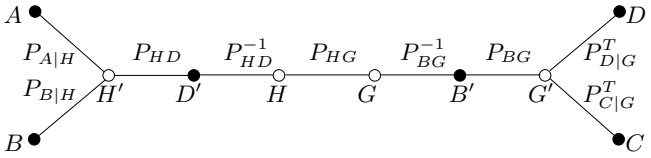
$$\sum_{H'} P(A|H')P(B|H')P(H', D')P(A, B, D').$$

Similarly, the elimination of G' gives us $P(B', C, D)$.

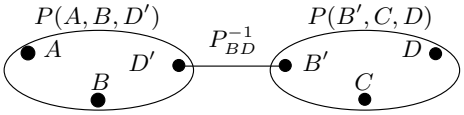
The elimination of H and G also involves three potentials. In matrix form, they are $\phi_{D'H} = P_{HD}^{-1}$, P_{HG} and $\phi_{GB'} = P_{BG}^{-1}$. Note that in M1 we have $P_{HD} = P_{HG}P_{D|G}^{T}$ and $P_{BG} = P_{B|H}P_{HG}$. So, the elimination of H and G gives us the following potential:



(a) M2 — Result of splitting H and G in the model of Figure 2.



(b) M3 — Results of replacing the identity edges in M2 with pairs edges that are inverse of each other.



(c) M4 — Result of eliminating H', G', H and G from M3. Two of the potentials are given in function form, while the third in matrix form.

Fig. 5. Transforms applied on the model of Figure 2

$$\begin{aligned}\phi_{D'H} P_{HG} \phi_{GB'} &= P^{-1}_{HD} P_{HG} P^{-1}_{BG} \\ &= (P^T_{D|G})^{-1} P^{-1}_{HG} P_{HG} P^{-1}_{HG} P^{-1}_{B|H} \\ &= (P^T_{D|G})^{-1} P^{-1}_{HG} (P_{B|H})^{-1} \\ &= (P_{B|H} P_{HG} P^T_{D|G})^{-1} = P^{-1}_{BD}.\end{aligned}$$

This is the matrix representation of the potential for the edge (D', B'), i.e., $\phi_{D'B'} = P^{-1}_{BD}$.

Finally, the elimination of D' and B' in M4 gives us the distribution $P(A, B, C, D)$. For specific values a, b, c and d of the corresponding variables, denote the probability $P(A{=}a, B{=}b, C{=}c, D{=}d)$ as P_{abcd}. It is clear that

$$P_{abcd} = P^T_{abD} P^{-1}_{BD} P_{Bcd}, \tag{5}$$

where P_{abD} and P_{Bcd} are column vectors obtained from the joint distributions $P(A, B, D)$ and $P(B, C, D)$ in the way described in Section 2.4.

The following theorem summarizes the foregoing derivations, which construct inverse edges and potentials for variable elimination:

Theorem 3. *In the quartet model of Figure 2, suppose all the variables have equal cardinality and all probability matrices have full rank. Then the distribution*

$P(A, B, C, D)$ can be computed from $P(A, B, D)$, $P(B, D)$, $P(B, C, D)$ using Equation (5).

4.2 Observed Variables with Unequal Cardinalities

Next we generalize Theorem 3 to the case where the observed variable might have unequal cardinalities. The latent variables are still required to have equal cardinality and it must be no greater than the cardinality of any observed variable. We further require that the probability matrices $P_{A|H}$, $P_{B|H}$, P_{HG}, $P_{C|G}$ and $P_{D|G}$ have full column rank.

The technical issue to deal with in this case is that the matrices P_{HD}, P_{BG} and P_{BD} might not be invertible.

Let the cardinality of H and G be r, and those of B and D be s and t respectively. In model M1 we have

$$P_{BD} = P_{B|H} P_{HG} P_{D|G}^T.$$

Because all the matrices on the right hand side have full column rank, the rank of P_{BD} is r. Consider the singular decomposition of $P_{BD} = U \Lambda V^T$, where Λ is a $r \times r$ diagonal matrix, U and V are $s \times r$ and $t \times r$ column orthogonal matrices respectively. We have

$$\begin{aligned} P_{BD} &= P_{B|H} P_{HG} P_{D|G}^T \\ &= P_{B|H} P_{HG} P_{HG}^{-1} P_{HG} P_{D|G}^T \\ &= P_{BG} P_{HG}^{-1} P_{HD}. \end{aligned} \tag{6}$$

Consequently,

$$\begin{aligned} P_{BG} P_{HG}^{-1} P_{HD} &= U \Lambda V^T, \\ U^T P_{BG} P_{HG}^{-1} P_{HD} V &= \Lambda. \end{aligned}$$

This implies that $U^T P_{BG}$ and $P_{HD} V$ are invertible.

Construct the model M3 as in the previous subsection, except that we set the potential matrices for the edges (D', H) and (G, B') as follows:

$$\phi_{D'H} = V(P_{HD} V)^{-1}, \phi_{GB'} = (U^T P_{BG})^{-1} U^T.$$

It is clear that the product of P_{HD} and $V(P_{HD} V)^{-1}$ is an identity matrix. So, the edge (D', H) is the inverse edge of (H', D'). It is also clear that the product of $(U^T P_{BG})^{-1} U^T$ and P_{BG} is an identity matrix. So, the edge (G, B') is the inverse edge of (B', G').

When we move from M3 to M4, everything is the same as in the previous subsection, except that the elimination of H and G now involves different potential matrices. The result is

$$\begin{aligned} \phi_{D'H} P_{HG} \phi_{GB'} &= V(P_{HD} V)^{-1} P_{HG} (U^T P_{BG})^{-1} U^T \\ &= V(U^T P_{BG} P_{HG}^{-1} P_{HD} V)^{-1} U^T \\ &= V(U^T P_{BD} V)^{-1} U^T, \end{aligned}$$

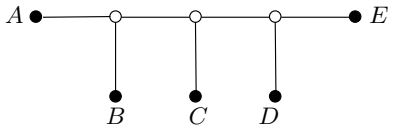


Fig. 6. A general latent tree model

where the last equality is due to Equation 6. This is the potential matrix for the edge (D', B').

Consequently, for any specific values a, b, c and d of the variables, we have

$$P_{abcd} = P_{abD}^T V (U^T P_{BD} V)^{-1} U^T P_{Bcd}. \tag{7}$$

Theorem 4. *In the quartet model of Figure 2, suppose the conditions specified in the first paragraph of this subsection hold. Then the distribution $P(A, B, C, D)$ can be computed from $P(A, B, D)$, $P(B, D)$, $P(B, C, D)$ using Equation (7), where U and V are from the singular decomposition of P_{BD}.*

Note that Equation (7) is same as Equation (5) except that the matrices U and V are used to deal with the issue that P_{BD} might not be invertible.

4.3 General Trees

To apply Theorems 3 and 4 in general trees, we divide all the variables into four groups S_1, S_2, S_3 and S_4 such that, when each group is viewed as a joint variable, the relationship among them is a quartet tree. By applying the theorems, we can compute $P(S_1, S_2, S_3, S_4)$ to from $P(S_1, S_2, S_4)$, $P(S_2, S_4)$, $P(S_2, S_3, S_4)$. In the process, we need to invert the matrix representation of $P(S_2, S_4)$. For computational efficiency, S_2 and S_4 should be singletons. The same strategy can then be repeated on $P(S_1, S_2, S_4)$ and $P(S_2, S_3, S_4)$ until all the distributions needed for the computation are for no more than 3 variables. Such distributions are directly estimated from data.

Let us illustrate the strategy using the model shown in Figure 6. First, partition the variables into four groups as follows: $S_1 = \{A\}$, $S_2 = \{B\}$, $S_3 = \{C, D\}$ and $S_4 = \{E\}$. This allows us to reduce $P(A, B, C, D, E)$ to $P(A, B, E)$, $P(B, E)$, and $P(B, C, D, E)$. The first two distributions involve no more than three variables and are directly estimated from data. To compute the last distribution, consider a restriction of the model onto the four variables involved. Let $S_1 = \{B\}$, $S_2 = \{C\}$, $S_3 = \{D\}$ and $S_4 = \{E\}$. This allows us to reduce $P(B, C, D, E)$ to $P(B, C, E)$, $P(C, D)$, and $P(C, D, E)$. All those distributions are estimated from data.

The description of a complete algorithm and the discussion of related issues are out the scope of this paper.

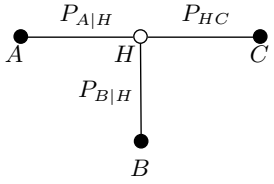


Fig. 7. A latent tree model with one latent variable

5 Equalities for Parameter Estimation

In this section, we derive a group of equations that are used to estimate parameters of LTMs [2].

5.1 Some Notations

Consider the LTM shown in Figure 7. Suppose all variables have equal cardinality. We parameterize the model with the distributions $P(A|H)$, $P(B|H)$ and $P(H, C)$. In matrix notation, they are $P_{A|H}$, $P_{B|H}$ and P_{HC}.

Let b be a particular value of B. For reasons to become clear later, we consider the joint probability $P(A{=}a, B{=}b, C{=}c)$ when variables $A = a$ and $C = c$. It is obvious that

$$P(A{=}a, B{=}b, C{=}c) = \sum_{H} P(A{=}a|H)P(B{=}b|H)P(H, C{=}c). \tag{8}$$

Use P_{AbC} to denote the matrix where the element at the a-th row and c-th column is P_{abc}. Use $P_{b|H}$ to denote the column vector where the element at the h-row is $P(B{=}b|H{=}h)$, which in turn is denoted as $P_{b|h}$. Use $diag(P_{b|H})$ to denote the diagonal matrix with the elements of the vector $P_{b|H}$ as the diagonal elements. Equation (8) can be rewritten in matrix form as follows:

$$P_{AbC} = P_{A|H} diag(P_{b|H}) P_{HC}. \tag{9}$$

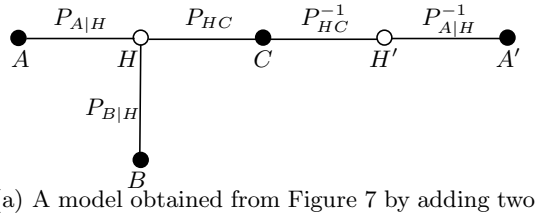
5.2 The Case of Equal Cardinality

Suppose all the probability matrices are invertible. Augment the model of Figure 7 with two edges (C, H') and (H', A'), where $|A'| = |A|$ and $|H'| = |H|$. Let the potential matrices for the two new edges be:

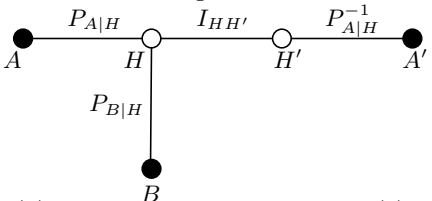
$$\phi_{CH'} = P_{HC}^{-1}, \phi_{H'A'} = P_{A|H}^{-1}.$$

Note that the edge (C, H') is the inverse edge of (H, C).

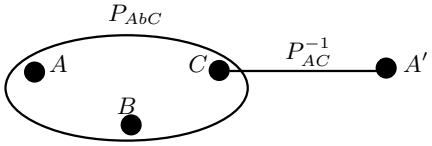
For the rest of this subsection, fix the value of B at b. Consider calculating the marginal potential $P(A, B{=}b, A')$ in the model of Figure 8 (a). In matrix form, it is $P_{AbA'}$.



(a) A model obtained from Figure 7 by adding two edges.



(b) Result of eliminating C from (a).



(c) Result of eliminating H and H' from (a).

Fig. 8. Operations on a model obtained by augmenting the model of Figure 7

One way to compute $P_{AbA'}$ is to first eliminate C and then eliminate H' and H. Eliminating C from the model of (a) results in the model of (b). Because the edge (C, H') is the inverse edge of (H, C), the potential for the edge (H, H') is an identity matrix $I_{HH'}$. Further eliminating H' results in a model that has the structure shown in Figure 7 with the variable C replaced with A' (ref. Theorem 1). The potential matrix for the edge (H, A') is $P_{A|H}^{-1}$. According to Equation (9), we have

$$P_{AbA'} = P_{A|H} diag(P_{b|H}) P_{A|H}^{-1}. \tag{10}$$

Another way to compute $P_{AbA'}$ is to first eliminate H and H', and then eliminate C. The elimination of H and H' leads to the model of (c). According to Equation (9), the elimination of H gives us the potential matrix P_{AbC}. The elimination of H' gives us the following potential matrix:

$$\phi_{CH'}\phi_{H'A'} = P_{HC}^{-1} P_{A|H}^{-1} = (P_{A|H} P_{HC})^{-1} = P_{AC}^{-1}.$$

Further eliminating C from the model of (c), we get

$$P_{AbA'} = P_{AbC} P_{AC}^{-1}. \tag{11}$$

Putting the two equations (10) and (11) together, we get the following theorem.

Theorem 5. *Suppose that all the variables in the LTM shown Figure 7 have equal cardinality and that all the probability matrices are invertible. Then,*

$$P_{A|H} diag(P_{b|H}) P_{A|H}^{-1} = P_{AbC} P_{AC}^{-1}. \tag{12}$$

Note that the two terms on the right hand side involve only observed variables. They can be directly estimated from data. On the other hand, the left hand side involves model parameters, and particularly the diagonal elements of $P_{b|H}$ are the eigenvalues of $P_{AbC}P_{AC}^{-1}$.

5.3 The Case of Unequal Cardinalities

Next we generalize Theorem 5 to the case where the observed variables might have unequal cardinalities. The cardinality of H latent variables is required to be no greater than the cardinality of any observed variable. We further require that the probability matrices $P_{A|H}$ and P_{HC}^T have full column rank.

The technical issue to deal with in this case is that the matrices $P_{A|H}$, P_{HC} and P_{AC} might not be invertible.

Let the cardinalities of H, A and C be r, s and t. In the model of Figure 7, we have $P_{AC} = P_{A|H}P_{HC}$. Because $P_{A|H}$ and P_{HC}^T have full column rank, the rank of P_{AC} is r. Consider the singular decomposition of $P_{AC} = U\Lambda V^T$, where Λ is a $r \times r$ diagonal matrix, U and V are $s \times r$ and $t \times r$ column orthogonal matrices respectively. By following the same line of reasoning as in Section 4.2, we can conclude that $U^T P_{A|H}$ and $P_{HC}V$ are invertible.

Note that $(U^T P_{A|H})^{-1}U^T$ is a $r \times s$ matrix, and $V(P_{HC}V)^{-1}$ is a $t \times r$ matrix. Construct the model of Figure 8 (a) in the same way as in the previous subsection, except that we set potential matrices of the edges (H', A') and (C, H') as follows:

$$\phi_{H'A'} = (U^T P_{A|H})^{-1}U^T, \phi_{CH'} = V(P_{HC}V)^{-1}.$$

Now consider the marginal potential matrix $P_{AbA'}$. One way to compute it is to first eliminate C and then eliminate H' and H. Eliminating C from the model of (a) results in the model of (b). Because the product of P_{HC} and $V(P_{HC}V)^{-1}$ is an identity matrix, the potential for the edge (H, H') an identity edge. Further eliminating H' results in a model that has the structure shown in Figure 7 with the variable C replaced with A' (ref. Theorem 1). The potential matrix for the edge (H, A') is $(U^T P_{A|H})^{-1}U^T$. According to Equation (9), we have

$$P_{AbA'} = P_{A|H}diag(P_{b|H})(U^T P_{A|H})^{-1}U^T. \tag{13}$$

Another way to compute $P_{AbA'}$ is to first eliminate H and H', and then eliminate C. The elimination of H and H' lead to the model of (c). As in the previous section, the elimination of H gives us the potential matrix P_{AbC}. The elimination of H' gives us the following potential matrix:

$$\begin{aligned}\phi_{CH'}\phi_{H'A'} &= V(P_{HC}V)^{-1}(U^T P_{A|H})^{-1}U^T \\ &= V(U^T P_{A|H}P_{HC}V)^{-1}U^T \\ &= V(U^T P_{AC}V)^{-1}U^T\end{aligned}$$

Further eliminating C from the model of (c), we get

$$P_{AbA'} = P_{AbC}V(U^T P_{AC}V)^{-1}U^T. \tag{14}$$

Putting the two equations (13) and (14) together, we get

$$P_{A|H}diag(P_{b|H})(U^T P_{A|H})^{-1}U^T = P_{AbC}V(U^T P_{AC}V)^{-1}U^T.$$

Consequently,

$$U^T P_{A|H}diag(P_{b|H})(U^T P_{A|H})^{-1} = U^T P_{AbC}V(U^T P_{AC}V)^{-1}. \quad (15)$$

Theorem 6. *Suppose that, in the LTM shown Figure 7, the cardinality of H is no greater than that of any observed variable, and that the probability matrices $P_{A|H}$ and P_{HC}^T have full column rank. Then, Equation (15) holds.*

Note that Equation (15) is the same as Equation (12) except that the column orthogonal matrices U and V are used to deal the issue that $P_{A|H}$ and P_{AC} might not be invertible.

5.4 Parameter Estimation

To use Equation (15) for parameter estimation, observe that the eigenvalues of the matrix on the right hand side are the diagonal elements of $diag(P_{b|H})$, which in turn are the values for the conditional distribution $P(B{=}b|H)$. To estimate $P(B|H)$, we can: (1) estimate P_{ABC} and P_{AC} from data; (2) compute the singular decomposition of the of P_{AC} to obtain the matrices U and V; (3) for each value b for B, form the matrix on the right hand side; and (4) calculate the eigenvalues of the matrix. Those eigenvalues are the values for the distribution $P(B|H)$.

If all the variables have equal cardinality, we can use Equation (12) instead. In this case, there is no need calculate the matrices U and V.

To see how the strategy can be applied to LTMs with multiple latent variables, consider the model of Figure 2. For simplicity, we assume all the variables have equal cardinality. Restricting the model to the variable A, B, C and H, we get the model of Figure 7. Using the strategy, we can estimate $P(A|H)$, $P(B|H)$ and $P(C|H)$. In similar fashion, we can estimate $P(C|G)$, $P(D|G)$ and $P(A|G)$. Then $P(G|H)$ can be calculated from $P(H|C)$ and $P(C|G)$ using the relationship $P_{G|H} = P_{C|G}^{-1}P_{C|H}$.

The description of a complete algorithm and the discussion of related issues are out the scope of this paper.

6 Conclusions

Starting from a latent tree model, we introduce inverse edges to obtain a generalized MRFs. Variables are then eliminated from the generalized MRFs in different orders to obtain equalities about the model. Two groups of equalities are

obtained. One group allows us to calculate the joint probability one of particular assignment of all observed variables without estimating the model parameters. The other group of equalities gives us a new method for estimating the model parameters, which has advantages over the commonly used EM algorithm. For example, it does not have the difficulty of getting trapped in local maxima.

References

1. Anandkumar, A., Ge, R., Hsu, D., Kakade, S.M., Telgarsky, M.: Tensor Decompositions for Learning Latent Variable Models (2012) (Preprint)
2. Anandkumar, A., Hsu, D., Kakade, S.M.: A Method of Moments for Mixture Models and Hidden Markov Models. In: An Abridged Version Appears in the Proc. of COLT (2012)
3. Bartholomew, D.J., Knott, M., Moustaki, I.: Latent variable models and factor analysis, 2nd edn. Wiley (1999)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38 (1977)
5. Hsu, D., Kakade, S., Zhang, T.: A spectral algorithm for learning hidden markov models. In: COLT (2009)
6. Lauritzen, S.L.: Graphical Models. Clarendon Press (1996)
7. Mossel, E., Roch, S.: Learning nonsingular phylogenies and hidden markov models. AOAP 16, 583–614 (2006)
8. Mourad, R., Sinoquet, C., Zhang, N.L., Liu, T.F., Leray, P.: A survey on latent tree models and applications. Journal of Artificial Intelligence Research 47, 157–203 (2013)
9. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT Press (2012)
10. Parikh, A.P., Song, L., Xing, E.P.: A Spectral Algorithm for Latent Tree Graphical Models. In: ICML (2011)
11. Zhang, N.L.: Hierarchical latent class models for cluster analysis. Journal of Machine Learning Research 5, 697–723 (2004)
12. Zhang, N.L., Poole, D.: Exploiting causal independence in Bayesian network inference. Journal of Artificial Intelligence Research 5, 301–328 (1996)