Journal of Machine Learning Research 12 (2011) 1617-1653 Submitted 5/10; Revised 2/11; Published 5/11

Learning High-Dimensional Markov Forest Distributions: Analysis of Error Rates

Vincent Y. F. Tan VTAN@WISC.EDU
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Center for Pervasive Communications and Computing
Electrical Engineering and Computer Science
University of California, Irvine
Irvine, CA 92697

Alan S. Willsky WILLSKY@MIT.EDU
Stochastic Systems Group
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Editor: Marina Meilă

Abstract

The problem of learning forest-structured discrete graphical models from i.i.d. samples is considered. An algorithm based on pruning of the Chow-Liu tree through adaptive thresholding is proposed. It is shown that this algorithm is both structurally consistent and risk consistent and the error probability of structure learning decays faster than any polynomial in the number of samples under fixed model size. For the high-dimensional scenario where the size of the model d and the number of edges k scale with the number of samples n, sufficient conditions on (n,d,k) are given for the algorithm to satisfy structural and risk consistencies. In addition, the extremal structures for learning are identified; we prove that the independent (resp., tree) model is the hardest (resp., easiest) to learn using the proposed algorithm in terms of error rates for structure learning.

Keywords: graphical models, forest distributions, structural consistency, risk consistency, method of types

1. Introduction

Graphical models (also known as Markov random fields) have a wide range of applications in diverse fields such as signal processing, coding theory and bioinformatics. See Lauritzen (1996), Wainwright and Jordan (2003) and references therein for examples. Inferring the structure and parameters of graphical models from samples is a starting point in all these applications. The structure of the model provides a quantitative interpretation of relationships amongst the given collection of random variables by specifying a set of conditional independence relationships. The parameters of the model quantify the strength of these interactions among the variables.

The challenge in learning graphical models is often compounded by the fact that typically only a small number of samples are available relative to the size of the model (dimension of data). This is referred to as the high-dimensional learning regime, which differs from classical statistics where a large number of samples of fixed dimensionality are available. As a concrete example, in order to analyze the effect of environmental and genetic factors on childhood asthma, clinician scientists in Manchester, UK have been conducting a longitudinal birth-cohort study since 1997 (Custovic et al., 2002; Simpson et al., 2010). The number of variables collected is of the order of $d \approx 10^6$ (dominated by the genetic data) but the number of children in the study is small ($n \approx 10^3$). The paucity of subjects in the study is due in part to the prohibitive cost of collecting high-quality clinical data from willing participants.

In order to learn high-dimensional graphical models, it is imperative to strike the right balance between data fidelity and overfitting. To ameliorate the effect of overfitting, the samples are often fitted to a *sparse graphical model* (Wainwright and Jordan, 2003), with a small number of edges. One popular and tractable class of sparse graphical models is the set of tree[1] models. When restricted to trees, the Chow-Liu algorithm (Chow and Liu, 1968; Chow and Wagner, 1973) provides an efficient implementation of the maximum-likelihood (ML) procedure to learn the structure from independent samples. However, in the high-dimensional regime, even a tree may overfit the data (Liu et al., 2011). In this paper, we consider learning high-dimensional, forest-structured (discrete) graphical models from a given set of samples.

For learning the forest structure, the ML (Chow-Liu) algorithm does not produce a consistent estimate since ML favors richer model classes and hence, outputs a tree in general. We propose a consistent algorithm called CLThres, which has a thresholding mechanism to prune "weak" edges from the Chow-Liu tree. We provide tight bounds on the *overestimation* and *underestimation* errors, that is, the error probability that the output of the algorithm has more or fewer edges than the true model.

1.1 Main Contributions

This paper contains three main contributions. Firstly, we propose an algorithm named CLThres and prove that it is structurally consistent when the true distribution is forest-structured. Secondly, we prove that CLThres is risk consistent, meaning that the risk under the estimated model converges to the risk of the *forest projection*[2] of the underlying distribution, which may not be a forest. We also provide precise convergence rates for structural and risk consistencies. Thirdly, we provide conditions for the consistency of CLThres in the high-dimensional setting.

We first prove that CLThres is structurally consistent, i.e., as the number of samples grows for a fixed model size, the probability of learning the incorrect structure (set of edges), decays to zero for a fixed model size. We show that the error rate is in fact, dominated by the rate of decay of the overestimation error probability.[3] We use an information-theoretic technique known as the *method of types* (Cover and Thomas, 2006, Ch. 11) as well as a recently-developed technique known as Euclidean information theory (Borade and Zheng, 2008). We provide an upper bound on the error probability by using convex duality to find a surprising connection between the overestimation error

1. A *tree* is a *connected*, acyclic graph. We use the term *proper forest* to denote the set of *disconnected*, acyclic graphs.
2. The forest projection is the forest-structured graphical model that is closest in the KL-divergence sense to the true distribution. We define this distribution formally in (12).
3. The overestimation error probability is the probability that the number of edges learned exceeds the true number of edges. The underestimation error is defined analogously.

rate and a semidefinite program (Vandenberghe and Boyd, 1996) and show that the overestimation error in structure learning decays faster than any polynomial in n for a fixed data dimension d.

We then consider the high-dimensional scenario and provide sufficient conditions on the growth of (n,d) (and also the true number of edges k) to ensure that CLThres is structurally consistent. We prove that even if d grows faster than any polynomial in n (and in fact close to exponential in n), structure estimation remains consistent. As a corollary from our analyses, we also show that for CLThres, independent models (resp., tree models) are the "hardest" (resp., "easiest") to learn in the sense that the asymptotic error rate is the highest (resp., lowest), over all models with the same scaling of (n,d). Thus, the empty graph and connected trees are the extremal forest structures for learning. We also prove that CLThres is risk consistent, i.e., the risk of the estimated forest distribution converges to the risk of the forest projection of the true model at a rate of $O_p(d\log d/n^{1-\gamma})$ for any $\gamma > 0$. We compare and contrast this rate to existing results such as Liu et al. (2011). Note that for this result, the true probability model does not need to be a forest-structured distribution. Finally, we use CLThres to learn forest-structured distributions given synthetic and real-world data sets and show that in the finite-sample case, there exists an inevitable trade-off between the underestimation and overestimation errors.

1.2 Related Work

There are many papers that discuss the learning of graphical models from data. See Dudik et al. (2004), Lee et al. (2006), Abbeel et al. (2006), Wainwright et al. (2006), Meinshausen and Buehlmann (2006), Johnson et al. (2007), and references therein. Most of these methods pose the learning problem as a parameterized convex optimization problem, typically with a regularization term to enforce sparsity in the learned graph. Consistency guarantees in terms of n and d (and possibly the maximum degree) are provided. Information-theoretic limits for learning graphical models have also been derived in Santhanam and Wainwright (2008). In Zuk et al. (2006), bounds on the error rate for learning the structure of Bayesian networks using the Bayesian Information Criterion (BIC) were provided. Bach and Jordan (2003) learned tree-structured models for solving the independent component analysis (ICA) problem. A PAC analysis for learning thin junction trees was given in Chechetka and Guestrin (2007). Meilă and Jordan (2000) discussed the learning of graphical models from a different perspective; namely that of learning mixtures of trees via an expectation-maximization procedure.

By using the theory of large-deviations (Dembo and Zeitouni, 1998; Den Hollander, 2000), we derived and analyzed the error exponent for learning trees for discrete (Tan et al., 2011) and Gaussian (Tan et al., 2010a) graphical models. The error exponent is a quantitative measure of performance of the learning algorithm since a larger exponent implies a faster decay of the error probability. However, the analysis does not readily extend to learning forest models and furthermore it was for the scenario when number of variables d does not grow with the number of samples n. In addition, we also posed the structure learning problem for trees as a composite hypothesis testing problem (Tan et al., 2010b) and derived a closed-form expression for the Chernoff-Stein exponent in terms of the mutual information on the bottleneck edge.

In a paper that is closely related to ours, Liu et al. (2011) derived consistency (and sparsistency) guarantees for learning tree and forest models. The pairwise joint distributions are modeled using kernel density estimates, where the kernels are Hölder continuous. This differs from our approach since we assume that each variable can only take finitely many values, leading to stronger results on

error rates for structure learning via the method of types, a powerful proof technique in information theory and statistics. We compare our convergence rates to these related works in Section 6. Furthermore, the algorithm suggested in both papers uses a subset (usually half) of the data set to learn the full tree model and then uses the remaining subset to prune the model based on the log-likelihood on the held-out set. We suggest a more direct and consistent method based on thresholding, which uses the *entire* data set to learn and prune the model without recourse to validation on a held-out data set. It is well known that validation is both computationally expensive (Bishop, 2008, pp. 33) and a potential waste of valuable data which may otherwise be employed to learn a better model. In Liu et al. (2011), the problem of estimating forests with restricted component sizes was considered and was proven to be NP-hard. We do not restrict the component size in this paper but instead attempt to learn the model with the minimum number of edges which best fits the data.

Our work is also related to and inspired by the vast body of literature in information theory and statistics on Markov order estimation. In these works, the authors use various regularization and model selection schemes to find the optimal order of a Markov chain (Merhav et al., 1989; Finesso et al., 1996; Csiszár and Shields, 2000), hidden Markov model (Gassiat and Boucheron, 2003) or exponential family (Merhav, 1989). We build on some of these ideas and proof techniques to identify the correct set of edges (and in particular the number of edges) in the forest model and also to provide strong theoretical guarantees of the rate of convergence of the estimated forest-structured distribution to the true one.

1.3 Organization of Paper

This paper is organized as follows: We define the mathematical notation and formally state the problem in Section 2. In Section 3, we describe the algorithm in full detail, highlighting its most salient aspect—the thresholding step. We state our main results on error rates for structure learning in Section 4 for a fixed forest-structured distribution. We extend these results to the high-dimensional case when (n,d,k) scale in Section 5. Extensions to rates of convergence of the estimated distribution, that is, the order of risk consistency, are discussed briefly in Section 6. Numerical simulations on synthetic and real data are presented in Section 7. Finally, we conclude the discussion in Section 8. The proofs of the majority of the results are provided in the appendices.

2. Preliminaries and Problem Formulation

Let $G = (V,E)$ be an undirected graph with vertex (or node) set $V := \{1,\ldots,d\}$ and edge set $E \subset \binom{V}{2}$ and let $\mathrm{nbd}(i) := \{j \in V : (i,j) \in E\}$ be the set of neighbors of vertex i. Let the set of labeled *trees* (connected, acyclic graphs) with d nodes be $\mathcal{T}^d$ and let the set of *forests* (acyclic graphs) with k edges and d nodes be $\mathcal{T}^d_k$ for $0 \le k \le d-1$. The set of forests includes all the trees. We reserve the term *proper forests* for the set of disconnected acylic graphs $\cup_{k=0}^{d-2}\mathcal{T}^d_k$. We also use the notation $\mathcal{F}^d := \cup_{k=0}^{d-1}\mathcal{T}^d_k$ to denote the set of labeled forests with d nodes.

A *graphical model* (Lauritzen, 1996) is a family of multivariate probability distributions (probability mass functions) in which each distribution factorizes according to a given undirected graph and where each variable is associated to a node in the graph. Let $\mathcal{X} = \{1,\ldots,r\}$ (where $2 \le r < \infty$) be a finite set and $\mathcal{X}^d$ the d-fold Cartesian product of the set $\mathcal{X}$. As usual, let $\mathcal{P}(\mathcal{X}^d)$ denote the probability simplex over the alphabet $\mathcal{X}^d$. We say that the random vector $\mathbf{X} = (X_1,\ldots,X_d)$ with

distribution $Q \in \mathcal{P}(\mathcal{X}^d)$ is *Markov on* the graph $G = (V, E)$ if

$$Q(x_i|x_{\text{nbd}(i)}) = Q(x_i|x_{V\setminus i}), \qquad \forall i \in V, \tag{1}$$

where $x_{V\setminus i}$ is the collection of variables excluding variable i. Equation (1) is known as the *local Markov property* (Lauritzen, 1996). In this paper, we always assume that graphs are *minimal representations* for the corresponding graphical model, that is, if Q is Markov on G, then G has the smallest number of edges for the conditional independence relations in (1) to hold. We say the distribution Q is a *forest-structured distribution* if it is Markov on a forest. We also use the notation $\mathcal{D}(\mathcal{T}_k^d) \subset \mathcal{P}(\mathcal{X}^d)$ to denote the set of d-variate distributions Markov on a forest with k edges. Similarly, $\mathcal{D}(\mathcal{F}^d)$ is the set of forest-structured distributions.

Let $P \in \mathcal{D}(\mathcal{T}_k^d)$ be a discrete forest-structured distribution Markov on $T_P = (V, E_P) \in \mathcal{T}_k^d$ (for some $k = 0, \ldots, d-1$). It is known that the joint distribution P factorizes as follows (Lauritzen, 1996; Wainwright and Jordan, 2003):

$$P(\mathbf{x}) = \prod_{i\in V} P_i(x_i) \prod_{(i,j)\in E_P} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i)P_j(x_j)},$$

where $\{P_i\}_{i\in V}$ and $\{P_{i,j}\}_{(i,j)\in E_P}$ are the node and pairwise marginals which are assumed to be positive everywhere.

The mutual information (MI) of two random variables X_i and X_j with joint distribution $P_{i,j}$ is the function $I(\cdot) : \mathcal{P}(\mathcal{X}^2) \to [0, \log r]$ defined as

$$I(P_{i,j}) := \sum_{(x_i,x_j)\in\mathcal{X}^2} P_{i,j}(x_i, x_j) \log \frac{P_{i,j}(x_i, x_j)}{P_i(x_i)P_j(x_j)}. \tag{2}$$

This notation for mutual information differs from the usual $I(X_i; X_j)$ used in Cover and Thomas (2006); we emphasize the dependence of I on the joint distribution $P_{i,j}$. The *minimum mutual information* in the forest, denoted as $I_{\min} := \min_{(i,j)\in E_P} I(P_{i,j})$ will turn out to be a fundamental quantity in the subsequent analysis. Note from our minimality assumption that $I_{\min} > 0$ since all edges in the forest have positive mutual information (none of the edges are degenerate). When we consider the scenario where d grows with n in Section 5, we assume that $I_{\min}$ is *uniformly* bounded away from zero.

2.1 Problem Statement

We now state the basic problem formally. We are given a set of i.i.d. samples, denoted as $\mathbf{x}^n := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Each sample $\mathbf{x}_l = (x_{l,1}, \ldots, x_{l,d}) \in \mathcal{X}^d$ is drawn independently from $P \in \mathcal{D}(\mathcal{T}_k^d)$ a forest-structured distribution. From these samples, and the prior knowledge that the undirected graph is acyclic (but not necessarily connected), estimate the true set of edges E_P as well as the true distribution P consistently.

3. The Forest Learning Algorithm: CLThres

We now describe our algorithm for estimating the edge set E_P and the distribution P. This algorithm is a modification of the celebrated Chow-Liu algorithm for maximum-likelihood (ML) learning of

tree-structured distributions (Chow and Liu, 1968). We call our algorithm CLThres which stands for *Chow-Liu with Thresholding*.

The inputs to the algorithm are the set of samples $\mathbf{x}^n$ and a *regularization* sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ (to be specified precisely later) that typically decays to zero, that is, $\lim_{n\to\infty}\varepsilon_n = 0$. The outputs are the estimated edge set, denoted $\widehat{E}_{\widehat{k}_n}$, and the estimated distribution, denoted P^*.

1. Given $\mathbf{x}^n$, calculate the set of *pairwise empirical distributions*[4] (or *pairwise types*) $\{\widehat{P}_{i,j}\}_{i,j\in V}$. This is just a normalized version of the counts of each observed symbol in $\mathcal{X}^2$ and serves as a set of sufficient statistics for the estimation problem. The dependence of $\widehat{P}_{i,j}$ on the samples $\mathbf{x}^n$ is suppressed.

2. Form the set of *empirical mutual information* quantities:

$$I(\widehat{P}_{i,j}) := \sum_{(x_i,x_j)\in\mathcal{X}^2} \widehat{P}_{i,j}(x_i,x_j)\log\frac{\widehat{P}_{i,j}(x_i,x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)},$$

   for $1 \le i,j \le d$. This is a consistent estimator of the true mutual information in (2).

3. Run a max-weight spanning tree (MWST) algorithm (Prim, 1957; Kruskal, 1956) to obtain an estimate of the edge set:

$$\widehat{E}_{d-1} := \operatorname*{argmax}_{E:T=(V,E)\in\mathcal{T}^d} \sum_{(i,j)\in E} I(\widehat{P}_{i,j}).$$

   Let the estimated edge set be $\widehat{E}_{d-1} := \{\widehat{e}_1,\ldots,\widehat{e}_{d-1}\}$ where the edges $\widehat{e}_i$ are sorted according to decreasing empirical mutual information values. We index the edge set by $d-1$ to emphasize that it has $d-1$ edges and hence is connected. We denote the sorted empirical mutual information quantities as $I(\widehat{P}_{\widehat{e}_1}) \ge \ldots \ge I(\widehat{P}_{\widehat{e}_{d-1}})$. These first three steps constitute the Chow-Liu algorithm (Chow and Liu, 1968).

4. Estimate the true number of edges using the *thresholding estimator*:

$$\widehat{k}_n := \operatorname*{argmin}_{1\le j\le d-1} \left\{ I(\widehat{P}_{\widehat{e}_j}) : I(\widehat{P}_{\widehat{e}_j}) \ge \varepsilon_n, I(\widehat{P}_{\widehat{e}_{j+1}}) \le \varepsilon_n \right\}. \tag{3}$$

   If there exists an empirical mutual information $I(\widehat{P}_{\widehat{e}_j})$ such that $I(\widehat{P}_{\widehat{e}_j}) = \varepsilon_n$, break the tie arbitrarily.[5]

5. Prune the tree by retaining only the top $\widehat{k}_n$ edges, that is, define the *estimated edge set* of the forest to be

$$\widehat{E}_{\widehat{k}_n} := \{\widehat{e}_1,\ldots,\widehat{e}_{\widehat{k}_n}\},$$

   where $\{\widehat{e}_i : 1 \le i \le d-1\}$ is the ordered edge set defined in Step 3. Define the estimated forest to be $\widehat{T}_{\widehat{k}_n} := (V,\widehat{E}_{\widehat{k}_n})$.

4. In this paper, the terms *empirical distribution* and *type* are used interchangeably.
5. Here were allow a bit of imprecision by noting that the non-strict inequalities in (3) simplify the subsequent analyses because the constraint sets that appear in optimization problems will be closed, hence compact, insuring the existence of optimizers.

6. Finally, define the estimated distribution P^* to be the *reverse I-projection* (Csiszár and Matúš, 2003) of the joint type $\widehat{P}$ onto $\widehat{T}_{\widehat{k}_n}$, that is,

$$P^*(\mathbf{x}) := \operatorname*{argmin}_{Q \in \mathcal{D}(\widehat{T}_{\widehat{k}_n})} D(\widehat{P}||Q).$$

It can easily be shown that the projection can be expressed in terms of the marginal and pairwise joint types:

$$P^*(\mathbf{x}) = \prod_{i \in V} \widehat{P}_i(x_i) \prod_{(i,j) \in \widehat{E}_{\widehat{k}_n}} \frac{\widehat{P}_{i,j}(x_i, x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)}.$$

Intuitively, CLThres first constructs a connected tree $(V, \widehat{E}_{d-1})$ via Chow-Liu (in Steps 1–3) before pruning the weak edges (with small mutual information) to obtain the final structure $\widehat{E}_{\widehat{k}_n}$. The estimated distribution P^* is simply the ML estimate of the parameters subject to the constraint that P^* is Markov on the learned tree $\widehat{T}_{\widehat{k}_n}$.

Note that if Step 4 is omitted and $\widehat{k}_n$ is defined to be $d-1$, then CLThres simply reduces to the Chow-Liu ML algorithm. Of course Chow-Liu, which outputs a tree, is guaranteed to fail (not be structurally consistent) if the number of edges in the true model $k < d-1$, which is the problem of interest in this paper. Thus, Step 4, a model selection step, is essential in estimating the true number of edges k. This step is a generalization of the test for independence of discrete memoryless sources discussed in Merhav (1989). In our work, we exploit the fact that the empirical mutual information $I(\widehat{P}_{\widehat{e}_j})$ corresponding to a pair of independent variables $\widehat{e}_j$ will be very small when n is large, thus a thresholding procedure using the (appropriately chosen) regularization sequence $\{\varepsilon_n\}$ will remove these edges. In fact, the subsequent analysis allows us to conclude that Step 4, in a formal sense, *dominates* the error probability in structure learning. CLThres is also efficient as shown by the following result.

Proposition 1 (Complexity of CLThres) CLThres *runs in time* $O((n+\log d)d^2)$.

Proof The computation of the sufficient statistics in Steps 1 and 2 requires $O(nd^2)$ operations. The MWST algorithm in Step 3 requires at most $O(d^2 \log d)$ operations (Prim, 1957). Steps 4 and 5 simply require the sorting of the empirical mutual information quantities on the learned tree which only requires $O(\log d)$ computations. ■

4. Structural Consistency For Fixed Model Size

In this section, we keep d and k fixed and consider a probability model P, which is assumed to be Markov on a forest in $\mathcal{T}_k^d$. This is to gain better insight into the problem before we analyze the high-dimensional scenario in Section 5 where d and k scale[6] with the sample size n. More precisely, we are interested in quantifying the rate at which the probability of the error event of structure learning

$$\mathcal{A}_n := \left\{ \mathbf{x}^n \in (\mathcal{X}^d)^n : \widehat{E}_{\widehat{k}_n}(\mathbf{x}^n) \neq E_P \right\} \tag{4}$$

6. In that case P must also scale, that is, we learn a *family* of models as d and k scale.

decays to zero as n tends to infinity. Recall that $\widehat{E}_{\widehat{k}_n}$, with cardinality $\widehat{k}_n$, is the learned edge set by using CLThres. As usual, P^n is the n-fold product probability measure corresponding to the forest-structured distribution P.

Before stating the main result of this section in Theorem 3, we first state an auxiliary result that essentially says that if one is provided with oracle knowledge of $I_{\min}$, the minimum mutual information in the forest, then the problem is greatly simplified.

Proposition 2 (Error Rate with knowledge of $I_{\min}$) *Assume that $I_{\min}$ is known in* CLThres*. Then by letting the regularization sequence be $\varepsilon_n = I_{\min}/2$ for all n, we have*

$$\lim_{n\to\infty} \frac{1}{n} \log P^n(\mathcal{A}_n) < 0, \tag{5}$$

that is, the error probability decays exponentially fast.

The proof of this theorem and all other results in the sequel can be found in the appendices.

Thus, the primary difficulty lies in estimating $I_{\min}$ or equivalently, the number of edges k. Note that if k is known, a simple modification to the Chow-Liu procedure by imposing the constraint that the final structure contains k edges will also yield exponential decay as in (5). However, in the realistic case where both $I_{\min}$ and k are unknown, we show in the rest of this section that we can design the regularization sequence ε_n in such a way that the rate of decay of $P^n(\mathcal{A}_n)$ decays almost exponentially fast.

4.1 Error Rate for Forest Structure Learning

We now state one of the main results in this paper. We emphasize that the following result is stated for a fixed forest-structured distribution $P \in \mathcal{D}(\mathcal{T}_k^d)$ so d and k are also fixed natural numbers.

Theorem 3 (Error Rate for Structure Learning) *Assume that the regularization sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ satisfies the following two conditions:*

$$\lim_{n\to\infty} \varepsilon_n = 0, \qquad \lim_{n\to\infty} \frac{n\varepsilon_n}{\log n} = \infty. \tag{6}$$

Then, if the true model $T_P = (V, E_P)$ is a proper forest ($k < d-1$), there exists a constant $C_P \in (1,\infty)$ such that

$$-C_P \leq \liminf_{n\to\infty} \frac{1}{n\varepsilon_n} \log P^n(\mathcal{A}_n) \tag{7}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n\varepsilon_n} \log P^n(\mathcal{A}_n) \leq -1. \tag{8}$$

Finally, if the true model $T_P = (V, E_P)$ is a tree ($k = d-1$), then

$$\lim_{n\to\infty} \frac{1}{n} \log P^n(\mathcal{A}_n) < 0, \tag{9}$$

that is, the error probability decays exponentially fast.

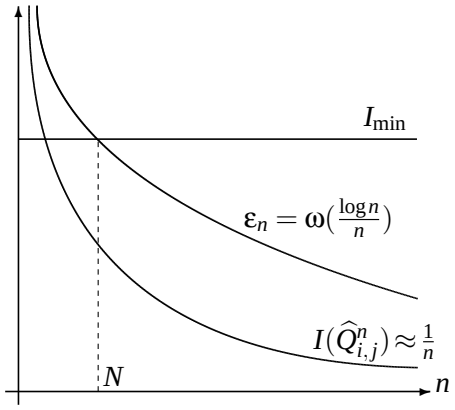

Figure 1: Graphical interpretation of the condition on ε_n. As $n \to \infty$, the regularization sequence ε_n will be smaller than $I_{\min}$ and larger than $I(\widehat{Q}^n_{i,j})$ with high probability.

4.2 Interpretation of Result

From (8), the rate of decay of the error probability for proper forests is subexponential but nonetheless can be made faster than any polynomial for an appropriate choice of ε_n. The reason for the subexponential rate is because of our lack of knowledge of $I_{\min}$, the minimum mutual information in the true forest T_P. For trees, the rate[7] is exponential ($\doteq \exp(-nF)$ for some positive constant F). Learning proper forests is thus, strictly "harder" than learning trees. The condition on ε_n in (6) is needed for the following intuitive reasons:

1. Firstly, (6) ensures that for all sufficiently large n, we have $\varepsilon_n < I_{\min}$. Thus, the true edges will be correctly identified by CLThres implying that with high probability, there will not be underestimation as $n \to \infty$.

2. Secondly, for two independent random variables X_i and X_j with distribution $Q_{i,j} = Q_i Q_j$, the sequence[8] $\sigma(I(\widehat{Q}^n_{i,j})) = \Theta(1/n)$, where $\widehat{Q}^n_{i,j}$ is the joint empirical distribution of n i.i.d. samples drawn from $Q_{i,j}$. Since the regularization sequence $\varepsilon_n = \omega(\log n/n)$ has a slower rate of decay than $\sigma(I(\widehat{Q}^n_{i,j}))$, $\varepsilon_n > I(\widehat{Q}^n_{i,j})$ with high probability as $n \to \infty$. Thus, with high probability there will not be overestimation as $n \to \infty$.

See Figure 1 for an illustration of this intuition. The formal proof follows from a method of types argument and we provide an outline in Section 4.3. A convenient choice of ε_n that satisfies (6) is

$$\varepsilon_n := n^{-\beta}, \qquad \forall \beta \in (0,1). \tag{10}$$

Note further that the upper bound in (8) is also independent of P since it is equal to -1 for all P. Thus, (8) is a *universal* result for all forest distributions $P \in \mathcal{D}(\mathcal{F}^d)$. The intuition for this

7. We use the asymptotic notation from information theory $\doteq$ to denote equality to first order in the exponent. More precisely, for two positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ we say that $a_n \doteq b_n$ iff $\lim_{n\to\infty} n^{-1}\log(a_n/b_n) = 0$.
8. The notation $\sigma(Z)$ denotes the standard deviation of the random variable Z. The fact that the standard deviation of the empirical MI $\sigma(I(\widehat{Q}^n_{i,j}))$ decays as $1/n$ can be verified by Taylor expanding $I(\widehat{Q}^n_{i,j})$ around $Q_{i,j} = Q_i Q_j$ and using the fact that the ML estimate converges at a rate of $n^{-1/2}$ (Serfling, 1980).

universality is because in the large-n regime, the typical way an error occurs is due to overestimation. The overestimation error results from testing whether pairs of random variables are independent and our asymptotic bound for the error probability of this test does not depend on the true distribution P.

The lower bound C_P in (7), defined in the proof in Appendix B, means that we cannot hope to do much better using CLThres if the original structure (edge set) is a proper forest. Together, (7) and (8) imply that the rate of decay of the error probability for structure learning is tight to within a constant factor in the exponent. We believe that the error rates given in Theorem 3 cannot, in general, be improved without knowledge of $I_{\min}$. We state a converse (a necessary lower bound on sample complexity) in Theorem 7 by treating the unknown forest graph as a uniform random variable over all possible forests of fixed size.

4.3 Proof Idea

The method of proof for Theorem 3 involves using the Gallager-Fano bounding technique (Fano, 1961, pp. 24) and the union bound to decompose the overall error probability $P^n(\mathcal{A}_n)$ into three distinct terms: (i) the rate of decay of the error probability for learning the top k edges (in terms of the mutual information quantities) correctly—known as the *Chow-Liu error*, (ii) the rate of decay of the *overestimation error* $\{\widehat{k}_n > k\}$ and (iii) the rate of decay of the *underestimation error* $\{\widehat{k}_n < k\}$. Each of these terms is upper bounded using a method of types (Cover and Thomas, 2006, Ch. 11) argument. It turns out, as is the case with the literature on Markov order estimation (e.g., Finesso et al., 1996), that bounding the overestimation error poses the greatest challenge. Indeed, we show that the underestimation and Chow-Liu errors have exponential decay in n. However, the overestimation error has subexponential decay ($\approx \exp(-n\varepsilon_n)$).

The main technique used to analyze the overestimation error relies on *Euclidean information theory* (Borade and Zheng, 2008) which states that if two distributions ν_0 and ν_1 (both supported on a common finite alphabet $\mathcal{Y}$) are close entry-wise, then various information-theoretic measures can be approximated locally by quantities related to Euclidean norms. For example, the KL-divergence $D(\nu_0 \,||\, \nu_1)$ can be approximated by the square of a weighted Euclidean norm:

$$D(\nu_0 \,||\, \nu_1) = \frac{1}{2} \sum_{a \in \mathcal{Y}} \frac{(\nu_0(a) - \nu_1(a))^2}{\nu_0(a)} + o(\|\nu_0 - \nu_1\|_\infty^2). \tag{11}$$

Note that if $\nu_0 \approx \nu_1$, then $D(\nu_0 \,||\, \nu_1)$ is close to the sum in (11) and the $o(\|\nu_0 - \nu_1\|_\infty^2)$ term can be neglected. Using this approximation and Lagrangian duality (Bertsekas, 1999), we reduce a non-convex I-projection (Csiszár and Matúš, 2003) problem involving information-theoretic quantities (such as divergence) to a relatively simple *semidefinite program* (Vandenberghe and Boyd, 1996) which admits a closed-form solution. Furthermore, the approximation in (11) becomes *exact* as $n \to \infty$ (i.e., $\varepsilon_n \to 0$), which is the asymptotic regime of interest. The full details of the proof can be found Appendix B.

4.4 Error Rate for Learning the Forest Projection

In our discussion thus far, P has been assumed to be Markov on a forest. In this subsection, we consider the situation when the underlying unknown distribution P is not forest-structured but we wish to learn its best forest approximation. To this end, we define the projection of P onto the set of

forests (or *forest projection*) to be

$$\widetilde{P} := \operatorname*{argmin}_{Q \in \mathcal{D}(\mathcal{F}^d)} D(P||Q). \tag{12}$$

If there are multiple optimizing distribution, choose a projection $\widetilde{P}$ that is minimal, that is, its graph $T_{\widetilde{P}} = (V, E_{\widetilde{P}})$ has the *fewest number of edges* such that (12) holds. If we redefine the event $\mathcal{A}_n$ in (4) to be $\widetilde{\mathcal{A}}_n := \{\widehat{E}_{\widehat{k}_n} \neq E_{\widetilde{P}}\}$, we have the following analogue of Theorem 3.

Corollary 4 (Error Rate for Learning Forest Projection) *Let P be an arbitrary distribution and the event $\widetilde{\mathcal{A}}_n$ be defined as above. Then the conclusions in* (7)–(9) *in Theorem 3 hold if the regularization sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ satisfies* (6)*.*

5. High-Dimensional Structural Consistency

In the previous section, we considered learning a fixed forest-structured distribution P (and hence fixed d and k) and derived bounds on the error rate for structure learning. However, for most problems of practical interest, the number of data samples is small compared to the data dimension d (see the asthma example in the introduction). In this section, we prove sufficient conditions on the scaling of (n,d,k) for structure learning to remain consistent. We will see that even if d and k are much larger than n, under some reasonable regularity conditions, structure learning remains consistent.

5.1 Structure Scaling Law

To pose the learning problem formally, we consider a *sequence* of structure learning problems indexed by the number of data points n. For the particular problem indexed by n, we have a data set $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of size n where each sample $\mathbf{x}_l \in \mathcal{X}^d$ is drawn independently from an unknown d-variate forest-structured distribution $P^{(d)} \in \mathcal{D}(\mathcal{T}_k^d)$, which has d nodes and k edges and where d and k depend on n. This *high-dimensional* setup allows us to model and subsequently analyze how d and k can scale with n while maintaining consistency. We will sometimes make the dependence of d and k on n explicit, that is, $d = d_n$ and $k = k_n$.

In order to be able to learn the structure of the models we assume that

$$\text{(A1)} \quad I_{\inf} := \inf_{d\in\mathbb{N}} \min_{(i,j)\in E_{P^{(d)}}} I(P_{i,j}^{(d)}) > 0, \tag{13}$$

$$\text{(A2)} \quad \kappa := \inf_{d\in\mathbb{N}} \min_{x_i,x_j\in\mathcal{X}} P_{i,j}^{(d)}(x_i,x_j) > 0. \tag{14}$$

That is, assumptions (A1) and (A2) insure that there exists *uniform* lower bounds on the minimum mutual information and the minimum entry in the pairwise probabilities in the forest models as the size of the graph grows. These are typical regularity assumptions for the high-dimensional setting. See Wainwright et al. (2006) and Meinshausen and Buehlmann (2006) for example. We again emphasize that the proposed learning algorithm CLThres has knowledge of neither $I_{\inf}$ nor κ. Equipped with (A1) and (A2) and assuming the asymptotic behavior of ε_n in (6), we claim the following theorem for CLThres.

Theorem 5 (Structure Scaling Law) *There exists two finite, positive constants C_1, C_2 such that if*

$$n > \max\left\{(2\log(d-k))^{1+\zeta}, C_1 \log d, C_2 \log k\right\}, \tag{15}$$

for any $\zeta > 0$, then the error probability of incorrectly learning the sequence of edge sets $\{E_{P^{(d)}}\}_{d\in\mathbb{N}}$ tends to zero as $(n,d,k) \to \infty$. When the sequence of forests are trees, $n > C\log d$ (where $C := \max\{C_1, C_2\}$) suffices for high-dimensional structure recovery.

Thus, if the model parameters (n,d,k) all grow with n but $d = o(\exp(n/C_1))$, $k = o(\exp(n/C_2))$ and $d-k = o(\exp(n^{1-\beta}/2))$ (for all $\beta > 0$), consistent structure recovery is possible in high dimensions. In other words, the number of nodes d can grow faster than any polynomial in the sample size n. In Liu et al. (2011), the bivariate densities are modeled by functions from a Hölder class with exponent α and it was mentioned (in Theorem 4.3) that the number of variables can grow like $o(\exp(n^{\alpha/(1+\alpha)}))$ for structural consistency. Our result is somewhat stronger but we model the pairwise joint distributions as (simpler) probability mass functions (the alphabet $\mathcal{X}$ is a finite set).

5.2 Extremal Forest Structures

In this subsection, we study the extremal structures for learning, that is, the structures that, roughly speaking, lead to the largest and smallest error probabilities for structure learning. Define the sequence

$$h_n(P) := \frac{1}{n\varepsilon_n} \log P^n(\mathcal{A}_n), \quad \forall n \in \mathbb{N}. \tag{16}$$

Note that h_n is a function of both the number of variables $d = d_n$ and the number of edges $k = k_n$ in the models $P^{(d)}$ since it is a sequence indexed by n. In the next result, we assume (n,d,k) satisfies the scaling law in (15) and answer the following question: How does h_n in (16) depend on the number of edges k_n for a given d_n? Let $P_1^{(d)}$ and $P_2^{(d)}$ be two sequences of forest-structured distributions with a common number of nodes d_n and number of edges $k_n(P_1^{(d)})$ and $k_n(P_2^{(d)})$ respectively.

Corollary 6 (Extremal Forests) *Assume that* CLThres *is employed as the forest learning algorithm. As $n \to \infty$, $h_n(P_1^{(d)}) \le h_n(P_2^{(d)})$ whenever $k_n(P_1^{(d)}) \ge k_n(P_2^{(d)})$ implying that h_n is maximized when $P^{(d)}$ are product distributions (i.e., $k_n = 0$) and minimized when $P^{(d)}$ are tree-structured distributions (i.e., $k_n = d_n - 1$). Furthermore, if $k_n(P_1^{(d)}) = k_n(P_2^{(d)})$, then $h_n(P_1^{(d)}) = h_n(P_2^{(d)})$.*

Note that the corollary is intimately tied to the proposed algorithm CLThres. We are not claiming that such a result holds for all other forest learning algorithms. The intuition for this result is the following: We recall from the discussion after Theorem 3 that the overestimation error dominates the probability of error for structure learning. Thus, the performance of CLThres degrades with the number of missing edges. If there are very few edges (i.e., k_n is very small relative to d_n), the CLThres estimator is more likely to overestimate the number of edges as compared to if there are many edges (i.e., k_n/d_n is close to 1). We conclude that a distribution which is Markov on an *empty graph* (all variables are independent) is the *hardest* to learn (in the sense of Corollary 6 above). Conversely, *trees* are the *easiest* structures to learn.

5.3 Lower Bounds on Sample Complexity

Thus far, our results are for a specific algorithm CLThres for learning the structure of Markov forest distributions. At this juncture, it is natural to ask whether the scaling laws in Theorem 5 are the best

possible over all algorithms (estimators). To answer this question, we limit ourselves to the scenario where the true graph T_P is a uniformly distributed chance variable[9] with probability measure $\mathbb{P}$. Assume two different scenarios:

(a) T_P is drawn from the uniform distribution on $\mathcal{T}_k^d$, that is, $\mathbb{P}(T_P = t) = 1/|\mathcal{T}_k^d|$ for all forests $t \in \mathcal{T}_k^d$. Recall that $\mathcal{T}_k^d$ is the set of labeled forests with d nodes and k edges.

(b) T_P is drawn from the uniform distribution on $\mathcal{F}^d$, that is, $\mathbb{P}(T_P = t) = 1/|\mathcal{F}^d|$ for all forests $t \in \mathcal{F}^d$. Recall that $\mathcal{F}^d$ is the set of labeled forests with d nodes.

This following result is inspired by Theorem 1 in Bresler et al. (2008). Note that an *estimator* or *algorithm* $\widehat{T}^d$ is simply a map from the set of samples $(\mathcal{X}^d)^n$ to a set of graphs (either $\mathcal{T}_k^d$ or $\mathcal{F}^d$). We emphasize that the following result is stated with the assumption that we are *averaging* over the random choice of the true graph T_P.

Theorem 7 (Lower Bounds on Sample Complexity) *Let* $\rho < 1$ *and* $r := |\mathcal{X}|$. *In case (a) above, if*

$$n < \rho \, \frac{(k-1)\log d}{d \log r}, \tag{17}$$

then $\mathbb{P}(\widehat{T}^d \neq T_P) \to 1$ *for any estimator* $\widehat{T}^d : (\mathcal{X}^d)^n \to \mathcal{T}_k^d$. *Alternatively, in case (b), if*

$$n < \rho \, \frac{\log d}{\log r}, \tag{18}$$

then $\mathbb{P}(\widehat{T}^d \neq T_P) \to 1$ *for any estimator* $\widehat{T}^d : (\mathcal{X}^d)^n \to \mathcal{F}^d$.

This result, a *strong converse*, states that $n = \Omega(\frac{k}{d}\log d)$ is *necessary* for any estimator with oracle knowledge of k to succeed. Thus, we need at least logarithmically many samples in d if the fraction k/d is kept constant as the graph size grows even if k *is known precisely* and does not have to be estimated. Interestingly, (17) says that if k is large, then we need more samples. This is because there are fewer forests with a small number of edges as compared to forests with a large number of edges. In contrast, the performance of CLThres degrades when k is small because it is more sensitive to the overestimation error. Moreover, if the estimator does not know k, then (18) says that $n = \Omega(\log d)$ is *necessary* for successful recovery. We conclude that the set of scaling requirements prescribed in Theorem 5 is almost optimal. In fact, if the true structure T_P is a tree, then Theorem 7 for CLThres says that the (achievability) scaling laws in Theorem 5 are indeed optimal (up to constant factors in the O and Ω-notation) since $n > (2\log(d-k))^{1+\zeta}$ in (15) is trivially satisfied. Note that if T_P is a tree, then the Chow-Liu ML procedure or CLThres results in the sample complexity $n = O(\log d)$ (see Theorem 5).

6. Risk Consistency

In this section, we develop results for risk consistency to study how fast the parameters of the estimated distribution converge to their true values. For this purpose, we define the *risk* of the estimated distribution P^* (with respect to the true probability model P) as

$$\mathcal{R}_n(P^*) := D(P\,||\,P^*) - D(P\,||\,\widetilde{P}), \tag{19}$$

9. The term *chance variable*, attributed to Gallager (2001), describes random quantities $Y : \Omega \to W$ that take on values in arbitrary alphabets W. In contrast, a random variable X maps the sample space Ω to the reals $\mathbb{R}$.

where $\widetilde{P}$ is the forest projection of P defined in (12). Note that the original probability model P does not need to be a forest-structured distribution in the definition of the risk. Indeed, if P is Markov on a forest, (19) reduces to $\mathcal{R}_n(P^*) = D(P||P^*)$ since the second term is zero. We quantify the rate of decay of the risk when the number of samples n tends to infinity. For $\delta > 0$, we define the event

$$\mathcal{C}_{n,\delta} := \left\{ \mathbf{x}^n \in (\mathcal{X}^d)^n : \frac{\mathcal{R}_n(P^*)}{d} > \delta \right\}. \tag{20}$$

That is, $\mathcal{C}_{n,\delta}$ is the event that the *average risk* $\mathcal{R}_n(P^*)/d$ exceeds some constant δ. We say that the estimator P^* (or an algorithm) is δ-*risk consistent* if the probability of $\mathcal{C}_{n,\delta}$ tends to zero as $n \to \infty$. Intuitively, achieving δ-risk consistency is easier than achieving structural consistency since the learned model P^* can be close to the true forest-projection $\widetilde{P}$ in the KL-divergence sense even if their structures differ.

In order to quantify the rate of decay of the risk in (19), we need to define some stochastic order notation. We say that a sequence of random variables $Y_n = O_p(g_n)$ (for some deterministic positive sequence $\{g_n\}$) if for every $\varepsilon > 0$, there exists a $B = B_\varepsilon > 0$ such that $\limsup_{n\to\infty} \Pr(|Y_n| > Bg_n) < \varepsilon$. Thus, $\Pr(|Y_n| > Bg_n) \geq \varepsilon$ holds for only finitely many n.

We say that a reconstruction algorithm has *risk consistency of order* (or *rate*) g_n if $\mathcal{R}_n(P^*) = O_p(g_n)$. The definition of the order of risk consistency involves the true model P. Intuitively, we expect that as $n \to \infty$, the estimated distribution P^* converges to the projection $\widetilde{P}$ so $\mathcal{R}_n(P^*) \to 0$ in probability.

6.1 Error Exponent for Risk Consistency

In this subsection, we consider a fixed distribution P and state consistency results in terms of the event $\mathcal{C}_{n,\delta}$. Consequently, the model size d and the number of edges k are fixed. This lends insight into deriving results for the order of the risk consistency and provides intuition for the high-dimensional scenario in Section 6.2.

Theorem 8 (Error Exponent for δ-Risk Consistency) *For* CLThres, *there exists a constant* $\delta_0 > 0$ *such that for all* $0 < \delta < \delta_0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P^n(\mathcal{C}_{n,\delta}) \leq -\delta. \tag{21}$$

The corresponding lower bound is

$$\liminf_{n\to\infty} \frac{1}{n} \log P^n(\mathcal{C}_{n,\delta}) \geq -\delta d. \tag{22}$$

The theorem states that if δ is sufficiently small, the decay rate of the probability of $\mathcal{C}_{n,\delta}$ is exponential, hence clearly CLThres is δ-risk consistent. Furthermore, the bounds on the error exponent associated to the event $\mathcal{C}_{n,\delta}$ are *independent* of the parameters of P and only depend on δ and the dimensionality d. Intuitively, (21) is true because if we want the risk of P^* to be at most δd, then each of the empirical pairwise marginals $\widehat{P}_{i,j}$ should be δ-close to the true pairwise marginal $\widetilde{P}_{i,j}$. Note also that for $\mathcal{C}_{n,\delta}$ to occur with high probability, the edge set does not need to be estimated correctly so there is no dependence on k.

6.2 The High-Dimensional Setting

We again consider the high-dimensional setting where the tuple of parameters (n, d_n, k_n) tend to infinity and we have a sequence of learning problems indexed by the number of data points n. We again assume that (13) and (14) hold and derive sufficient conditions under which the probability of the event $\mathcal{C}_{n,\delta}$ tends to zero for a sequence of d-variate distributions $\{P^{(d)} \in \mathcal{P}(\mathcal{X}^d)\}_{d\in\mathbb{N}}$. The proof of Theorem 8 leads immediately to the following corollary.

Corollary 9 (δ-Risk Consistency Scaling Law) *Let $\delta > 0$ be a sufficiently small constant and $a \in (0, \delta)$. If the number of variables in the sequence of models $\{P^{(d)}\}_{d\in\mathbb{N}}$ satisfies $d_n = o(\exp(an))$, then* CLThres *is δ-risk consistent for $\{P^{(d)}\}_{d\in\mathbb{N}}$.*

Interestingly, this sufficient condition on how number of variables d should scale with n for consistency is very similar to Theorem 5. In particular, if d is polynomial in n, then CLThres is both structurally consistent as well as δ-risk consistent. We now study the order of the risk consistency of CLThres as the model size d grows.

Theorem 10 (Order of Risk Consistency) *The risk of the sequence of estimated distributions $\{(P^{(d)})^*\}_{d\in\mathbb{N}}$ with respect to $\{P^{(d)}\}_{d\in\mathbb{N}}$ satisfies*

$$\mathcal{R}_n((P^{(d)})^*) = O_p\left(\frac{d\log d}{n^{1-\gamma}}\right), \tag{23}$$

for every $\gamma > 0$, that is, the risk consistency for CLThres *is of order $(d\log d)/n^{1-\gamma}$.*

Note that since this result is stated for the high-dimensional case, $d = d_n$ is a sequence in n but the dependence on n is suppressed for notational simplicity in (23). This result implies that if $d = o(n^{1-2\gamma})$ then CLThres is risk consistent, that is, $\mathcal{R}_n((P^{(d)})^*) \to 0$ in probability. Note that this result is not the same as the conclusion of Corollary 9 which refers to the probability that the average risk is greater than a fixed constant δ. Also, the order of convergence given in (23) does not depend on the true number of edges k. This is a consequence of the result in (21) where the upper bound on the exponent associated to the event $\mathcal{C}_{n,\delta}$ is independent of the parameters of P.

The order of the risk, or equivalently the rate of convergence of the estimated distribution to the forest projection, is almost linear in the number of variables d and inversely proportional to n. We provide three intuitive reasons to explain why this is plausible: (i) the dimension of the sufficient statistics in a tree-structured graphical model is of order $O(d)$, (ii) the ML estimator of the natural parameters of an exponential family converge to their true values at the rate of $O_p(n^{-1/2})$ (Serfling, 1980, Sec. 4.2.2), and (iii) locally, the KL-divergence behaves like the square of a weighted Euclidean norm of the natural parameters (Cover and Thomas, 2006, Equation (11.320)).

We now compare Theorem 10 to the corresponding results in Liu et al. (2011). In these recent papers, it was shown that by modeling the bivariate densities $\widehat{P}_{i,j}$ as functions from a Hölder class with exponent $\alpha > 0$ and using a reconstruction algorithm based on validation on a held-out data set, the risk decays at a rate[10] of $\widetilde{O}_p(dn^{-\alpha/(1+2\alpha)})$, which is slower than the order of risk consistency in (23). This is due to the need to compute the bivariate densities via kernel density estimation. Furthermore, we model the pairwise joint distributions as discrete probability mass functions and not continuous probability density functions, hence there is no dependence on Hölder exponents.

10. The $\widetilde{O}_p(\cdot)$ notation suppresses the dependence on factors involving logarithms.

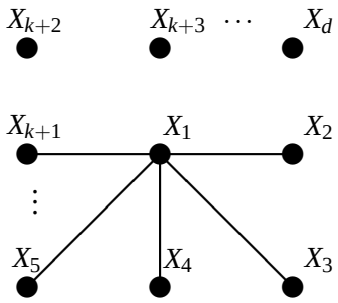

Figure 2: The forest-structured distribution Markov on d nodes and k edges. Variables $X_{k+1},\ldots,X_d$ are not connected to the main star graph.

7. Numerical Results

In this section, we perform numerical simulations on synthetic and real data sets to study the effect of a finite number of samples on the probability of the event $\mathcal{A}_n$ defined in (4). Recall that this is the error event associated to an incorrect learned structure.

7.1 Synthetic Data Sets

In order to compare our estimate to the ground truth graph, we learn the structure of distributions that are Markov on the forest shown in Figure 2. Thus, a subgraph (nodes $1,\ldots,k+1$) is a (connected) star while nodes $k+2,\ldots,d-1$ are not connected to the star. Each random variable X_j takes on values from a binary alphabet $\mathcal{X}=\{0,1\}$. Furthermore, $P_j(x_j)=0.5$ for $x_j=0,1$ and all $j\in V$. The conditional distributions are governed by the "binary symmetric channel":

$$P_{j|1}(x_j|x_1)=\begin{cases}0.7 & x_j=x_1\\ 0.3 & x_j\neq x_1\end{cases}$$

for $j=2,\ldots,k+1$. We further assume that the regularization sequence is given by $\varepsilon_n:=n^{-\beta}$ for some $\beta\in(0,1)$. Recall that this sequence satisfies the conditions in (6). We will vary β in our experiments to observe its effect on the overestimation and underestimation errors.

In Figure 3, we show the simulated error probability as a function of the sample size n for a $d=101$ node graph (as in Figure 2) with $k=50$ edges. The error probability is estimated based on 30,000 independent runs of CLThres (over different data sets $\mathbf{x}^n$). We observe that the error probability is minimized when $\beta\approx 0.625$. Figure 4 show the simulated overestimation and underestimation errors for this experiment. We see that as $\beta\to 0$, the overestimation (resp., underestimation) error is likely to be small (resp., large) because the regularization sequence ε_n is large. When the number of samples is relatively small as in this experiment, both types of errors contribute significantly to the overall error probability. When $\beta\approx 0.625$, we have the best tradeoff between overestimation and underestimation for this particular experimental setting.

Even though we mentioned that β in (10) should be chosen to be close to zero so that the error probability of structure learning decays as rapidly as possible, this example demonstrates that when given a finite number of samples, β should be chosen to balance the overestimation and underestimation errors. This does not violate Theorem 3 since Theorem 3 is an asymptotic result and refers to the typical way an error occurs in the limit as $n\to\infty$. Indeed, when the number of

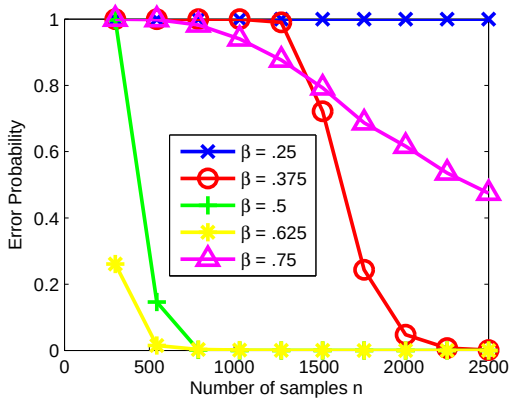

Figure 3: The error probability of structure learning for $\beta \in (0,1)$.

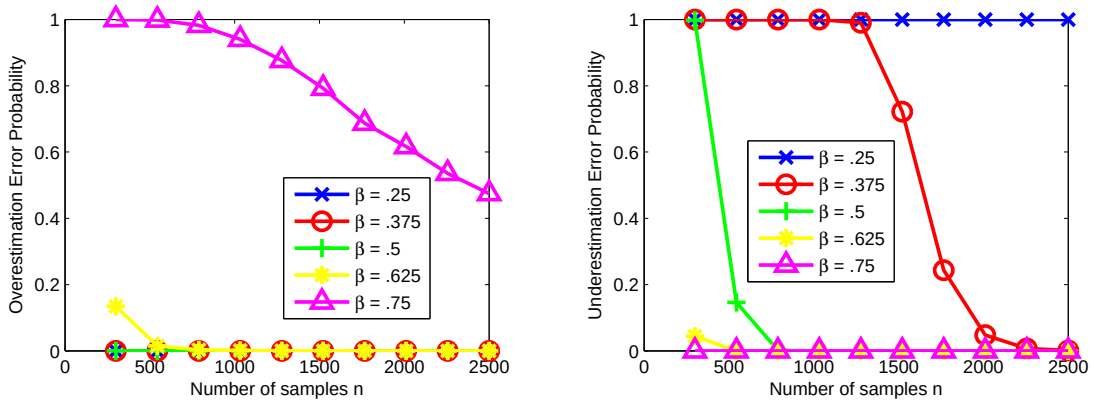

Figure 4: The overestimation and underestimation errors for $\beta \in (0,1)$.

samples is very large, it is shown that the overestimation error dominates the overall probability of error and so one should choose β to be close to zero. The question of how best to select optimal β when given only a finite number of samples appears to be a challenging one. We use cross-validation as a proxy to select this parameter for the real-world data sets in the next section.

In Figure 5, we fix the value of β at 0.625 and plot the KL-divergence $D(P||P^*)$ as a function of the number of samples. This is done for a forest-structured distribution P whose graph is shown in Figure 2 and with $d = 21$ nodes and $k = 10$ edges. The mean, minimum and maximum KL-divergences are computed based on 50 independent runs of CLThres. We see that $\log D(P||P^*)$ decays linearly. Furthermore, the slope of the mean curve is approximately -1, which is in agreement with (23). This experiment shows that if we want to reduce the KL-divergence between the estimated and true models by a constant factor $A > 0$, we need to increase the number of samples by roughly the same factor A.

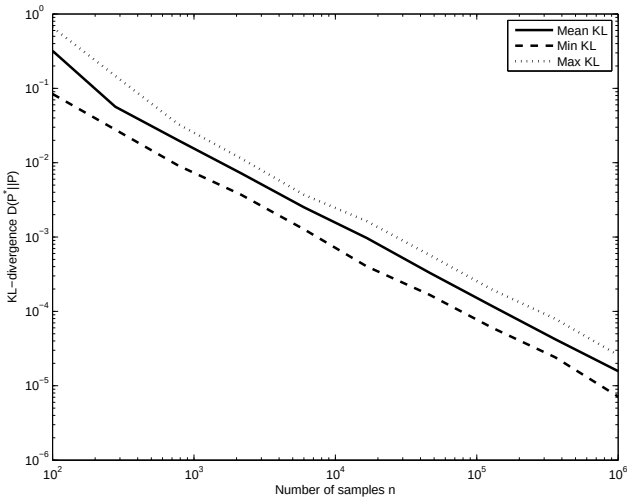

Figure 5: Mean, minimum and maximum (across 50 different runs) of the KL-divergence between the estimated model P^* and the true model P for a $d = 21$ node graph with $k = 10$ edges.

7.2 Real Data Sets

We now demonstrate how well forests-structured distributions can model two real data sets[11] which are obtained from the UCI Machine Learning Repository (Newman et al., 1998). The first data set we used is known as the SPECT Heart data set, which describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images on normal and abnormal patients. The data set contains $d = 22$ binary variables and $n = 80$ training samples. There are also 183 test samples. We learned a forest-structured distributions using the 80 training samples for different $\beta \in (0,1)$ and subsequently computed the log-likelihood of both the training and test samples. The results are displayed in Figure 6. We observe that, as expected, the log-likelihood of the training samples increases monotonically with β. This is because there are more edges in the model when β is large improving the modeling ability. However, we observe that there is overfitting when β is large as evidenced by the decrease in the log-likelihood of the 183 test samples. The optimal value of β in terms of the log-likelihood for this data set is ≈ 0.25, but surprisingly an approximation with an empty graph[12] also yields a high log-likelihood score on the test samples. This implies that according to the available data, the variables are nearly independent. The forest graph for $\beta = 0.25$ is shown in Figure 7(a) and is very sparse.

The second data set we used is the Statlog Heart data set containing physiological measurements of subjects with and without heart disease. There are 270 subjects and $d = 13$ discrete and continuous attributes, such as gender and resting blood pressure. We quantized the continuous attributes into two bins. Those measurements that are above the mean are encoded as 1 and those below the mean as 0. Since the raw data set is not partitioned into training and test sets, we learned forest-structured models based on a randomly chosen set of $n = 230$ training samples and then computed

11. These data sets are typically employed for binary classification but we use them for modeling purposes.
12. When $\beta = 0$ we have an empty graph because all empirical mutual information quantities in this experiment are smaller than 1.

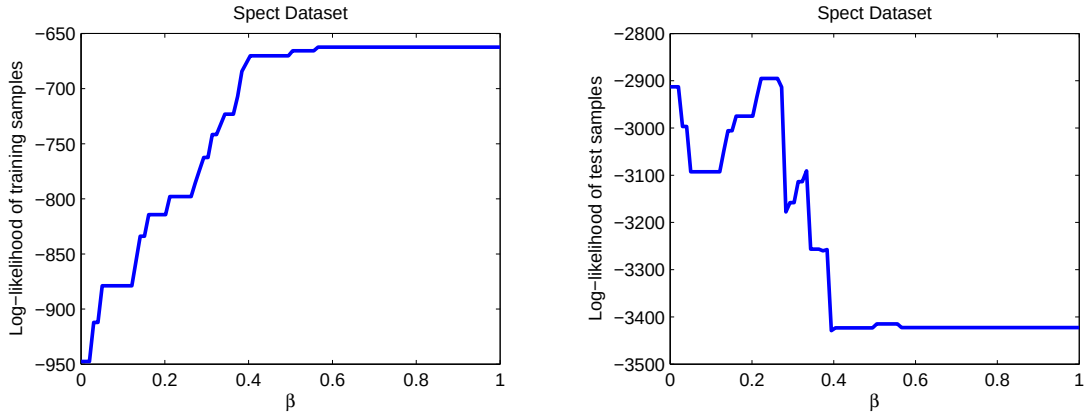

Figure 6: Log-likelihood scores on the SPECT data set

the log-likelihood of these training and 40 remaining test samples. We then chose an additional 49 randomly partitioned training and test sets and performed the same learning task and computation of log-likelihood scores. The mean of the log-likelihood scores over these 50 runs is shown in Figure 8. We observe that the log-likelihood on the test set is maximized at $\beta \approx 0.53$ and the tree approximation ($\beta \approx 1$) also yields a high likelihood score. The forest learned when $\beta = 0.53$ is shown in Figure 7(b). Observe that two nodes (ECG and Cholesterol) are disconnected from the main graph because their mutual information values with other variables are below the threshold. In contrast, HeartDisease, the label for this data set, has the highest degree, that is, it influences and is influenced by many other covariates. The strengths of the interactions between HeartDisease and its neighbors are also strong as evidenced by the bold edges.

From these experiments, we observe that some data sets can be modeled well as proper forests with very few edges while others are better modeled as distributions that are almost tree-structured (see Figure 7). Also, we need to choose β carefully to balance between data fidelity and overfitting. In contrast, our asymptotic result in Theorem 3 says that ε_n should be chosen according to (6) so that we have structural consistency. When the number of data points n is large, β in (10) should be chosen to be small to ensure that the learned edge set is equal to the true one (assuming the underlying model is a forest) with high probability as the overestimation error dominates.

8. Conclusion

In this paper, we proposed an efficient algorithm CLThres for learning the parameters and the structure of forest-structured graphical models. We showed that the asymptotic error rates associated to structure learning are nearly optimal. We also provided the rate at which the error probability of structure learning tends to zero and the order of the risk consistency. One natural question that arises from our analyses is whether β in (10) can be selected automatically in the finite-sample regime. There are many other open problems that could possibly leverage on the proof techniques employed here. For example, we are currently interested to analyze the learning of general graphical models using similar thresholding-like techniques on the empirical correlation coefficients. The analyses could potentially leverage on the use of the method of types. We are currently exploring this promising line of research.

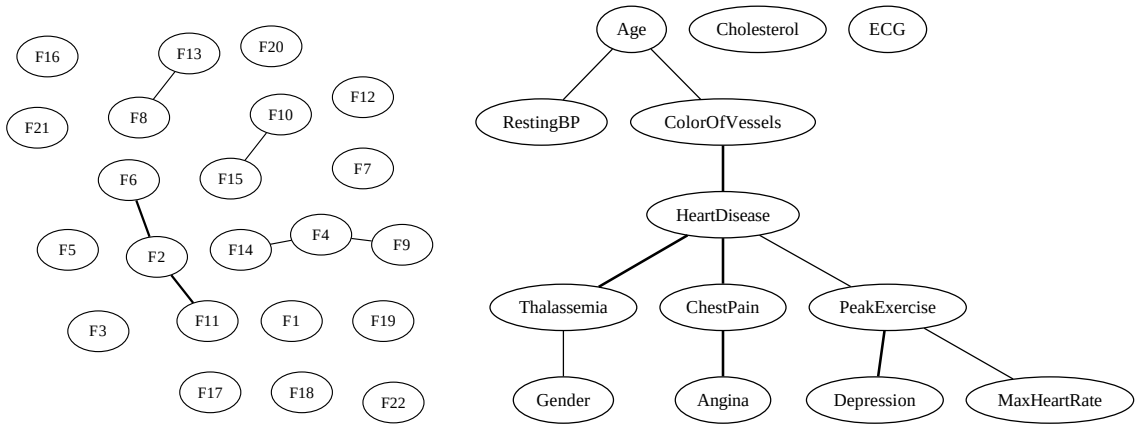

Figure 7: Learned forest graph of the (a) SPECT data set for $\beta = 0.25$ and (b) HEART data set for $\beta = 0.53$. Bold edges denote higher mutual information values. The features names are not provided for the SPECT data set.

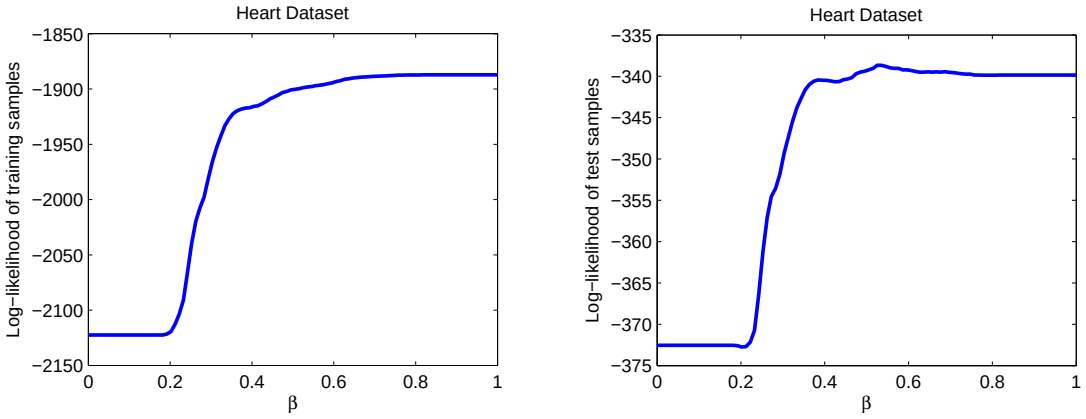

Figure 8: Log-likelihood scores on the HEART data set

Acknowledgments

This work was supported by a AFOSR funded through Grant FA9559-08-1-1080, a MURI funded through ARO Grant W911NF-06-1-0076 and a MURI funded through AFOSR Grant FA9550-06-1-0324. V. Tan is also funded by A*STAR, Singapore. The authors would like to thank Sanjoy Mitter, Lav Varshney, Matt Johnson and James Saunderson for discussions. The authors would also like to thank Rui Wu (UIUC) for pointing out an error in the proof of Theorem 3.

Appendix A. Proof of Proposition 2

Proof (*Sketch*) The proof of this result hinges on the fact that both the overestimation and underestimation errors decay to zero exponentially fast when the threshold is chosen to be $I_{\min}/2$. This threshold is able to differentiate between true edges (with MI larger than $I_{\min}$) from non-edges (with MI smaller than $I_{\min}$) with high probability for n sufficiently large. The error for learning the top k edges of the forest also decays exponentially fast (Tan et al., 2011). Thus, (5) holds. The full details of the proof follow in a straightforward manner from Appendix B which we present next. ■

Appendix B. Proof of Theorem 3

Define the event $\mathcal{B}_n := \{\widehat{E}_k \neq E_P\}$, where $\widehat{E}_k = \{\widehat{e}_1, \ldots, \widehat{e}_k\}$ is the set of top k edges (see Step 3 of CLThres for notation). This is the Chow-Liu error as mentioned in Section 4.3. Let $\mathcal{B}_n^c$ denote the complement of $\mathcal{B}_n$. Note that in $\mathcal{B}_n^c$, the estimated edge set depends on k, the true model order, which is *a-priori* unknown to the learner. Further define the constant

$$K_P := \lim_{n\to\infty} -\frac{1}{n}\log P^n(\mathcal{B}_n). \tag{24}$$

In other words, K_P is the error exponent for learning the forest structure incorrectly assuming the true model order k is known and Chow-Liu terminates after the addition of exactly k edges in the MWST procedure (Kruskal, 1956). The existence of the limit in (24) and the positivity of K_P follow from the main results in Tan et al. (2011).

We first state a result which relies on the Gallager-Fano bound (Fano, 1961, pp. 24). The proof will be provided at the end of this appendix.

Lemma 11 (Reduction to Model Order Estimation) *For every $\eta \in (0, K_P)$, there exists a $N \in \mathbb{N}$ sufficiently large such that for every $n > N$, the error probability $P^n(\mathcal{A}_n)$ satisfies*

$$(1-\eta)P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c) \leq P^n(\mathcal{A}_n) \tag{25}$$

$$\leq P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c) + 2\exp(-n(K_P - \eta)). \tag{26}$$

Proof (*of Theorem 3*) We will prove (i) the upper bound in (8) (ii) the lower bound in (7) and (iii) the exponential rate of decay in the case of trees (9).

B.1 Proof of Upper Bound in Theorem 3

We now bound the error probability $P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c)$ in (26). Using the union bound,

$$P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c) \leq P^n(\widehat{k}_n > k|\mathcal{B}_n^c) + P^n(\widehat{k}_n < k|\mathcal{B}_n^c). \tag{27}$$

The first and second terms are known as the *overestimation* and *underestimation* errors respectively. We will show that the underestimation error decays exponentially fast. The overestimation error decays only subexponentially fast and so its rate of decay dominates the overall rate of decay of the error probability for structure learning.

B.1.1 UNDERESTIMATION ERROR

We now bound these terms staring with the underestimation error. By the union bound,

$$\begin{aligned} P^n(\widehat{k}_n < k|\mathcal{B}_n^c) &\leq (k-1) \max_{1\leq j\leq k-1} P^n(\widehat{k}_n = j|\mathcal{B}_n^c) \\ &= (k-1)P^n(\widehat{k}_n = k-1|\mathcal{B}_n^c), \end{aligned} \tag{28}$$

where (28) follows because $P^n(\widehat{k}_n = j|\mathcal{B}_n^c)$ is maximized when $j = k-1$. This is because if, to the contrary, $P^n(\widehat{k}_n = j|\mathcal{B}_n^c)$ were to be maximized at some other $j \leq k-2$, then there exists at least two edges, call them $e_1, e_2 \in E_P$ such that events $\mathcal{E}_1 := \{I(\widehat{P}_{e_1}) \leq \varepsilon_n\}$ and $\mathcal{E}_2 := \{I(\widehat{P}_{e_2}) \leq \varepsilon_n\}$ occur. The probability of this joint event is smaller than the individual probabilities, that is, $P^n(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \min\{P^n(\mathcal{E}_1), P^n(\mathcal{E}_2)\}$. This is a contradiction.

By the rule for choosing $\widehat{k}_n$ in (3), we have the upper bound

$$P^n(\widehat{k}_n = k-1|\mathcal{B}_n^c) = P^n(\exists e \in E_P \text{ s.t. } I(\widehat{P}_e) \leq \varepsilon_n) \leq k \max_{e\in E_P} P^n(I(\widehat{P}_e) \leq \varepsilon_n), \tag{29}$$

where the inequality follows from the union bound. Now, note that if $e \in E_P$, then $I(P_e) > \varepsilon_n$ for n sufficiently large (since $\varepsilon_n \to 0$). Thus, by Sanov's theorem (Cover and Thomas, 2006, Ch. 11), $P^n(I(\widehat{P}_e) \leq \varepsilon_n)$ can be upper bounded as

$$P^n(I(\widehat{P}_e) \leq \varepsilon_n) \leq (n+1)^{r^2} \exp\left(-n \min_{Q\in\mathcal{P}(\mathcal{X}^2)} \{D(Q||P_e) : I(Q) \leq \varepsilon_n\}\right). \tag{30}$$

Define the good rate function (Dembo and Zeitouni, 1998) in (30) to be $L : \mathcal{P}(\mathcal{X}^2) \times [0,\infty) \to [0,\infty)$, which is given by

$$L(P_e; a) := \min_{Q\in\mathcal{P}(\mathcal{X}^2)} \{D(Q||P_e) : I(Q) \leq a\}. \tag{31}$$

Clearly, $L(P_e; a)$ is continuous in a. Furthermore it is monotonically decreasing in a for fixed P_e. Thus by using the continuity of $L(P_e;\cdot)$ we can assert: To every $\eta > 0$, there exists a $N \in \mathbb{N}$ such that for all $n > N$ we have $L(P_e;\varepsilon_n) > L(P_e;0) - \eta$. As such, we can further upper bound the error probability in (30) as

$$P^n(I(\widehat{P}_e) \leq \varepsilon_n) \leq (n+1)^{r^2} \exp\left(-n(L(P_e;0) - \eta)\right). \tag{32}$$

By using the fact that $I_{\min} > 0$, the exponent $L(P_e;0) > 0$ and thus, we can put the pieces in (28), (29) and (32) together to show that the underestimation error is upper bounded as

$$P^n(\widehat{k}_n < k|\mathcal{B}_n^c) \leq k(k-1)(n+1)^{r^2} \exp\left(-n \min_{e\in E_P}(L(P_e;0) - \eta)\right). \tag{33}$$

Hence, if k is constant, the underestimation error $P^n(\widehat{k}_n < k|\mathcal{B}_n^c)$ decays to zero exponentially fast as $n \to \infty$, that is, the normalized logarithm of the underestimation error can be bounded as

$$\limsup_{n\to\infty} \frac{1}{n} \log P^n(\widehat{k}_n < k|\mathcal{B}_n^c) \leq -\min_{e\in E_P}(L(P_e;0) - \eta).$$

The above statement is now independent of n. Hence, we can take the limit as $\eta \to 0$ to conclude that:

$$\limsup_{n\to\infty} \frac{1}{n} \log P^n(\widehat{k}_n < k | \mathcal{B}_n^c) \leq -L_P. \tag{34}$$

The exponent $L_P := \min_{e\in E_P} L(P_e;0)$ is positive because we assumed that the model is minimal and so $I_{\min} > 0$, which ensures the positivity of the rate function $L(P_e;0)$ for each true edge $e \in E_P$.

B.1.2 OVERESTIMATION ERROR

Bounding the overestimation error is harder. It follows by first applying the union bound:

$$\begin{aligned} P^n(\widehat{k}_n > k | \mathcal{B}_n^c) &\leq (d-k-1) \max_{k+1\leq j\leq d-1} P^n(\widehat{k}_n = j | \mathcal{B}_n^c) \\ &= (d-k-1) P^n(\widehat{k}_n = k+1 | \mathcal{B}_n^c), \end{aligned} \tag{35}$$

where (35) follows because $P^n(\widehat{k}_n = j|\mathcal{B}_n^c)$ is maximized when $j = k+1$ (by the same argument as for the underestimation error). Apply the union bound again, we have

$$P^n(\widehat{k}_n = k+1 | \mathcal{B}_n^c) \leq (d-k-1) \max_{e\in V\times V: I(P_e)=0} P^n(I(\widehat{P}_e) \geq \varepsilon_n). \tag{36}$$

From (36), it suffices to bound $P^n(I(\widehat{P}_e) \geq \varepsilon_n)$ for any pair of independent random variables (X_i, X_j) and $e = (i,j)$. We proceed by applying the upper bound in Sanov's theorem (Cover and Thomas, 2006, Ch. 11) to $P^n(I(\widehat{P}_e) \geq \varepsilon_n)$ which yields

$$P^n(I(\widehat{P}_e) \geq \varepsilon_n) \leq (n+1)^{r^2} \exp\left(-n \min_{Q\in\mathcal{P}(\mathcal{X}^2)} \{ D(Q||P_e) : I(Q) \geq \varepsilon_n \} \right), \tag{37}$$

for all $n \in \mathbb{N}$. Our task now is to lower bound the good rate function in (37), which we denote as $M : \mathcal{P}(\mathcal{X}^2) \times [0,\infty) \to [0,\infty)$:

$$M(P_e; b) := \min_{Q\in\mathcal{P}(\mathcal{X}^2)} \{ D(Q||P_e) : I(Q) \geq b \}. \tag{38}$$

Note that $M(P_e;b)$ is monotonically increasing and continuous in b for fixed P_e. Because the sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ tends to zero, when n is sufficiently large, ε_n is arbitrarily small and we are in the so-called *very-noisy learning regime* (Borade and Zheng, 2008; Tan et al., 2011), where the optimizer to (38), denoted as Q_n^*, is very close to P_e. See Figure 9.

Thus, when n is large, the KL-divergence and mutual information can be approximated as

$$D(Q_n^*||P_e) = \frac{1}{2}\mathbf{v}^T \mathbf{\Pi}_e \mathbf{v} + o(\|\mathbf{v}\|^2), \tag{39}$$

$$I(Q_n^*) = \frac{1}{2}\mathbf{v}^T \mathbf{H}_e \mathbf{v} + o(\|\mathbf{v}\|^2), \tag{40}$$

where[13] $\mathbf{v} := \mathrm{vec}(Q_n^*) - \mathrm{vec}(P_e) \in \mathbb{R}^{r^2}$. The $r^2 \times r^2$ matrices $\mathbf{\Pi}_e$ and $\mathbf{H}_e$ are defined as

$$\mathbf{\Pi}_e := \mathrm{diag}(1/\mathrm{vec}(P_e)), \tag{41}$$

$$\mathbf{H}_e := \nabla^2_{\mathrm{vec}(Q)} I(\mathrm{vec}(Q))\big|_{Q=P_e}. \tag{42}$$

13. The operator $\mathrm{vec}(\mathbf{C})$ vectorizes a matrix in a column oriented way. Thus, if $\mathbf{C} \in \mathbb{R}^{l\times l}$, $\mathrm{vec}(\mathbf{C})$ is a length-l^2 vector with the columns of $\mathbf{C}$ stacked one on top of another ($\mathbf{C}(:)$ in Matlab).

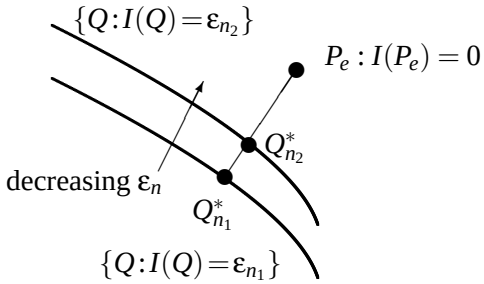

Figure 9: As $\varepsilon_n \to 0$, the projection of P_e onto the constraint set $\{Q : I(Q) \geq \varepsilon_n\}$, denoted Q^*_n (the optimizer in (38)), approaches P_e. The approximations in (39) and (40) become increasingly accurate as ε_n tends to zero. In the figure, $n_2 > n_1$ and $\varepsilon_{n_1} > \varepsilon_{n_2}$ and the curves are the (sub-)manifold of distributions such that the mutual information is constant, that is, the mutual information level sets.

In other words, $\mathbf{\Pi}_e$ is the diagonal matrix that contains the reciprocal of the elements of $\mathrm{vec}(P_e)$ on its diagonal. $\mathbf{H}_e$ is the Hessian[14] of $I(\mathrm{vec}(Q^*_n))$, viewed as a function of $\mathrm{vec}(Q^*_n)$ and evaluated at P_e. As such, the exponent for overestimation in (38) can be approximated by a *quadratically constrained quadratic program* (QCQP), where $\mathbf{z} := \mathrm{vec}(Q) - \mathrm{vec}(P_e)$:

$$\widetilde{M}(P_e;\varepsilon_n) = \min_{\mathbf{z}\in\mathbb{R}^{r^2}} \frac{1}{2}\mathbf{z}^T \mathbf{\Pi}_e \mathbf{z},$$
$$\text{subject to} \quad \frac{1}{2}\mathbf{z}^T \mathbf{H}_e \mathbf{z} \geq \varepsilon_n, \quad \mathbf{z}^T \mathbf{1} = 0. \tag{43}$$

Note that the constraint $\mathbf{z}^T\mathbf{1} = 0$ does not necessarily ensure that Q is a probability distribution so $\widetilde{M}(P_e;\varepsilon_n)$ is an approximate lower bound to the true rate function $M(P_e;\varepsilon_n)$, defined in (38). We now argue that the approximate rate function $\widetilde{M}$ in (43), can be lower bounded by a quantity that is proportional to ε_n. To show this, we resort to Lagrangian duality (Bertsekas, 1999, Ch. 5). It can easily be shown that the *Lagrangian dual* corresponding to the primal in (43) is

$$g(P_e;\varepsilon_n) := \varepsilon_n \max_{\mu\geq 0}\{\mu : \mathbf{\Pi}_e \succeq \mu \mathbf{H}_e\}. \tag{44}$$

We see from (44) that $g(P_e;\varepsilon_n)$ is proportional to ε_n. By weak duality (Bertsekas, 1999, Proposition 5.1.3), any dual feasible solution provides a lower bound to the primal, that is,

$$g(P_e;\varepsilon_n) \leq \widetilde{M}(P_e;\varepsilon_n). \tag{45}$$

Note that strong duality (equality in (45)) does not hold in general due in part to the non-convex constraint set in (43). Interestingly, our manipulations lead lower bounding $\widetilde{M}$ by (44), which is a (convex) semidefinite program (Vandenberghe and Boyd, 1996).

Now observe that the approximations in (39) and (40) are accurate in the limit of large n because the optimizing distribution Q^*_n becomes increasingly close to P_e. By continuity of the optimization

14. The first two terms in the Taylor expansion of the mutual information $I(\mathrm{vec}(Q^*_n))$ in (40) vanish because (i) $I(P_e) = 0$ and (ii) $(\mathrm{vec}(Q^*_n) - \mathrm{vec}(P_e))^T \nabla_{\mathrm{vec}(Q)} I(\mathrm{vec}(P_e)) = 0$. Indeed, if we expand $I(\mathrm{vec}(Q))$ around a product distribution, the constant and linear terms vanish (Borade and Zheng, 2008). Note that $\mathbf{H}_e$ in (42) is an indefinite matrix because $I(\mathrm{vec}(Q))$ is not convex.

problems in (perturbations of) the objective and the constraints, $\widetilde{M}(P_e;\varepsilon_n)$ and $M(P_e;\varepsilon_n)$ are close when n is large, that is,

$$\lim_{n\to\infty}\frac{\widetilde{M}(P_e;\varepsilon_n)}{M(P_e;\varepsilon_n)}=1. \tag{46}$$

This can be seen from (39) in which the ratio of the KL-divergence to its approximation $\mathbf{v}^T\mathbf{\Pi}_e\mathbf{v}/2$ is unity in the limit as $\|\mathbf{v}\|\to 0$. The same holds true for the ratio of the mutual information to its approximation $\mathbf{v}^T\mathbf{H}_e\mathbf{v}/2$ in (40). By applying the continuity statement in (46) to the upper bound in (37), we can conclude that for every $\eta>0$, there exists a $N\in\mathbb{N}$ such that

$$P^n(I(\widehat{P}_e)\geq\varepsilon_n)\leq(n+1)^{r^2}\exp\left(-n\widetilde{M}(P_e;\varepsilon_n)(1-\eta)\right),$$

for all $n>N$. Define the constant

$$c_P:=\min_{e\in V\times V:I(P_e)=0}\ \max_{\mu\geq 0}\{\mu:\mathbf{\Pi}_e\succeq\mu\mathbf{H}_e\}. \tag{47}$$

By (44), (45) and the definition of c_P in (47),

$$P^n(I(\widehat{P}_e)\geq\varepsilon_n)\leq(n+1)^{r^2}\exp\left(-n\varepsilon_n c_P(1-\eta)\right). \tag{48}$$

Putting (35), (36) and (48) together, we see that the overestimation error

$$P^n(\widehat{k}_n>k|\mathcal{B}_n^c)\leq(d-k-1)^2(n+1)^{r^2}\exp\left(-n\varepsilon_n c_P(1-\eta)\right). \tag{49}$$

Note that the above probability tends to zero by the assumption that $n\varepsilon_n/\log n\to\infty$ in (6). Thus, we have consistency overall (since the underestimation, Chow-Liu and now the overestimation errors all tend to zero). Thus, by taking the normalized logarithm (normalized by $n\varepsilon_n$), the lim sup in n (keeping in mind that d and k are constant), we conclude that

$$\limsup_{n\to\infty}\frac{1}{n\varepsilon_n}\log P^n(\widehat{k}_n>k|\mathcal{B}_n^c)\leq -c_P(1-\eta). \tag{50}$$

Now by take $\eta\to 0$, it remains to prove that $c_P=1$ for all P. For this purpose, it suffices to show that the optimal solution to the optimization problem in (44), denoted μ^*, is equal to one for all $\mathbf{\Pi}_e$ and $\mathbf{H}_e$. Note that μ^* can be expressed in terms of eigenvalues:

$$\mu^*=\left(\max\left\{\mathrm{eig}(\mathbf{\Pi}_e^{-1/2}\mathbf{H}_e\mathbf{\Pi}_e^{-1/2})\right\}\right)^{-1}, \tag{51}$$

where $\mathrm{eig}(\mathbf{A})$ denotes the set of real eigenvalues of the symmetric matrix $\mathbf{A}$. By using the definitions of $\mathbf{\Pi}_e$ and $\mathbf{H}_e$ in (41) and (42) respectively, we can verify that the matrix $\mathbf{I}-\mathbf{\Pi}_e^{-1/2}\mathbf{H}_e\mathbf{\Pi}_e^{-1/2}$ is positive semidefinite with an eigenvalue at zero. This proves that the largest eigenvalue of $\mathbf{\Pi}_e^{-1/2}\mathbf{H}_e\mathbf{\Pi}_e^{-1/2}$ is one and hence from (51), $\mu^*=1$. The proof of the upper bound in (8) is completed by combining the estimates in (26), (34) and (50).

B.2 Proof of Lower Bound in Theorem 3

The key idea is to bound the overestimation error using a modification of the lower bound in Sanov's theorem. Denote the set of types supported on a finite set $\mathcal{Y}$ with denominator n as $\mathcal{P}_n(\mathcal{Y})$ and the *type class* of a distribution $Q \in \mathcal{P}_n(\mathcal{Y})$ as

$$\mathsf{T}_n(Q) := \{y^n \in \mathcal{Y}^n : \widehat{P}(\cdot; y^n) = Q(\cdot)\},$$

where $\widehat{P}(\cdot; y^n)$ is the empirical distribution of the sequence $y^n = (y_1, \ldots, y_n)$. The following bounds on the type class are well known (Cover and Thomas, 2006, Ch. 11).

Lemma 12 (Probability of Type Class) *For any $Q \in \mathcal{P}_n(\mathcal{Y})$ and any distribution P, the probability of the type class $\mathsf{T}_n(Q)$ under P^n satisfies:*

$$(n+1)^{-|\mathcal{Y}|} \exp(-nD(Q||P)) \leq P^n(\mathsf{T}_n(Q)) \leq \exp(-nD(Q||P)). \tag{52}$$

To prove the lower bound in (7), assume that $k < d-1$ and note that the error probability $P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c)$ can be lower bounded by $P^n(I(\widehat{P}_e) \geq \varepsilon_n)$ for any node pair e such that $I(P_e) = 0$. We seek to lower bound the latter probability by appealing to (52). Now choose a sequence of distributions $Q^{(n)} \in \{Q \in \mathcal{P}_n(\mathcal{X}^2) : I(Q) \geq \varepsilon_n\}$ such that

$$\lim_{n\to\infty} \left| M(P_e; \varepsilon_n) - D(Q^{(n)}||P_e) \right| = 0.$$

This is possible because the set of types is dense in the probability simplex (Dembo and Zeitouni, 1998, Lemma 2.1.2(b)). Thus,

$$\begin{aligned} P^n(I(\widehat{P}_e) \geq \varepsilon_n) &= \sum_{Q \in \mathcal{P}_n(\mathcal{X}^2): I(Q) \geq \varepsilon_n} P^n(\mathsf{T}_n(Q)) \\ &\geq P^n(\mathsf{T}_n(Q^{(n)})) \\ &\geq (n+1)^{-r^2} \exp(-nD(Q^{(n)}||P_e)), \end{aligned} \tag{53}$$

where (53) follows from the lower bound in (52). Note from (46) that the following convergence holds: $|\widetilde{M}(P_e; \varepsilon_n) - M(P_e; \varepsilon_n)| \to 0$. Using this and the fact that if $|a_n - b_n| \to 0$ and $|b_n - c_n| \to 0$ then, $|a_n - c_n| \to 0$ (triangle inequality), we also have

$$\lim_{n\to\infty} \left| \widetilde{M}(P_e; \varepsilon_n) - D(Q^{(n)}||P_e) \right| = 0.$$

Hence, continuing the chain in (53), for any $\eta > 0$, there exists a $N \in \mathbb{N}$ such that for all $n > N$,

$$P^n(I(\widehat{P}_e) \geq \varepsilon_n) \geq (n+1)^{-r^2} \exp(-n(\widetilde{M}(P_e; \varepsilon_n) + \eta)). \tag{54}$$

Note that an upper bound for $\widetilde{M}(P_e; \varepsilon_n)$ in (43) is simply given by the objective evaluated at any feasible point. In fact, by manipulating (43), we see that the upper bound is also proportional to ε_n, that is,

$$\widetilde{M}(P_e; \varepsilon_n) \leq C_{P_e} \varepsilon_n,$$

where $C_{P_e} \in (0,\infty)$ is some constant[15] that depends on the matrices $\mathbf{\Pi}_e$ and $\mathbf{H}_e$. Define $C_P := \max_{e\in V\times V:I(P_e)=0} C_{P_e}$. Continuing the lower bound in (54), we obtain

$$P^n(I(\widehat{P}_e) \geq \varepsilon_n) \geq (n+1)^{-r^2} \exp(-n\varepsilon_n(C_P+\eta)),$$

for n sufficiently large. Now take the normalized logarithm and the lim inf to conclude that

$$\liminf_{n\to\infty} \frac{1}{n\varepsilon_n} \log P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c) \geq -(C_P+\eta). \tag{55}$$

Substitute (55) into the lower bound in (25). Now the resulting inequality is independent of n and we can take $\eta \to 0$ to complete the proof of the lower bound in Theorem 3.

B.3 Proof of the Exponential Rate of Decay for Trees in Theorem 3

For the claim in (9), note that for n sufficiently large,

$$P^n(\mathcal{A}_n) \geq \max\{(1-\eta)P^n(\widehat{k}_n \neq k_n|\mathcal{B}_n^c), P^n(\mathcal{B}_n)\}, \tag{56}$$

from Lemma 11 and the fact that $\mathcal{B}_n \subseteq \mathcal{A}_n$. Equation (56) gives us a lower bound on the error probability in terms of the Chow-Liu error $P^n(\mathcal{B}_n)$ and the underestimation and overestimation errors $P^n(\widehat{k}_n \neq k_n|\mathcal{B}_n^c)$. If $k = d-1$, the overestimation error probability is identically zero, so we only have to be concerned with the underestimation error. Furthermore, from (34) and a corresponding lower bound which we omit, the underestimation error event satisfies $P^n(\widehat{k}_n < k|\mathcal{B}_n^c) \doteq \exp(-nL_P)$. Combining this fact with the definition of the error exponent K_P in (24) and the result in (56) establishes (9). Note that the relation in (56) and our preceding upper bounds ensure that the limit in (9) exists. ■

Proof (*of Lemma 11*) We note that $P^n(\mathcal{A}_n|\widehat{k}_n \neq k) = 1$ and thus,

$$P^n(\mathcal{A}_n) \leq P^n(\widehat{k}_n \neq k) + P^n(\mathcal{A}_n|\widehat{k}_n = k). \tag{57}$$

By using the definition of K_P in (24), the second term in (57) is precisely $P^n(\mathcal{B}_n)$ therefore,

$$P^n(\mathcal{A}_n) \leq P^n(\widehat{k}_n \neq k) + \exp(-n(K_P-\eta)), \tag{58}$$

for all $n > N_1$. We further bound $P^n(\widehat{k}_n \neq k)$ by conditioning on the event $\mathcal{B}_n^c$. Thus, for $\eta > 0$,

$$\begin{aligned} P^n(\widehat{k}_n \neq k) &\leq P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c) + P^n(\mathcal{B}_n) \\ &\leq P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c) + \exp(-n(K_P-\eta)), \end{aligned} \tag{59}$$

for all $n > N_2$. The upper bound result follows by combining (58) and (59). The lower bound follows by the chain

$$\begin{aligned} P^n(\mathcal{A}_n) \geq P^n(\widehat{k}_n \neq k) &\geq P^n(\{\widehat{k}_n \neq k\} \cap \mathcal{B}_n^c) \\ &= P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c)P^n(\mathcal{B}_n^c) \geq (1-\eta)P^n(\widehat{k}_n \neq k|\mathcal{B}_n^c), \end{aligned}$$

which holds for all $n > N_3$ since $P^n(\mathcal{B}_n^c) \to 1$. Now the claims in (25) and (26) follow by taking $N := \max\{N_1, N_2, N_3\}$. ■

15. We can easily remove the constraint $\mathbf{z}^T\mathbf{1}$ in (43) by a simple change of variables to only consider those vectors in the subspace orthogonal to the all ones vector so we ignore it here for simplicity. To obtain C_{P_e}, suppose the matrix $\mathbf{W}_e$ diagonalizes $\mathbf{H}_e$, that is, $\mathbf{H}_e = \mathbf{W}_e^T\mathbf{D}_e\mathbf{W}_e$, then one can, for example, choose $C_{P_e} = \min_{i:[\mathbf{D}_e]_{i,i}>0}[\mathbf{W}_e^T\mathbf{\Pi}_e\mathbf{W}_e]_{i,i}$.

Appendix C. Proof of Corollary 4

Proof This claim follows from the fact that three errors (i) Chow-Liu error (ii) underestimation error and (iii) overestimation error behave in exactly the same way as in Theorem 3. In particular, the Chow-Liu error, that is, the error for the learning the top k edges in the forest projection model $\widetilde{P}$ decays with error exponent K_P. The underestimation error behaves as in (34) and the overestimation error as in (50). ■

Appendix D. Proof of Theorem 5

Proof Given assumptions (A1) and (A2), we claim that the underestimation exponent $L_{P^{(d)}}$, defined in (34), is uniformly bounded away from zero, that is,

$$L := \inf_{d\in\mathbb{N}} L_{P^{(d)}} = \inf_{d\in\mathbb{N}} \min_{e\in E_{P^{(d)}}} L(P_e^{(d)};0) \tag{60}$$

is positive. Before providing a formal proof, we provide a plausible argument to show that this claim is true. Recall the definition of $L(P_e;0)$ in (31). Assuming that the joint $P_e = P_{i,j}$ is close to a product distribution or equivalently if its mutual information $I(P_e)$ is small (which is the worst-case scenario),

$$L(P_e;0) \approx \min_{Q\in\mathcal{P}(\mathcal{X}^2)} \{D(P_e\,||\,Q) : I(Q) = 0\} \tag{61}$$

$$= D(P_e\,||\,P_iP_j) = I(P_e) \geq I_{\text{inf}} > 0, \tag{62}$$

where in (61), the arguments in the KL-divergence have been swapped. This is because when $Q \approx P_e$ entry-wise, $D(Q\,||\,P_e) \approx D(P_e\,||\,Q)$ in the sense that their difference is small compared to their absolute values (Borade and Zheng, 2008). In (62), we used the fact that the reverse I-projection of P_e onto the set of product distributions is P_iP_j. Since I_{inf} is constant, this proves the claim, that is, $L > 0$.

More formally, let

$$B_{\kappa'} := \{Q_{i,j} \in \mathcal{P}(\mathcal{X}^2) : Q_{i,j}(x_i,x_j) \geq \kappa', \forall x_i, x_j \in \mathcal{X}\}$$

be the set of joint distributions whose entries are bounded away from zero by $\kappa' > 0$. Now, consider a pair of joint distributions $P_e^{(d)}, \widetilde{P}_e^{(d)} \in B_{\kappa'}$ whose minimum values are uniformly bounded away from zero as assumed in (A2). Then there exists a Lipschitz constant (independent of d) $U \in (0,\infty)$ such that for all d,

$$|I(P_e^{(d)}) - I(\widetilde{P}_e^{(d)})| \leq U \|\text{vec}(P_e^{(d)}) - \text{vec}(\widetilde{P}_e^{(d)})\|_1, \tag{63}$$

where $\|\cdot\|_1$ is the vector ℓ_1 norm. In fact, $U := \max_{Q\in B_{\kappa'}} \|\nabla I(\text{vec}(Q))\|_\infty$ is the Lipschitz constant of $I(\cdot)$ which is uniformly bounded because the joint distributions $P_e^{(d)}$ and $\widetilde{P}_e^{(d)}$ are assumed to be uniformly bounded away from zero. Suppose, to the contrary, $L = 0$. Then by the definition of the infimum in (60), for every $\varepsilon > 0$, there exists a $d \in \mathbb{N}$ and a corresponding $e \in E_{P^{(d)}}$ such that if Q^* is the optimizer in (31),

$$\varepsilon > D(Q^*\,||\,P_e^{(d)}) \overset{(a)}{\geq} \frac{\|\text{vec}(P_e^{(d)}) - \text{vec}(Q^*)\|_1^2}{2\log 2} \overset{(b)}{\geq} \frac{|I(P_e^{(d)}) - I(Q^*)|^2}{(2\log 2)U^2} \overset{(c)}{\geq} \frac{I_{\text{inf}}^2}{(2\log 2)U^2},$$

where (a) follows from Pinsker's inequality (Cover and Thomas, 2006, Lemma 11.6.1), (b) is an application of (63) and the fact that if $P_e^{(d)} \in B_\kappa$ is uniformly bounded from zero (as assumed in (14)) so is the associated optimizer Q^* (i.e., in $B_{\kappa''}$ for some possibly different uniform $\kappa'' > 0$). Statement (c) follows from the definition of $I_{\inf}$ and the fact that Q^* is a product distribution, that is, $I(Q^*) = 0$. Since ε can be chosen to be arbitrarily small, we arrive at a contradiction. Thus L in (60) is positive. Finally, we observe from (33) that if $n > (3/L)\log k$ the underestimation error tends to zero because (33) can be further upper bounded as

$$P^n(\widehat{k}_n < k|\mathcal{B}_n^c) \leq (n+1)^{r^2}\exp(2\log k - nL) < (n+1)^{r^2}\exp\left(\frac{2}{3}nL - nL\right) \to 0$$

as $n \to \infty$. Take $C_2 = 3/L$ in (15).

Similarly, given the same assumptions, the error exponent for structure learning $K_{P^{(d)}}$, defined in (24), is also uniformly bounded away from zero, that is,

$$K := \inf_{d\in\mathbb{N}} K_{P^{(d)}} > 0.$$

Thus, if $n > (4/K)\log d$, the error probability associated to estimating the top k edges (event $\mathcal{B}_n$) decays to zero along similar lines as in the case of the underestimation error. Take $C_1 = 4/K$ in (15).

Finally, from (49), if $n\varepsilon_n > 2\log(d-k)$, then the overestimation error tends to zero. Since from (6), ε_n can take the form $n^{-\beta}$ for $\beta > 0$, this is equivalent to $n^{1-\beta} > 2\log(d-k)$, which is the same as the first condition in (15), namely $n > (2\log(d-k))^{1+\zeta}$. By (26) and (27), these three probabilities constitute the overall error probability when learning the sequence of forest structures $\{E_{P^{(d)}}\}_{d\in\mathbb{N}}$. Thus the conditions in (15) suffice for high-dimensional consistency. ■

Appendix E. Proof of Corollary 6

Proof First note that $k_n \in \{0,\ldots,d_n-1\}$. From (49), we see that for n sufficiently large, the sequence $h_n(P) := (n\varepsilon_n)^{-1}\log P^n(\mathcal{A}_n)$ is upper bounded by

$$-1 + \frac{2}{n\varepsilon_n}\log(d_n - k_n - 1) + \frac{r^2\log(n+1)}{n\varepsilon_n}. \tag{64}$$

The last term in (64) tends to zero by (6). Thus $h_n(P) = O((n\varepsilon_n)^{-1}\log(d_n - k_n - 1))$, where the implied constant is 2 by (64). Clearly, this sequence is maximized (resp., minimized) when $k_n = 0$ (resp., $k_n = d_n - 1$). Equation (64) also shows that the sequence h_n is monotonically decreasing in k_n. ■

Appendix F. Proof of Theorem 7

Proof We first focus on part (a). Part (b) follows in a relatively straightforward manner. Define

$$\widehat{T}_{\mathrm{MAP}}(\mathbf{x}^n) := \operatorname*{argmax}_{t\in\mathcal{T}_k^d} \mathbb{P}(T_P = t|\mathbf{x}^n)$$

to be the maximum a-posteriori (MAP) decoding rule.[16] By the optimality of the MAP rule, this lower bounds the error probability of any other estimator. Let $\mathcal{W} := \widehat{T}_{\mathrm{MAP}}((\mathcal{X}^d)^n)$ be the range of the function $\widehat{T}_{\mathrm{MAP}}$, that is, a forest $t \in \mathcal{W}$ if and only if there exists a sequence $\mathbf{x}^n$ such that $\widehat{T}_{\mathrm{MAP}} = t$. Note that $\mathcal{W} \cup \mathcal{W}^c = \mathcal{T}_k^d$. Then, consider the lower bounds:

$$\begin{aligned}
\mathbb{P}(\widehat{T} \neq T_P) &= \sum_{t \in \mathcal{T}_k^d} \mathbb{P}(\widehat{T} \neq T_P | T_P = t)\mathbb{P}(T_P = t) \\
&\geq \sum_{t \in \mathcal{W}^c} \mathbb{P}(\widehat{T} \neq T_P | T_P = t)\mathbb{P}(T_P = t) \\
&= \sum_{t \in \mathcal{W}^c} \mathbb{P}(T_P = t) = 1 - \sum_{t \in \mathcal{W}} \mathbb{P}(T_P = t) \qquad (65) \\
&= 1 - \sum_{t \in \mathcal{W}} |\mathcal{T}_k^d|^{-1} \qquad (66) \\
&\geq 1 - r^{nd} |\mathcal{T}_k^d|^{-1}, \qquad (67)
\end{aligned}$$

where in (65), we used the fact that $\mathbb{P}(\widehat{T} \neq T_P | T_P = t) = 1$ if $t \in \mathcal{W}^c$, in (66), the fact that $\mathbb{P}(T_P = t) = 1/|\mathcal{T}_k^d|$. In (67), we used the observation $|\mathcal{W}| \leq (|\mathcal{X}^d|)^n = r^{nd}$ since the function $\widehat{T}_{\mathrm{MAP}} : (\mathcal{X}^d)^n \to \mathcal{W}$ is surjective. Now, the number of labeled forests with k edges and d nodes is (Aigner and Ziegler, 2009, pp. 204) $|\mathcal{T}_k^d| \geq (d-k)d^{k-1} \geq d^{k-1}$. Applying this lower bound to (67), we obtain

$$\mathbb{P}(\widehat{T} \neq T_P) \geq 1 - \exp\left(nd \log r - (k-1)\log d\right) > 1 - \exp\left((\rho - 1)(k-1)\log d\right), \qquad (68)$$

where the second inequality follows by choice of n in (17). The estimate in (68) converges to 1 as $(k,d) \to \infty$ since $\rho < 1$. The same reasoning applies to part (b) but we instead use the following estimates of the cardinality of the set of forests (Aigner and Ziegler, 2009, Ch. 30):

$$(d-2)\log d \leq \log |\mathcal{F}^d| \leq (d-1)\log(d+1). \qquad (69)$$

Note that we have lower bounded $|\mathcal{F}^d|$ by the number trees with d nodes which is d^{d-2} by Cayley's formula (Aigner and Ziegler, 2009, Ch. 30). The upper bound[17] follows by a simple combinatorial argument which is omitted. Using the lower bound in (69), we have

$$\mathbb{P}(\widehat{T} \neq T_P) \geq 1 - \exp(nd \log r) \exp(-(d-2)\log d) > 1 - d^2 \exp((\rho - 1) d \log d), \qquad (70)$$

with the choice of n in (18). The estimate in (70) converges to 1, completing the proof. ■

Appendix G. Proof of Theorem 8

Proof We assume that P is Markov on a forest since the extension to non-forest-structured P is a straightforward generalization. We start with some useful definitions. Recall from Appendix B that $\mathcal{B}_n := \{\widehat{E}_k \neq E_P\}$ is the event that the top k edges (in terms of mutual information) in the edge set $\widehat{E}_{d-1}$ are not equal to the edges in E_P. Also define $\widetilde{\mathcal{C}}_{n,\delta} := \{D(P^* || P) > \delta d\}$ to be the event that the divergence between the learned model and the true (forest) one is greater than δd. We will

16. In fact, this proof works for *any* decoding rule, and not just the MAP rule. We focus on the MAP rule for concreteness.
17. The purpose of the upper bound is to show that our estimates of $|\mathcal{F}^d|$ in (69) are reasonably tight.

Figure 10: In $\widehat{E}_{\widehat{k}_n}$ (left), nodes 1 and 5 are the roots. The parents are defined as $\pi(i;\widehat{E}_{\widehat{k}_n}) = i-1$ for $i = 2,3,4,6$ and $\pi(i;\widehat{E}_{\widehat{k}_n}) = \emptyset$ for $i = 1,5$. In E_P (right), the parents are defined as $\pi(i;E_P) = i-1$ for $i = 2,3,4$ but $\pi(i;E_P) = \emptyset$ for $i = 1,5,6$ since $(5,6),(\emptyset,1),(\emptyset,5) \notin E_P$.

see that $\widetilde{\mathcal{C}}_{n,\delta}$ is closely related to the event of interest $\mathcal{C}_{n,\delta}$ defined in (20). Let $\mathcal{U}_n := \{\widehat{k}_n < k\}$ be the underestimation event. Our proof relies on the following result, which is similar to Lemma 11, hence its proof is omitted.

Lemma 13 *For every $\eta > 0$, there exists a $N \in \mathbb{N}$ such that for all $n > N$, the following bounds on $P^n(\widetilde{\mathcal{C}}_{n,\delta})$ hold:*

$$(1-\eta)P^n(\widetilde{\mathcal{C}}_{n,\delta}|\mathcal{B}_n^c,\mathcal{U}_n^c) \leq P^n(\widetilde{\mathcal{C}}_{n,\delta}) \tag{71}$$

$$\leq P^n(\widetilde{\mathcal{C}}_{n,\delta}|\mathcal{B}_n^c,\mathcal{U}_n^c) + \exp(-n(\min\{K_P,L_P\}-\eta)). \tag{72}$$

Note that the exponential term in (72) comes from an application of the union bound and the "largest-exponent-wins" principle in large-deviations theory (Den Hollander, 2000). From (71) and (72) we see that it is possible to bound the probability of $\widetilde{\mathcal{C}}_{n,\delta}$ by providing upper and lower bounds for $P^n(\widetilde{\mathcal{C}}_{n,\delta}|\mathcal{B}_n^c,\mathcal{U}_n^c)$. In particular, we show that the upper bound equals $\exp(-n\delta)$ to first order in the exponent. This will lead directly to (21). To proceed, we rely on the following lemma, which is a generalization of a well-known result (Cover and Thomas, 2006, Ch. 11). We defer the proof to the end of the section.

Lemma 14 (Empirical Divergence Bounds) *Let X,Y be two random variables whose joint distribution is $P_{X,Y} \in \mathcal{P}(\mathcal{X}^2)$ and $|\mathcal{X}| = r$. Let $(x^n,y^n) = \{(x_1,y_1),\ldots,(x_n,y_n)\}$ be n independent and identically distributed observations drawn from $P_{X,Y}$. Then, for every n,*

$$P_{X,Y}^n(D(\widehat{P}_{X|Y}\,||\,P_{X|Y}) > \delta) \leq (n+1)^{r^2}\exp(-n\delta), \tag{73}$$

where $\widehat{P}_{X|Y} = \widehat{P}_{X,Y}/\widehat{P}_Y$ is the conditional type of (x^n,y^n). Furthermore,

$$\liminf_{n\to\infty} \frac{1}{n}\log P_{X,Y}^n(D(\widehat{P}_{X|Y}\,||\,P_{X|Y}) > \delta) \geq -\delta. \tag{74}$$

It is worth noting that the bounds in (73) and (74) are independent of the distribution $P_{X,Y}$ (cf. discussion after Theorem 8). We now proceed with the proof of Theorem 8. To do so, we consider the directed representation of a tree distribution Q (Lauritzen, 1996):

$$Q(\mathbf{x}) = \prod_{i\in V} Q_{i|\pi(i)}(x_i|x_{\pi(i)}), \tag{75}$$

where $\pi(i)$ is the parent of i in the edge set of Q (assuming a fixed root). Using (75) and conditioned on the fact that the top k edges of the graph of P^* are the same as those in E_P (event $\mathcal{B}_n^c$) and

underestimation does not occur (event $\mathcal{U}_n^c$), the KL-divergence between P^* (which is a function of the samples $\mathbf{x}^n$ and hence of n) and P can be expressed as a sum over d terms:

$$D(P^*||P) = \sum_{i\in V} D(\widehat{P}_{i|\pi(i;\widehat{E}_{\widehat{k}_n})}||P_{i|\pi(i;E_P)}), \tag{76}$$

where the parent of node i in $\widehat{E}_{\widehat{k}_n}$, denoted $\pi(i;\widehat{E}_{\widehat{k}_n})$, is defined by arbitrarily choosing a root in each component tree of the forest $\widehat{T}_{\widehat{k}_n} = (V,\widehat{E}_{\widehat{k}_n})$. The parents of the chosen roots are empty sets. The parent of node i in E_P are "matched" to those in $\widehat{E}_{\widehat{k}_n}$, that is, defined as $\pi(i;E_P) := \pi(i;\widehat{E}_{\widehat{k}_n})$ if $(i,\pi(i;\widehat{E}_{\widehat{k}_n})) \in E_P$ and $\pi(i;E_P) := \emptyset$ otherwise. See Figure 10 for an example. Note that this can be done because $\widehat{E}_{\widehat{k}_n} \supseteq E_P$ by conditioning on the events $\mathcal{B}_n^c$ and $\mathcal{U}_n^c = \{\widehat{k}_n \geq k\}$. Then, the error probability $P^n(\widetilde{\mathcal{C}}_{n,\delta}|\mathcal{B}_n^c,\mathcal{U}_n^c)$ in (72) can be upper bounded as

$$\begin{aligned}
P^n(\widetilde{\mathcal{C}}_{n,\delta}|\mathcal{B}_n^c,\mathcal{U}_n^c) &= P^n\left(\sum_{i\in V} D(\widehat{P}_{i|\pi(i;\widehat{E}_{\widehat{k}_n})}||P_{i|\pi(i;E_P)}) > \delta d \Big| \mathcal{B}_n^c,\mathcal{U}_n^c\right) && (77)\\
&= P^n\left(\frac{1}{d}\sum_{i\in V} D(\widehat{P}_{i|\pi(i;\widehat{E}_{\widehat{k}_n})}||P_{i|\pi(i;E_P)}) > \delta \Big| \mathcal{B}_n^c,\mathcal{U}_n^c\right)\\
&\leq P^n\left(\max_{i\in V}\left\{D(\widehat{P}_{i|\pi(i;\widehat{E}_{\widehat{k}_n})}||P_{i|\pi(i;E_P)})\right\} > \delta \Big| \mathcal{B}_n^c,\mathcal{U}_n^c\right) && (78)\\
&\leq \sum_{i\in V} P^n\left(D(\widehat{P}_{i|\pi(i;\widehat{E}_{\widehat{k}_n})}||P_{i|\pi(i;E_P)}) > \delta \Big| \mathcal{B}_n^c,\mathcal{U}_n^c\right) && (79)\\
&\leq \sum_{i\in V} (n+1)^{r^2}\exp(-n\delta) = d(n+1)^{r^2}\exp(-n\delta), && (80)
\end{aligned}$$

where Equation (77) follows from the decomposition in (76). Equation (78) follows from the fact that if the arithmetic mean of d positive numbers exceeds δ, then the maximum exceeds δ. Equation (79) follows from the union bound. Equation (80), which holds for all $n \in \mathbb{N}$, follows from the upper bound in (73). Combining (72) and (80) shows that if $\delta < \min\{K_P, L_P\}$,

$$\limsup_{n\to\infty} \frac{1}{n}\log P^n(\widetilde{\mathcal{C}}_{n,\delta}) \leq -\delta.$$

Now recall that $\widetilde{\mathcal{C}}_{n,\delta} = \{D(P^*||P) > \delta d\}$. In order to complete the proof of (21), we need to swap the arguments in the KL-divergence to bound the probability of the event $\mathcal{C}_{n,\delta} = \{D(P||P^*) > \delta d\}$. To this end, note that for $\varepsilon > 0$ and n sufficiently large, $|D(P^*||P) - D(P||P^*)| < \varepsilon$ with high probability since the two KL-divergences become close ($P^* \approx P$ w.h.p. as $n \to \infty$). More precisely, the probability of $\{|D(P^*||P) - D(P||P^*)| \geq \varepsilon\} = \{o(\|P - P^*\|_\infty^2) \geq \varepsilon\}$ decays exponentially with some rate $M_P > 0$. Hence,

$$\limsup_{n\to\infty} \frac{1}{n}\log P^n(D(P||P^*) > \delta d) \leq -\delta, \tag{81}$$

if $\delta < \min\{K_P, L_P, M_P\}$. If P is not Markov on a forest, (81) holds with the forest projection $\widetilde{P}$ in place of P, that is,

$$\limsup_{n\to\infty} \frac{1}{n}\log P^n(D(\widetilde{P}||P^*) > \delta d) \leq -\delta. \tag{82}$$

The Pythagorean relationship (Simon, 1973; Bach and Jordan, 2003) states that

$$D(P||P^*) = D(P||\widetilde{P}) + D(\widetilde{P}||P^*) \tag{83}$$

which means that the risk is $\mathcal{R}_n(P^*) = D(\widetilde{P}||P^*)$. Combining this fact with (82) implies the assertion of (21) by choosing $\delta_0 := \min\{K_P, L_P, M_P\}$.

Now we exploit the lower bound in Lemma 14 to prove the lower bound in Theorem 8. The error probability $P^n(\widetilde{\mathcal{C}}_{n,\delta}|\mathcal{B}_n^c, \mathcal{U}_n^c)$ in (72) can now be lower bounded by

$$P^n(\widetilde{\mathcal{C}}_{n,\delta}|\mathcal{B}_n^c, \mathcal{U}_n^c) \geq \max_{i\in V} P^n\left(D(\widehat{P}_{i|\pi(i;\widehat{E}_{\widehat{k}_n})}||P_{i|\pi(i;E_P)}) > \delta d \,\middle|\, \mathcal{B}_n^c, \mathcal{U}_n^c\right) \tag{84}$$

$$\geq \exp(-n(\delta d + \eta)), \tag{85}$$

where (84) follows from the decomposition in (77) and (85) holds for every η for sufficiently large n by (74). Using the same argument that allows us to swap the arguments of the KL-divergence as in the proof of the upper bound completes the proof of (22). ■

Proof (*of Lemma 14*) Define the δ-*conditional-typical set with respect to* $P_{X,Y} \in \mathcal{P}(\mathcal{X}^2)$ as

$$\mathcal{S}^{\delta}_{P_{X,Y}} := \{(x^n, y^n) \in (\mathcal{X}^2)^n : D(\widehat{P}_{X|Y}||P_{X|Y}) \leq \delta\},$$

where $\widehat{P}_{X|Y}$ is the conditional type of (x^n, y^n). We now estimate the $P^n_{X,Y}$-probability of the δ-conditional-atypical set, that is, $P^n_{X,Y}((\mathcal{S}^{\delta}_{P_{X,Y}})^c)$

$$P^n_{X,Y}((\mathcal{S}^{\delta}_{P_{X,Y}})^c) = \sum_{(x^n,y^n)\in\mathcal{X}^2 : D(\hat{P}_{X|Y}||P_{X|Y})>\delta} P^n_{X,Y}((x^n, y^n)) \tag{86}$$

$$= \sum_{Q_{X,Y}\in\mathcal{P}_n(\mathcal{X}^2) : D(Q_{X|Y}||P_{X|Y})>\delta} P^n_{X,Y}(\mathsf{T}_n(Q_{X,Y})) \tag{87}$$

$$\leq \sum_{Q_{X,Y}\in\mathcal{P}_n(\mathcal{X}^2) : D(Q_{X|Y}||P_{X|Y})>\delta} \exp(-nD(Q_{X,Y}||P_{X,Y})) \tag{88}$$

$$\leq \sum_{Q_{X,Y}\in\mathcal{P}_n(\mathcal{X}^2) : D(Q_{X|Y}||P_{X|Y})>\delta} \exp(-nD(Q_{X|Y}||P_{X|Y})) \tag{89}$$

$$\leq \sum_{Q_{X,Y}\in\mathcal{P}_n(\mathcal{X}^2) : D(Q_{X|Y}||P_{X|Y})>\delta} \exp(-n\delta) \tag{90}$$

$$\leq (n+1)^{r^2} \exp(-n\delta), \tag{91}$$

where (86) and (87) are the same because summing over sequences is equivalent to summing over the corresponding type classes since every sequence in each type class has the same probability (Cover and Thomas, 2006, Ch. 11). Equation (88) follows from the method of types result in Lemma 12. Equation (89) follows from the KL-divergence version of the chain rule, namely,

$$D(Q_{X,Y}||P_{X,Y}) = D(Q_{X|Y}||P_{X|Y}) + D(Q_Y||P_Y)$$

and non-negativity of the KL-divergence $D(Q_Y||P_Y)$. Equation (90) follows from the fact that $D(Q_{X|Y}||P_{X|Y}) > \delta$ for $Q_{X,Y} \in (\mathcal{S}^{\delta}_{P_{X,Y}})^c$. Finally, (91) follows the fact that the number of types with denominator n and alphabet $\mathcal{X}^2$ is upper bounded by $(n+1)^{r^2}$. This concludes the proof of (73).

We now prove the lower bound in (74). To this end, construct a sequence of distributions $\{Q^{(n)}_{X,Y} \in \mathcal{P}_n(\mathcal{X}^2)\}_{n\in\mathbb{N}}$ such that $Q^{(n)}_Y = P_Y$ and $D(Q^{(n)}_{X|Y}||P_{X|Y}) \to \delta$. Such a sequence exists by the denseness of types in the probability simplex (Dembo and Zeitouni, 1998, Lemma 2.1.2(b)). Now we lower bound (87):

$$P^n_{X,Y}((\mathcal{S}^{\delta}_{P_{X,Y}})^c) \geq P^n_{X,Y}(\mathsf{T}_n(Q^{(n)}_{X,Y})) \geq (n+1)^{-r^2}\exp(-nD(Q^{(n)}_{X,Y}\,||\,P_{X,Y})). \tag{92}$$

Taking the normalized logarithm and liminf in n on both sides of (92) yields

$$\liminf_{n\to\infty}\frac{1}{n}\log P^n_{X,Y}((\mathcal{S}^{\delta}_{P_{X,Y}})^c) \geq \liminf_{n\to\infty}\left\{-D(Q^{(n)}_{X|Y}||P_{X|Y}) - D(Q^{(n)}_Y||P_Y)\right\} = -\delta.$$

This concludes the proof of Lemma 14. ■

Appendix H. Proof of Corollary 9

Proof If the dimension $d = o(\exp(n\delta))$, then the upper bound in (80) is asymptotically majorized by $\mathrm{poly}(n)o(\exp(na))\exp(-n\delta) = o(\exp(n\delta))\exp(-n\delta)$, which can be made arbitrarily small for n sufficiently large. Thus the probability tends to zero as $n\to\infty$. ■

Appendix I. Proof of Theorem 10

Proof In this proof, we drop the superscript (d) for all distributions P for notational simplicity but note that $d = d_n$. We first claim that $D(P^*\,||\,\widetilde{P}) = O_p(d\log d/n^{1-\gamma})$. Note from (72) and (80) that by taking $\delta = (\tau\log d)/n^{1-\gamma}$ (for any $\tau > 0$),

$$P^n\left(\frac{n^{1-\gamma}}{d\log d}D(P^*\,||\,\widetilde{P}) > \tau\right) \leq d(n+1)^{r^2}\exp(-\tau n^{\gamma}\log d) + \exp(-\Theta(n)) = o_n(1). \tag{93}$$

Therefore, the scaled sequence of random variables $\frac{n^{1-\gamma}}{d\log d}D(P^*\,||\,\widetilde{P})$ is stochastically bounded (Serfling, 1980) which proves the claim.[18]

Now, we claim that $D(\widetilde{P}\,||\,P^*) = O_p(d\log d/n^{1-\gamma})$. A simple calculation using Pinsker's Inequality and Lemma 6.3 in Csiszár and Talata (2006) yields

$$D(\widehat{P}_{X,Y}\,||\,P_{X,Y}) \leq \frac{c}{\kappa}D(P_{X,Y}\,||\,\widehat{P}_{X,Y}),$$

where $\kappa := \min_{x,y}P_{X,Y}(x,y)$ and $c = 2\log 2$. Using this fact, we can use (73) to show that for all n sufficiently large,

$$P^n_{X,Y}(D(P_{X|Y}\,||\,\widehat{P}_{X|Y}) > \delta) \leq (n+1)^{r^2}\exp(-n\delta\kappa/c),$$

that is, if the arguments in the KL-divergence in (73) are swapped, then the exponent is reduced by a factor proportional to κ. Using this fact and the assumption in (14) (uniformity of the minimum

18. In fact, we have in fact proven the stronger assertion that $D(P^*\,||\,\widetilde{P}) = o_p(d\log d/n^{1-\gamma})$ since the right-hand-side of (93) converges to zero.

entry in the pairwise joint $\kappa > 0$), we can replicate the proof of the result in (80) with $\delta\kappa/c$ in place of δ giving

$$P^n(D(P||P^*) > \delta) \leq d(n+1)^{r^2} \exp(-n\delta\kappa/c).$$

We then arrive at a similar result to (93) by taking $\delta = (\tau \log d)/n^{1-\gamma}$. We conclude that $D(\widetilde{P}||P^*) = O_p(d \log d/n^{1-\gamma})$. This completes the proof of the claim.

Equation (23) then follows from the definition of the risk in (19) and from the Pythagorean theorem in (83). This implies the assertion of Theorem 10. ■

References

P. Abbeel, D. Koller, and A. Y. Ng. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, Dec 2006.

M. Aigner and G. M. Ziegler. *Proofs From THE BOOK*. Springer, 2009.

F. Bach and M. I. Jordan. Beyond independent components: trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233, 2003.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2008.

S. Borade and L. Zheng. Euclidean information theory. In *IEEE International Zurich Seminar on Communications*, pages 14–17, 2008.

G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: Some observations and algorithms. In *11th International workshop APPROX 2008 and 12th International workshop RANDOM*, pages 343–356., 2008.

A. Chechetka and C. Guestrin. Efficient principled learning of thin junction trees. In *Advances of Neural Information Processing Systems (NIPS)*, 2007.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Infomation Theory*, 14(3):462–467, May 1968.

C. K. Chow and T. Wagner. Consistency of an estimate of tree-dependent probability distributions . *IEEE Transactions in Information Theory*, 19(3):369 – 371, May 1973.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

I. Csiszár and F. Matúš. Information projections revisited. *IEEE Transactions on Infomation Theory*, 49(6):1474–1490, 2003.

I. Csiszár and P. Shields. The consistency of the BIC Markov order estimator. *Ann. Statist.*, 28(6): 1601–1619, 2000.

I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information Theory*, 52(3):1007–16, 2006.

A. Custovic, B. M. Simpson, C. S. Murray, L. Lowe, and A. Woodcock. The national asthma campaign Manchester asthma and allergy study. *Pediatr Allergy Immunol*, 13:32–37, 2002.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2nd edition, 1998.

F. Den Hollander. *Large Deviations (Fields Institute Monographs, 14)*. American Mathematical Society, Feb 2000.

M. Dudik, S. J. Phillips, and R. E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Conference on Learning Theory (COLT)*, 2004.

R. M. Fano. *Transmission of Information*. New York: Wiley, 1961.

L. Finesso, C. C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Trans. on Info Th.*, 42(5):1488–1497, 1996.

R. G. Gallager. Claude E. Shannon: A retrospective on his life, work and impact. *IEEE Trans. on Info. Th.*, 47:2687–95, Nov 2001.

E. Gassiat and S. Boucheron. Optimal error exponents in hidden Markov models order estimation. *IEEE Transactions on Infomation Theory*, 49(4):964–980, Apr 2003.

J. Johnson, V. Chandrasekaran, and A. S. Willsky. Learning Markov structure by maximum entropy relaxation. In *Artificial Intelligence and Statistics (AISTATS)*, 2007.

J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1), Feb 1956.

S. Lauritzen. *Graphical Models*. Oxford University Press, USA, 1996.

S. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using L1-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman. Forest density estimation. *Journal of Machine Learning Research*, 12:907–951, March 2011.

M. Meilă and M. I. Jordan. Learning with mixtures of trees. *Journal of Machine Learning Research*, 1:1–48, Oct 2000.

N. Meinshausen and P. Buehlmann. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

N. Merhav. The estimation of the model order in exponential families. *IEEE Transactions on Infomation Theory*, 35(5):1109–1115, 1989.

N. Merhav, M. Gutman., and J. Ziv. On the estimation of the order of a Markov chain and universal data compression. *IEEE Transactions on Infomation Theory*, 35:1014–1019, 1989.

D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI Repository of Machine Learning Databases, University of California, Irvine, 1998.

R. C. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36, 1957.

N. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. In *Proc. of IEEE Intl. Symp. on Info. Theory*, Toronto, Canada, July 2008.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley-Interscience, Nov 1980.

G. Simon. Additivity of information in exponential family probability laws. *Amer. Statist. Assoc.*, 68(478–482), 1973.

A. Simpson, V. Y. F. Tan, J. M. Winn, M. Svensén, C. M. Bishop, D. E. Heckerman, I. Buchan, and A. Custovic. Beyond atopy: Multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med*, 2010.

V. Y. F. Tan, A. Anandkumar, and A. S. Willsky. Learning Gaussian tree models: Analysis of error exponents and extremal structures. *IEEE Transactions on Signal Processing*, 58(5):2701 – 2714, May 2010a.

V. Y. F. Tan, A. Anandkumar, and A. S. Willsky. Error exponents for composite hypothesis testing of Markov forest distributions. In *Proc. of Intl. Symp. on Info. Th.*, June 2010b.

V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky. A large-deviation analysis for the maximum-likelihood learning of Markov tree structures. *IEEE Transactions on Infomation Theory*, Mar 2011.

L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, Mar 1996.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, University of California, Berkeley, 2003.

M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances of Neural Information Processing Systems (NIPS)*, pages 1465–1472, 2006.

O. Zuk, S. Margel, and E. Domany. On the number of samples needed to learn the correct structure of a Bayesian network. In *Proc of Uncertainty in Artificial Intelligence (UAI)*, 2006.