Spectral Methods for Learning Multivariate Latent Tree Structure

Authored by:

Tong Zhang
Sham M. Kakade
Le Song
Daniel J. Hsu
Kamalika Chaudhuri
Animashree Anandkumar

Abstract

This work considers the problem of learning the structure of multivariate linear tree models, which include a variety of directed tree graphical models with continuous, discrete, and mixed latent variables such as linear-Gaussian models, hidden Markov models, Gaussian mixture models, and Markov evolutionary trees. The setting is one where we only have samples from certain observed variables in the tree, and our goal is to estimate the tree structure (i.e., the graph of how the underlying hidden variables are connected to each other and to the observed variables). We propose the Spectral Recursive Grouping algorithm, an efficient and simple bottom-up procedure for recovering the tree structure from independent samples of the observed variables. Our finite sample size bounds for exact recovery of the tree structure reveal certain natural dependencies on underlying statistical and structural properties of the underlying joint distribution. Furthermore, our sample complexity guarantees have no explicit dependence on the dimensionality of the observed variables, making the algorithm applicable to many high-dimensional settings. At the heart of our algorithm is a spectral quartet test for determining the relative topology of a quartet of variables from second-order statistics.

1 Paper Body

Graphical models are a central tool in modern machine learning applications, as they provide a natural methodology for succinctly representing high-dimensional distributions. As such, they have enjoyed much success in various AI and machine learning applications such as natural language processing, speech recognition, robotics, computer vision, and bioinformatics. The main statistical challenges associated with graphical models include estimation and inference. While

the body of techniques for probabilistic inference in graphical models is rather rich [1], current methods for tackling the more challenging problems of parameter and structure estimation are less developed and understood, especially in the presence of latent (hidden) variables. The problem of parameter estimation involves determining the model parameters from samples of certain observed variables. Here, the predominant approach is the expectation maximization (EM) algorithm, and only rather recently is the understanding of this algorithm improving [2, 3]. The problem of structure learning is to estimate the underlying graph of the graphical model. In general, structure learning is NP-hard and becomes even more challenging when some variables are unobserved [4]. The main approaches for structure estimation are either greedy or local search approaches [5, 6] or, more recently, based on convex relaxation [7]. 1

z1 z2
z3 h g
z4
{{z1 , z2 }, {z3 , z4 }} (a)
z1 z3
z2 h g
z1
z4
z4
{{z1 , z3 }, {z2 , z4 }} (b)
z2 h g
z3
{{z1 , z4 }, {z2 , z3 }} (c)
z1 z4
z2 h
z3
{{z1 , z2 , z3 , z4 }} (d)

Figure 1: The four possible (undirected) tree topologies over leaves {z1 , z2 , z3 , z4 }. This work focuses on learning the structure of multivariate latent tree graphical models. Here, the underlying graph is a directed tree (e.g., hidden Markov model, binary evolutionary tree), and only samples from a set of (multivariate) observed variables (the leaves of the tree) are available for learning the structure. Latent tree graphical models are relevant in many applications, ranging from computer vision, where one may learn object/scene structure from the co-occurrences of objects to aid image understanding [8]; to phylogenetics, where the central task is to reconstruct the tree of life from the genetic material of surviving species [9]. Generally speaking, methods for learning latent tree structure exploit structural properties afforded by the tree that are revealed through certain statistical tests over every choice of four variables in the tree. These quartet tests, which have origins in structural equation modeling [10, 11], are hypothesis tests of the relative configuration of four (possibly non-adjacent) nodes/variables in the tree (see Figure 1); they are also related to the four point condition associated with a corresponding additive tree metric induced by the distribution [12]. Some early methods for learning tree structure are

based on the use of exact correlation statistics or distance measurements (e.g., [13, 14]). Unfortunately, these methods ignore the crucial aspect of estimation error, which ultimately governs their sample complexity. Indeed, this (lack of) robustness to estimation error has been quantified for various algorithms (notably, for the popular Neighbor Joining algorithm [15, 16]), and therefore serves as a basis for comparing different methods. Subsequent work in the area of mathematical phylogenetics has focused on the sample complexity of evolutionary tree reconstruction [17, 15, 18, 19]. The basic model there corresponds to a directed tree over discrete random variables, and much of the recent effort deals exclusively in the regime for a certain model parameter (the Kesten-Stigum regime [20]) that allows for a sample complexity that is polylogarithmic in the number of leaves, as opposed to polynomial [18, 19]. Finally, recent work in machine learning has developed structure learning methods for latent tree graphical models that extend beyond the discrete distributions of evolutionary trees [21], thereby widening their applicability to other problem domains. This work extends beyond previous studies, which have focused on latent tree models with either discrete or scalar Gaussian variables, by directly addressing the multivariate setting where hidden and observed nodes may be random vectors rather than scalars. The generality of our techniques allows us to handle a much wider class of distributions than before, both in terms of the conditional independence properties imposed by the models (i.e., the random vector associated with a node need not follow a distribution that corresponds to a tree model), as well as other characteristics of the node distributions (e.g., some nodes in the tree could have discrete state spaces and others continuous, as in a Gaussian mixture model). We propose the Spectral Recursive Grouping algorithm for learning multivariate latent tree structure. The algorithm has at its core a multivariate spectral quartet test, which extends the classical quartet tests for scalar variables by applying spectral techniques from multivariate statistics (specifically canonical correlation analysis [22, 23]). Spectral methods have enjoyed recent success in the context of parameter estimation [24, 25, 26, 27]; our work shows that they are also useful for structure learning. We use the spectral quartet test in a simple modification of the recursive grouping algorithm of [21] to perform the tree reconstruction. The algorithm is essentially a robust method for reasoning about the results of quartet tests (viewed simply as hypothesis tests); the tests either confirm or reject hypotheses about the relative topology over quartets of variables. By carefully choosing which tests to consider and properly interpreting their results, the algorithm is able to recover the correct latent tree structure (with high probability) in a provably efficient manner, in terms of both computational and sample complexity. The recursive grouping procedure is similar to the short quartet method from phylogenetics [15], which also guarantees efficient reconstruction in the context of evolutionary trees. However, our method and analysis applies to considerably more general high-dimensional settings; for instance, our sample complexity bound is given in terms of natural correlation con2

ditions that generalize the more restrictive effective depth conditions of previous works [15, 21]. Finally, we note that while we do not directly address the

question of parameter estimation, provable parameter estimation methods may derived using the spectral techniques from [24, 25].

2 2.1

Preliminaries Latent variable tree models

Let T be a connected, directed tree graphical model with leaves Vobs := {x1 , x2 , . . . , xn } and internal nodes Vhid := {h1 , h2 , . . . , hm } such that every node has at most one parent. The leaves are termed the observed variables and the internal nodes hidden variables. Note that all nodes in this work generally correspond to multivariate random vectors; we will abuse terminology and still refer to these random vectors as random variables. For any h ? Vhid , let ChildrenT (h) ? VT denote the children of h in T. Each observed variable x ? Vobs is modeled as random vector in Rd , and each hidden variable h ? Vhid as a random vector in Rk . The joint distribution over all the variables VT := Vobs ? Vhid is assumed satisfy conditional independence properties specified by the tree structure over the variables. Specifically, for any disjoint subsets V1 , V2 , V3 ? VT such that V3 separates V1 from V2 in T, the variables in V1 are conditionally independent of those in V2 given V3 . 2.2

Structural and distributional assumptions

The class of models considered are specified by the following structural and distributional assumptions. Condition 1 (Linear conditional means). Fix any hidden variable h ? Vhid . For each hidden child g ? ChildrenT (h) ? Vhid , there exists a matrix A(g—h) ? Rk?k such that E[g—h] = A(g—h) h; and for each observed child x ? ChildrenT (h) ? Vobs , there exists a matrix C(x—h) ? Rd?k such that E[x—h] = C(x—h) h. We refer to the class of tree graphical models satisfying Condition 1 as linear tree models. Such models include a variety of continuous and discrete tree distributions (as well as hybrid combinations of the two, such as Gaussian mixture models) which are widely used in practice. Continuous linear tree models include linear-Gaussian models and Kalman filters. In the discrete case, suppose that the observed variables take on d values, and hidden variables take k values. Then, each variable is represented by a binary vector in {0, 1}s , where s = d for the observed variables and s = k for the hidden variables (in particular, if the variable takes value i, then the corresponding vector is the i-th coordinate vector), and any conditional distribution between the variables is represented by a linear relationship. Thus, discrete linear tree models include discrete hidden Markov models [25] and Markovian evolutionary trees [24]. In addition to the linearity, the following conditions are assumed in order to recover the hidden tree structure. For any matrix M , let ?t (M) denote its t-th largest singular value. Condition 2 (Rank condition). The variables in VT = Vhid ? Vobs obey the following rank conditions. 1. For all h ? Vhid , E[hh?] has rank k (i.e., ?k (E[hh?]) ¿ 0). 2. For all h ? Vhid and hidden child g ? ChildrenT (h) ? Vhid , A(g—h) has rank k. 3. For all h ? Vhid and observed child x ? ChildrenT (h) ? Vobs , C(x—h) has rank k. The rank condition is a generalization of parameter identifiability conditions in latent variable models [28, 24, 25] which rules out various (provably) hard instances in discrete variable settings [24]. 3

h4 h2

x3
h1 x1
T1
h3 x4
T2
x6
T3
x5
x2

Figure 2: Set of trees Fh4 = {T1 , T2 , T3 } obtained if h4 is removed. Condition 3 (Non-redundancy condition). Each hidden variable has at least three neighbors. Furthermore, there exists ?2max ¿ 0 such that for each pair of distinct hidden variables h, g ? Vhid , det(E[hg ?])2 ? ?2max ¡ 1. det(E[hh?]) det(E[gg ?]) The requirement for each hidden node to have three neighbors is natural; otherwise, the hidden node can be eliminated. The quantity ?max is a natural multivariate generalization of correlation. First, note that ?max ? 1, and that if ?max = 1 is achieved with some h and g, then h and g are completely correlated, implying the existence of a deterministic map between hidden nodes h and g; hence simply merging the two nodes into a single node h (or g) resolves this issue. Therefore the non-redundancy condition simply means that any two hidden nodes h and g cannot be further reduced to a single node. Clearly, this condition is necessary for the goal of identifying the correct tree structure, and it is satisfied as soon as h and g have limited correlation in just a single direction. Previous works [13, 29] show that an analogous condition ensures identifiability for general latent tree models (and in fact, the conditions are identical in the Gaussian case). Condition 3 is therefore a generalization of this condition suitable for the multivariate setting. Our learning guarantees also require a correlation condition that generalize the explicit depth conditions considered in the phylogenetics literature [15, 24]. To state this condition, first define Fh to be the set of subtrees of that remain after a hidden variable h ? Vhid is removed from T (see Figure 2). Also, for any subtree T ? of T, let Vobs [T ?] ? Vobs be the observed variables in T ? . Condition 4 (Correlation condition). There exists ?min ¿ 0 such that for all hidden variables h ? Vhid and all triples of subtrees {T1 , T2 , T3 } ? Fh in the forest obtained if h is removed from T, max

min

x1 ?Vobs [T1],x2 ?Vobs [T2],x3 ?Vobs [T3] {i,j}?{1,2,3}

?k (E[xi x? j]) ? ?min .

The quantity ?min is related to the effective depth of T, which is the maximum graph distance between a hidden variable and its closest observed variable [15, 21]. The effective depth is at most logarithmic in the number of variables (as achieved by a complete binary tree), though it can also be a constant if every hidden variable is close to an observed variable (e.g., in a hidden Markov model, the effective depth is 1, even though the true depth, or diameter, is m + 1). If the matrices giving the (conditionally) linear relationship between neighboring variables in T are all well-conditioned, then ?min is at worst exponentially small

in the effective depth, and therefore at worst polynomially small in the number of variables. Finally, also define ?max :=

max
{x1 ,x2 }?Vobs
{?1 (E[x1 x? 2])}

to be the largest spectral norm of any second-moment matrix between observed variables. Note ?max ? 1 in the discrete case, and, in the continuous case, ?max ? 1 if each observed random vector is in isotropic position. In this work, the Euclidean norm of a vector x is denoted by ?x?, and the (induced) spectral norm of a matrix A is denoted by ?A?, i.e., ?A? := ?1 (A) = sup{?Ax? : ?x? = 1}. 4

Algorithm 1 SpectralQuartetTest on observed variables {z1 , z2 , z3 , z4 }. ?i,j of the second-moment matrix Input: For each pair {i, j} ? {1, 2, 3, 4}, an empirical estimate ? ? E[zi zj] and a corresponding confidence parameter ?i,j ¿ 0. Output: Either a pairing {{zi , zj }, {zi? , zj ? }} or ?. 1: if there exists a partition of {z1 , z2 , z3 , z4 } = {zi , zj } ? {zi? , zj ? } such that k ?

s=1
?i,j) ? ?i,j]+ [?s (? ?i? ,j ?) ? ?i? ,j ?]+ ¿ [?s (?
then return the pairing {{zi , zj }, {zi? , zj ? }}.
k ?
?i? ,j) + ?i? ,j)(?s (? ?i,j ?) + ?i,j ?) (?s (?
s=1
2: else return ?.
3

Spectral quartet tests

This section describes the core of our learning algorithm, a spectral quartet test that determines topology of the subtree induced by four observed variables {z1 , z2 , z3 , z4 }. There are four possibilities for the induced subtree, as shown in Figure 1. Our quartet test either returns the correct induced subtree among possibilities in Figure 1(a)?(c); or it outputs ? to indicate abstinence. If the test returns ?, then no guarantees are provided on the induced subtree topology. If it does return a subtree, then the output is guaranteed to be the correct induced subtree (with high probability). The quartet test proposed is described in Algorithm 1 (SpectralQuartetTest). The notation [a]+ denotes max{0, a} and [t] (for an integer t) denotes the set {1, 2, . . . , t}.

The quartet test is defined with respect to four observed variables Z := {z1 , z2 , z3 , z4 }. For each ?i,j of the second-moment matrix pair of variables zi and zj , it takes as input an empirical estimate ? E[zi zj?], and confidence bound parameters ?i,j which are functions of N , the number of samples ?i,j ?s, a confidence parameter ?, and of properties of the distributions of zi and used to compute the ? zj . In practice, one uses a single threshold ? for all pairs, which is tuned by the algorithm. Our theoretical analysis also applies to this case. The output of the test is either ? or a pairing of the variables {{zi , zj }, {zi? , zj ? }}. For example, if the output is the pairing is {{z1 , z2 }, {z3 , z4 }}, then Figure 1(a) is the output topology. Even though the configuration in Figure 1(d) is a possibility, the spectral quartet test never returns {{z1 , z2 , z3

, z4 }}, as there is no correct pairing of Z. The topology {{z1 , z2 , z3 , z4 }} can be viewed as a degenerate case of {{z1 , z2 }, {z3 , z4 }} (say) where the hidden variables h and g are deterministically identical, and Condition 3 fails to hold with respect to h and g. 3.1

Properties of the spectral quartet test

With exact second moments: The spectral quartet test is motivated by the following lemma, which shows the relationship between the singular values of second-moment matrices of the zi ?s and the ?k induced topology among them in the latent tree. Let detk (M) := s=1 ?s (M) denote the product of the k largest singular values of a matrix M . Lemma 1 (Perfect quartet test). Suppose that the observed variables Z = {z1 , z2 , z3 , z4 } have the true induced tree topology shown in Figure 1(a), and the tree model satisfies Condition 1 and Condition 2. Then detk (E[z1 z3?])detk (E[z2 z4?]) detk (E[z1 z4?])detk (E[z2 z3?]) det(E[hg ?])2 = = ?1 det(E[hh?]) det(E[gg ?]) detk (E[z1 z2?])detk (E[z3 z4?]) detk (E[z1 z2?])detk (E[z3 z4?]) (1) and detk (E[z1 z3?])detk (E[z2 z4?]) = detk (E[z1 z4?])detk (E[z2 z3?]). This lemma shows that given the true second-moment matrices and assuming Condition 3, the inequality in (1) becomes strict and thus can be used to deduce the correct topology: the correct pairing is {{zi , zj }, {zi? , zj ? }} if and only if detk (E[zi zj?])detk (E[zi? zj??]) ¿ detk (E[zi? zj?])detk (E[zi zj??]). 5

Reliability: The next lemma shows that even if the singular values of E[zi zj?] are not known exactly, then with valid confidence intervals (that contain these singular values) a robust test can be constructed which is reliable in the following sense: if it does not output ?, then the output topology is indeed the correct topology. Lemma 2 (Reliability). Consider the setup of Lemma 1, and suppose that Figure 1(a) is the ?i,j) ? ?i,j ? correct topology. If for all pairs {zi , zj } ? Z and all s ? [k], ?s (? ? ? ?s (E[zi zj]) ? ?s (?i,j) + ?i,j , and if SpectralQuartetTest returns a pairing {{zi , zj }, {zi? , zj ? }}, then {{zi , zj }, {zi? , zj ? }} = {{z1 , z2 }, {z3 , z4 }}. In other words, the spectral quartet test never returns an incorrect pairing as long as the singular ?i,j . The lemma values of E[zi zj?] lie in an interval of length 2?i,j around the singular values of ? below shows how to set the ?i,j s as a function of N , ? and properties of the distributions of zi and zj so that this required event holds with probability at least 1 ? ?. We remark that any valid confidence intervals may be used; the one described below is particularly suitable when the observed variables are high-dimensional random vectors. Lemma 3 (Confidence intervals). Let Z = {z1 , z2 , z3 , z4 } be four random vectors. Let ?zi ? ? Mi ?i,j is computed using almost surely, and let ? ? (0, 1/6). If each empirical second-moment matrix ? N iid copies of zi and zj , and if E[?zi ?2 ?zj ?2] ? tr(E[zi zj?]E[zi zj?]?)

, ti,j := 1.55 ln(24d?i,j /?), max{?E[?zj ?2 zi zi?]?, ?E[?zi ?2 zj zj?]?} ? ? ? ?? ?? 2 max ?E[?zj ?2 zi zi?]?, ?E[?zi ?2 zj zj?]? ti,j Mi Mj ti,j ?i,j ? + , N 3N then with probability 1 ? ?, for all pairs {zi , zj } ? Z and all s ? [k], ?i,j) ? ?i,j ? ?s (E[zi z ?]) ? ?s (? ?i,j) + ?i,j . ?s (? d?i,j :=

j

(2)

Conditions for returning a correct pairing: The conditions under which Spec-

tralQuartetTest returns an induced topology (as opposed to ?) are now provided.

An important quantity in this analysis is the level of non-redundancy between the hidden variables h and g. Let det(E[hg ?])2 ?2 := . (3) det(E[hh?]) det(E[gg ?]) If Figure 1(a) is the correct induced topology among {z1 , z2 , z3 , z4 }, then the smaller ? is, the greater the gap between detk (E[z1 z2?])detk (E[z3 z4?]) and either of detk (E[z1 z3?])detk (E[z2 z4?]) and detk (E[z1 z4?])detk (E[z2 z3?]). Therefore, ? also governs how small the ?i,j need to be for the quartet test to return a correct pairing; this is quantified in Lemma 4. Note that Condition 3 implies ? ? ?max ¡ 1. Lemma 4 (Correct pairing). Suppose that (i) the observed variables Z = {z1 , z2 , z3 , z4 } have the true induced tree topology shown in Figure 1(a); (ii) the tree model satisfies Condition 1, Condition 2, and ? ¡ 1 (where ? is defined in (3)), and (iii) the confidence bounds in (2) hold for all {i, j} and all s ? [k]. If ? 1 ? 1 ?i,j ¡ ? min 1, ? 1 ? min{?k (E[zi zj?])} 8k ? {i,j} for each pair {i, j}, then SpectralQuartetTest returns the correct pairing {{z1 , z2 }, {z3 , z4 }}.

4

The Spectral Recursive Grouping algorithm

The Spectral Recursive Grouping algorithm, presented as Algorithm 2, uses the spectral quartet test discussed in the previous section to estimate the structure of a multivariate latent tree distribution from iid samples of the observed leaf variables.1 The algorithm is a modification of the recursive 1 ?x,y and threshold parameTo simplify notation, we assume that the estimated second-moment matrices ? ters ?x,y ? 0 for all pairs {x, y} ? Vobs are globally defined. In particular, we assume the spectral quartet tests use these quantities.

6

Algorithm 2 Spectral Recursive Grouping. ?x,y for all pairs {x, y} ? Vobs computed from N iid Input: Empirical second-moment matrices ? samples from the distribution over Vobs ; threshold parameters ?x,y for all pairs {x, y} ? Vobs . ? or ?failure?. Output: Tree structure T 1: let R := Vobs , and for all x ? R, T [x] := rooted single-node tree x and L[x] := {x}. 2: while —R— ¿ 1 do 3: let pair {u, v} ? {{? u, v?} ? R : Mergeable(R, L[?], u ?, v?) = true} be such that ?x,y) : (x, y) ? L[u] ? L[v]} is maximized. If no such pair exists, then halt max{?k (? and return ?failure?. 4: let result := Relationship(R, L[?], T [?], u, v). 5: if result = ?siblings? then 6: Create a new variable h, create subtree T [h] rooted at h by joining T [u] and T [v] to h with edges {h, u} and {h, v}, and set L[h] := L[u] ? L[v]. 7: Add h to R, and remove u and v from R. 8: else if result = ?u is parent of v? then 9: Modify subtree T [u] by joining T [v] to u with an edge {u, v}, and modify L[u] := L[u] ? L[v]. 10: Remove v from R. 11: else if result = ?v is parent of u? then 12: {Analogous to above case.} 13: end if 14: end while ? := T [h] where R = {h}. 15: Return T grouping (RG) procedure proposed in [21]. RG builds the tree in a bottom-up fashion, where the initial working set of variables are the observed variables. The variables in the working set always correspond to roots of disjoint subtrees of T discovered by the algorithm. (Note that because these subtrees are rooted, they naturally induce parent/child relationships, but these may differ from those

implied by the edge directions in T.) In each iteration, the algorithm determines which variables in the working set to combine. If the variables are combined as siblings, then a new hidden variable is introduced as their parent and is added to the working set, and its children are removed. If the variables are combined as neighbors (parent/child), then the child is removed from the working set. The process repeats until the entire tree is constructed. Our modification of RG uses the spectral quartet tests from Section 3 to decide which subtree roots in the current working set to combine. Note that because the test may return ? (a null result), our algorithm uses the tests to rule out possible siblings or neighbors among variables in the working set?this is encapsulated in the subroutine Mergeable (Algorithm 3), which tests quartets of observed variables (leaves) in the subtrees rooted at working set variables. For any pair {u, v} ? R submitted to the subroutine (along with the current working set R and leaf sets L[?]): ? Mergeable returns false if there is evidence (provided by a quartet test) that u and v should first be joined with different variables (u? and v ? , respectively) before joining with each other; and ? Mergeable returns true if no quartet test provides such evidence. The subroutine is also used by the subroutine Relationship (Algorithm 4) which determines whether a candidate pair of variables should be merged as neighbors (parent/child) or as siblings: essentially, to check if u is a parent of v, it checks if v is a sibling of each child of u. The use of unreliable estimates of long-range correlations is avoided by only considering highly-correlated variables as candidate pairs to merge (where correlation is measured using observed variables in their corresponding subtrees as proxies). This leads to a sample-efficient algorithm for recovering the hidden tree structure. The Spectral Recursive Grouping algorithm enjoys the following guarantee. Theorem 1. Let ? ? (0, 1). Assume the directed tree graphical model T over variables (random vectors) VT = Vobs ? Vhid satisfies Conditions 1, 2, 3, and 4. Suppose the Spectral Recursive 7

Algorithm 3 Subroutine Mergeable(R, L[?], u, v). Input: Set of nodes R; leaf sets L[v] for all v ? R; distinct u, v ? R. Output: true or false. 1: if there exists distinct u? , v ? ? R {u, v} and (x, y, x? , y ?) ? L[u] ? L[v] ? L[u?] ? L[v ?] s.t. SpectralQuartetTest({x, y, x? , y ? }) returns {{x, x? }, {y, y ? }} or {{x, y ? }, {x? , y}} then return false. 2: else return true. Algorithm 4 Subroutine Relationship(R, L[?], T [?], u, v). Input: Set of nodes R; leaf sets L[v] for all v ? R; rooted subtrees T [v] for all v ? R; distinct u, v ? R. Output: ?siblings?, ?u is parent of v? (?u ? v?), or ?v is parent of u? (?v ? u?). 1: if u is a leaf then assert u ?? v. 2: if v is a leaf then assert v ?? u. 3: let R[w] := (R {w}) ? {w ? : w ? is a child of w in T [w]} for each w ? {u, v}. 4: if there exists child u1 of u in T [u] s.t. Mergeable(R[u], L[?], u1 , v) = false then assert ?u ?? v?. 5: if there exists child v1 of v in T [v] s.t. Mergeable(R[v], L[?], u, v1) = false then assert ?v ?? u?. 6: if both ?u ?? v? and ?v ?? u? were asserted then return ?siblings?. 7: else if ?u ?? v? was asserted then return ?v is parent of u? (?v ? u?). 8: else return ?u is parent of v? (?u ? v?). Grouping algorithm (Algorithm 2) is provided N independent samples from the distribution over Vobs , and uses parameters given by ? 2Bxi ,xj txi ,xj Mxi Mxj txi ,xj ?xi ,xj := + (4) N 3N where

? ? ?? ?? 2 ? ? ? ? Bxi ,xj := max ?E[?xi ?2 xj x? , j] , E[?xj ? xi xi] d?xi ,xj :=

E[?xi ? ?xj ?] ? tr(E[xi x?]E[xj x? i]) ? ? j ?? ?? ?, ?E[?xi ?2 xj x?]? , max ?E[?xj ?2 xi x?] i j 2

2

Mxi ? ?xi ?

almost surely,

txi ,xj := 4 ln(4d?xi ,xj n/?).

Let B := maxxi ,xj ?Vobs {Bxi ,xj }, M := maxxi ?Vobs {Mxi }, t := maxxi ,xj ?Vobs {txi ,xj }. If N¿?

200 ? k 2 ? B ? t

2 ?min ? (1 ? ?max) ?max

?2 +

7 ? k ? M2 ? t

2 ?min ? (1 ? ?max) ?max

,

? with then with probability at least 1 ? ?, the Spectral Recursive Grouping algorithm returns a tree T the same undirected graph structure as T.

Consistency is implied by the above theorem with an appropriate scaling of ? with N . The theorem reveals that the sample complexity of the algorithm depends solely on intrinsic spectral properties of the distribution. Note that there is no explicit dependence on the dimensions of the observable variables, which makes the result applicable to high-dimensional settings. Acknowledgements Part of this work was completed while DH was at the Wharton School of the University of Pennsylvania and at Rutgers University. AA was supported by in part by the setup funds at UCI and the AFOSR Award FA9550-10-1-0310.

2 References

[1] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1-2):1?305, 2008.

8

[2] S. Dasgupta and L. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. Journal of Machine Learning Research, 8(Feb):203?226, 2007. [3] K. Chaudhuri, S. Dasgupta, and A. Vattani. Learning mixtures of Gaussians using the k-means algorithm, 2009. arXiv:0912.0086. [4] D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. Journal of Machine Learning Research, 5:1287?1330, 2004. [5] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 14(3):462?467, 1968. [6] N. Friedman, I. Nachman, and D. Pe?er. Learning Bayesian network structure from massive datasets: the ?sparse candidate? algorithm. In Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999. [7] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional Ising

model selection using ?1 regularized logistic regression. Annals of Statistics, 38(3):1287?1319, 2010. [8] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In IEEE Conference on Computer Vision and Pattern Recognition, 2010. [9] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1999. [10] J. Wishart. Sampling errors in the theory of two factors. British Journal of Psychology, 19:180?187, 1928. [11] K. Bollen. Structural Equation Models with Latent Variables. John Wiley & Sons, 1989. [12] P. Buneman. The recovery of trees from measurements of dissimilarity. In F. R. Hodson, D. G. Kendall, and P. Tautu, editors, Mathematics in the Archaeological and Historical Sciences, pages 387?395. 1971. [13] J. Pearl and M. Tarsi. Structuring causal trees. Journal of Complexity, 2(1):60?77, 1986. [14] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4:406?425, 1987. [15] P. L. Erd?os, L. A. Sz?ekely, M. A. Steel, and T. J. Warnow. A few logs suffice to build (almost) all trees: Part II. Theoretical Computer Science, 221:77?118, 1999. [16] M. R. Lacey and J. T. Chang. A signal-to-noise analysis of phylogeny estimation by neighbor-joining: insufficiency of polynomial length sequences. Mathematical Biosciences, 199(2):188?215, 2006. [17] P. L. Erd?os, L. A. Sz?ekely, M. A. Steel, and T. J. Warnow. A few logs suffice to build (almost) all trees (I). Random Structures and Algorithms, 14:153?184, 1999. [18] E. Mossel. Phase transitions in phylogeny. Transactions of the American Mathematical Society, 356(6):2379?2404, 2004. [19] C. Daskalakis, E. Mossel, and S. Roch. Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel?s conjecture. Probability Theory and Related Fields, 149(1?2):149?189, 2011. [20] H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional galtonwatson processes. Annals of Mathematical Statistics, 37:1463?1481, 1966. [21] M. J. Choi, V. Tan, A. Anandkumar, and A. Willsky. Learning latent tree graphical models. Journal of Machine Learning Research, 12:1771?1812, 2011. [22] M. S. Bartlett. Further aspects of the theory of multiple regression. Mathematical Proceedings of the Cambridge Philosophical Society, 34:33?40, 1938. [23] R. J. Muirhead and C. M. Waternaux. Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. Biometrika, 67(1):31?43, 1980. [24] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. Annals of Applied Probability, 16(2):583?614, 2006. [25] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. In Twenty-Second Annual Conference on Learning Theory, 2009. [26] S. M. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank hidden Markov models. In Thirteenth International Conference on Artificial Intelligence and Statistics, 2010. [27] L. Song, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. Hilbert space embeddings of hidden Markov models. In International Conference on Machine Learning, 2010. [28] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. The Annals of Statistics, 37(6A):3099?3132, 2009. [29] J. Pearl. Probabilis-

tic Reasoning in Intelligent Systems?Networks of Plausible Inference. Morgan Kaufmann, 1988. [30] D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices, 2011. arXiv:1104.1672.

9