# Learning Tree Structures from Noisy Data

**Konstantinos E. Nikolakakis**
Rutgers University

**Dionysios S. Kalogerias**
Princeton University

**Anand D. Sarwate**
Rutgers University

## Abstract

We provide high-probability sample complexity guarantees for exact structure recovery of tree-structured graphical models, when only noisy observations of the respective vertex emissions are available. We assume that the hidden variables follow either an Ising model or a Gaussian graphical model, and the observables are noise-corrupted versions of the hidden variables: We consider multiplicative $\pm 1$ binary noise for Ising models, and additive Gaussian noise for Gaussian models. Such hidden models arise naturally in a variety of applications such as physics, biology, computer science, and finance. We study the impact of measurement noise on the task of learning the underlying tree structure via the well-known *Chow-Liu algorithm*, and provide formal sample complexity guarantees for exact recovery. In particular, for a tree with $p$ vertices and probability of failure $\delta > 0$, we show that the number of necessary samples for exact structure recovery is of the order of $\mathcal{O}(\log(p/\delta))$ for Ising models (which remains the *same as in the noiseless case*), and $\mathcal{O}(\text{polylog}(p/\delta))$ for Gaussian models.

## 1 Introduction

Graphical models are a useful tool for modeling high-dimensional structured data. In particular, Markov random fields (MRFs) are undirected graphical models in which variables, represented by nodes in a graph, satisfy conditional independence properties, the so-called Markov properties. The graph models the dependence between variables: the set of the edges corresponds to (often physical) interactions between variables. There is a long and deep literature on graphical models (see [1] for a comprehensive introduction), which have also found wide applications in areas such as image processing and vision [2, 3, 4, 5, 6, 7], artificial intelligence more broadly [8, 9], signal processing [10, 11], and gene regulatory networks [12, 13].

An undirected graphical model, or Markov Random Field, is defined in terms of a hyper-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which models a joint distribution on variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ where $p = |\mathcal{V}|$. A tree-structured graphical model is one in which $\mathcal{G}$ is a tree. We denote the tree-structured model as $\mathrm{T} = (\mathcal{V}, \mathcal{E})$.

This paper studies concurrently two distinct problem settings. First, we consider *binary models* on $2p$ variables $(\mathbf{X}, \mathbf{Y})$, where the joint distribution $\mathrm{p}(\cdot)$ of $\mathbf{X}$ is a tree structured model distribution and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$ constitutes a noisy version of $\mathbf{X}$. Specifically, we assume that $\mathbf{X}$ follows a tree structured Ising model, whereas $\mathbf{Y}$ is the output of a binary symmetric channel with crossover probability $q$, and input $\mathbf{X}$. Under this setting, our objective is to exactly recover the underlying tree structure of the *hidden layer* $\mathbf{X}$ by only using noisy observables $\mathbf{Y}$. This is non-trivial, as $\mathbf{Y}$ does not itself follow any tree structure.

We also consider the case where $\mathbf{X}$ follows a *Gaussian tree-structured distribution*, and $\mathbf{Y}$ is the output $\mathbf{X}$ measured through an Additive White Gaussian Noise (AWGN) channel. This is similar to more traditional nonlinear filtering, where a Markov process of known distribution (and thus, of known structure) is observed through noisy measurements [14, 15, 16, 17, 18].

Under both settings, we use the *Chow-Liu algorithm* [19] to reconstruct the tree. The main contribution of this paper is to characterize the effect of observation noise on *hidden tree-structure estimation*.

**Notation.** Boldface indicates a vector or tuple. Calligraphic face is used for sets and trees. For an integer $n$, let $[n] \triangleq \{1, 2, \ldots n\}$. The cardinality of the set of nodes $\mathcal{V}$ is assumed to be equal to $p$, $|\mathcal{V}| = [p]$. The indicator function of a set $A$ is denoted as $\mathbf{1}_A$. For a pair $i, j \in [p]$, the correlation of two *hidden* random variables $X_i, X_j$ is denoted $\mu_{ij} \stackrel{\triangle}{=} \mathbb{E}[X_i X_j]$ and the corre-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

lation coefficient as $\rho_{i,j} \triangleq \mathbb{E}[X_i X_j]/\sqrt{\mathbb{E}[X_i^2]\mathbb{E}[X_j^2]}$. If $(i,j) \in \mathcal{E}$, then we write $\mu_e \triangleq \mathbb{E}[X_i X_j]$ and $e \equiv (i,j)$. For two nodes $w, \tilde{w}$ of a tree, the path$(w, \tilde{w})$ denotes the set of edges in the unique path with endpoints $w$ and $\tilde{w}$. The probability mass function of $\mathbf{X}$ is denoted as $\mathrm{p}(\cdot)$. We use the symbol † to indicate the corresponding quantity for the observable (noisy) layer. For instance, $\mathrm{p}_\dagger(\cdot)$ is the probability mass function of $\mathbf{Y}$ and $\mu_{i,j}^\dagger \triangleq \mathbb{E}[Y_i Y_j]$ corresponds to the correlation of variables $Y_i, Y_j$, where $Y_i$ generates noisy observations of $X_i$, for any $i \in \mathcal{V}$. Also, $\mathrm{BSC}(q)^p$ denotes a binary symmetric channel with crossover $q$ and block-length $p$.

### 1.1 Motivating Examples and Applications

Models for joint distributions characterized by pairwise variable interactions have found many applications, with the Ising model being a popular model for binary variables. Our work is primarily motivated by examples of Ising models *corrupted by noise*. In many cases, the underlying graph-structured process cannot be observed directly; instead, only a noisy version of the process is available; examples abound in physics, computer science, biology, medicine, psychology, social sciences, and finance. Some applications motivating this work include the following:

**1)** *Statistical mechanics of population, social and pedestrian dynamics* [20, 21]: The Ising model can be used to represent the statistical properties of the spreading of a feeling, behavior or the change of an emotional state among individuals in a crowd, where each individual interacts with his neighbors.

**2)** *Epidemic dynamics and epidemiological models* [22, 23]: Disease spread can be modeled through the Ising model, where each individual is susceptible (spin down) or ineffective (spin up).

**3)** *Neoplastic transitions* and related applications in biology [24]: Each cell interacts with neighboring cells. Different cases are studied in the literature, for instance, healthy versus cancerous cells, malignant versus benign cells, where both can be modeled as spin up and spin down observations. The probability of diagnostic error is not zero which gives rise to the hidden model that we consider.

**4)** *Differential Privacy* [25]: In computer science, differential privacy is used to guarantee privacy for individuals. A hidden model describes data gathered using a privacy-preserving mechanism such as randomized response (Section C, Appendix).

**5)** *Trading* and related applications in economics [26, 27]: The Ising model has been considered in the literature to model increasing (spin up) or decreasing (spin down) price trends in a market. In Section 5, we consider the closing prices of ten equities to demonstrate the performance of Chow-Liu algorithm.

### 1.2 Contribution

The main question asked by this paper is as follows: *What is the impact of observation noise on the sample complexity of learning a tree structured graphical model?* For the binary case, we sample variables $\mathbf{Y}$ generated by $\mathbf{X}$, which follows a tree-structured Ising model distribution and randomly flipping each sign independently with probability $q$. A typical example is classification, where a subset of the data might be misclassified. Then, corrupted data are observed, however, we are still able to retrieve the underlying structure by considering the appropriate number of samples. Under the Gaussian model assumption, we sample $\mathbf{Y}$, the output of an AWGN channel with Gaussian input $\mathbf{X}$.

This paper makes the following contributions:

- For the binary case, we provide a lower bound (Theorem 1) on the number of samples that are sufficient for the Chow-Liu algorithm to recover the tree structure of $\mathrm{p}(\cdot)$, with probability at least $1-\delta$. This generalizes the noiseless case, for which the Chow-Liu algorithm applied to simple pairwise correlation estimates is order-optimal [28], and has interesting implications for the applications mentioned earlier. Here, we explicitly show that the same algorithm, with appropriate minor modifications, *achieves nearly the same rate*, as the order of necessary number of samples remains the same as in the noiseless case, i.e., $\mathcal{O}(\log(p/\delta))$.
- We prove an upper bound (Theorem 2) on the number of binary samples necessary for any algorithm to recover a tree structure. The proof shows an intriguing connection between the hidden model and recent studies on strong data processing inequalities [29]. Future work could strengthen this upper bound but may require additional assumptions.
- For the Gaussian case, we provide a lower bound (Theorem 3) on the number of samples that are sufficient for the Chow-Liu algorithm to recover the underlying structure. This is a general result which reduces to the noiseless case as the noise variance goes to zero. In particular, we show that the order of the necessary number of samples is *polylogarithmic in* $p/\delta$, i.e., $\mathcal{O}(\log^4(p/\delta))$.

Our results strictly generalize the noiseless case [28,

Theorem 3.1, Theorem 3.2] for that of a hidden model. Our proof strategy is similar, but the hidden model presents additional complexity, which presents extended technical challenges, requiring the development of new arguments. In particular, the corresponding graph of the observable layer, $\mathbf{Y}$, is *no longer a tree*, so the Markov property does not hold for the observable nodes. In addition, closed-form bounds of the KL divergence can not be computed for the set of output distributions $\mathrm{p}_\dagger(\cdot)$. To overcome this, we combine Bresler's and Karzand's method [28] and a strong data processing inequality by Polyanskiy and Wu [29], to derive an upper bound on the sample complexity.

### 1.3 Related Work

We refer the reader to the textbook by Koller and Friedman [1] for background material and a recent review by Drton et al. [30]. In general, learning the graph structure of a graphical model from samples can be intractable, which has been shown by [31, 32]. For general graphs, neighborhood selection methods [33, 34, 35] estimate the conditional distribution for each vertex in order to learn the neighborhood of each node and therefore the full structure. These approaches may use greedy search or $\ell_1$ regularization. For Gaussian or Ising models, works have proposed $\ell_1$-regularization by [36], the GLasso [37, 38] or coordinate descent approaches by [39] to learn the structure by estimating the non-zero entries of the precision (or interaction) matrix. Model selection can also be done using score matching methods [40, 41, 42, 43] or Bayesian information criterion methods by [44, 45, 46]. Other works address non-Gaussian models such as elliptical distributions, t-distribution models, latent Gaussian data or even mixed data [47, 48, 49, 50, 51]. Latent variable models are considered by [52, 53, 54, 55, 56], when some variables of the graph are deterministically unobserved.

For tree- or forest-structured models the problem is significantly simpler: the Chow-Liu algorithm [19] provides an estimate of the tree or forest structure of the underlying graph. This has resulted in a number of sample complexity results for for tree-structure learning [28, 57, 58, 59, 60, 61]).

We differ from these models because we consider the case of *noisy observable data* (or a hidden model), in which the underlying structure does not appear in the marginal distribution of the observed variables. We generalize results known for the noiseless case [28] to this new setting. In the special case of a linear graph, our class of models also includes hidden Markov models (HMMs): one application of our results could be to testing if data follow an HMM.

Our model is similar to previous works considering hidden models with discrete exponential distribution and Gaussian noise [62]. They solve the parameter estimation problem by using moment matching and pseudolikelihood methods; the structure can be recovered indirectly using the estimated parameters. In contrast, we analyze the performance of Chow-Liu algorithm, because we can estimate the underlying structure with low sample complexity and low computational cost.

The performance of Chow-Liu algorithm for Gaussian tree structured models has also been studied in prior work by Tan et.al. [63], which shows which tree types are hardest to be estimated by using the Chow-Liu algorithm. More specifically, they show that, in some cases, "star"-shaped trees can be harder to be estimated than "chain"-shaped trees. In the worst case scenario, where the correlations can be small or large (outside of a critical region), the sample complexity is not affected by the shape of tree. Our results provide the exact bound for sufficient number of samples, as a function of the minimum and maximum correlation, and the power of noise, in the worst case scenario, thus the bound does not depend on the type of structure.

Alternative methods and algorithms for sparse structure estimation are based on Graphical LASSO [39], and variations of it [64, 65, 66, 67, 68, 69]. These methods have a competitive sample complexity, but their algorithms are computationally more expensive than the Chow-Liu algorithm. An interesting approach *for the noiseless case*, which uses the Chow-Liu algorithm, was recently proposed by Tavassolipour et.al. [70]. By observing only the sign of the Gaussian observations, they show that structure learning is feasible, while the sample complexity is $\mathcal{O}(\log(p/\delta))$. This result can lead to further improvements of our results for the Gaussian hidden model, and constitutes a subject currently under investigation.

## 2 Models and Problem Statement

We start by presenting the models under consideration and their properties.

### 2.1 Tree-structured Ising models

According to the Ising model, the hypergraph $\mathcal{G}$ is a simple undirected graph, indicating that the associated node variables have only pairwise and unitary interactions, and each variable takes values in $\{-1, +1\}$. The joint distribution for the *Ising model with zero external field* is given by

$$\mathrm{p}(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{(s,t)\in\mathcal{E}} \theta_{st} x_s x_t \right\}, \ \mathbf{x} \in \{-1, 1\}^p, \quad (1)$$

where $\{\theta_{st} : (s,t) \in \mathcal{E}\}$ are parameters of the model representing the interaction strength of the variables, and $Z(\cdot) \in (0,\infty)$ is the *partition function*. In this paper, the considered model allows only pairwise interactions between the nodes. These interactions are expressed through potential functions $\exp(\theta_{st} x_s x_t)$, which ensure that the Markov property holds with respect to the graph $G = (\mathcal{V}, \mathcal{E})$. From Lauritzen's prior work [71], we know that any distribution $\mathrm{p}(\cdot)$ which is Markov with respect to a tree $\mathrm{T} = (\mathcal{V}, \mathcal{E})$ factorizes as

$$\mathrm{p}(\mathbf{x}) = \prod_{i \in V} p\left(x_i\right) \prod_{(i,j)\in\mathcal{E}} \frac{p(x_i, x_j)}{p(x_i)p(x_j)}, \ \mathbf{x} \in \{-1,1\}^p. \quad (2)$$

For our analysis, as in the noiseless setting [28], we require the parameters $\theta_{st}$ to be bounded for edges $(s,t) \in \mathcal{E}$ in (1), as follows.

**Assumption 1.** *There exist $\alpha$ and $\beta$, such that $0 < \alpha \leq |\theta_{st}| \leq \beta < \infty$ for all $(s,t) \in \mathcal{E}$.*

We also use the notation $\mathcal{P}_{\mathrm{T}}(\alpha, \beta)$ to denote the *set of tree structured Ising models satisfying* Assumption 1.

### 2.2 Hidden Ising Model

For the binary case, we consider a hidden Markov random field with hidden layer $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(\alpha, \beta)$, under Assumption 1. The observed variables $\mathbf{Y}$ are formed by setting $Y_r = N_r X_r$, for all $r \in \mathcal{V}$, where $\{N_r\}$ are i.i.d. $\mathsf{Rademacher}(q)$ random variables. Under this model, the observables $\mathbf{Y} \sim \mathrm{p}_\dagger(\cdot)$ are the outputs of $\mathbf{X}$ when passed through a memoryless binary symmetric channel (BSC) with crossover probability $q$, or $\mathrm{BSC}(q)^p$.

### 2.3 Tree structured Gaussian models

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ be a Gaussian random vector with distribution $\mathcal{N}(0, \Sigma)$. The nonzero entries of the precision matrix $\Sigma^{-1}$ indicate the existence of the corresponding edge in the underlying graph. The following assumption holds on the Gaussian data.

**Assumption 2.** *The variances of variables $X_i$ are equal to 1, for all $i \in \mathcal{V}$. Furthermore, there exist numbers $\rho_m, \rho_M$, such that*

$$0 < \rho_m \leq |\rho_{i,j}| \leq \rho_M < 1, \quad \forall (i,j) \in \mathcal{E}. \quad (3)$$

Hereafter, we use the notation $\boldsymbol{\mathcal{N}}_{\mathrm{T}}^{m,M}$ to denote the *set of tree structured Gaussian distributions satisfying* Assumption 2. We also use the notation $\mu_{i,j}$ for the correlation coefficient, since $\rho_{i,j} = \mathbb{E}[X_i X_j]$ under Assumption 2.

### 2.4 Hidden Gaussian Model

For the Gaussian setting, and for $\mathbf{N} \sim \mathcal{N}(0, \sigma^2 \mathcal{I})$, the noisy output variables of the hidden model are taken as $\mathbf{X} + \mathbf{N} = \tilde{\mathbf{Y}} \sim \mathcal{N}(0, \Sigma + \sigma^2 \mathcal{I})$. Then the correlation coefficient of the observable data is

$$\rho_{i,j}^\dagger = \mathbb{E}\left[\frac{\tilde{Y}_i}{\sqrt{1+\sigma^2}} \frac{\tilde{Y}_j}{\sqrt{1+\sigma^2}}\right], \quad \forall i,j \in \mathcal{V}. \quad (4)$$

The random variables $Y_i \triangleq \tilde{Y}_i / \sqrt{1+\sigma^2}$ are normalized Gaussian with variance equal to 1. To simplify the analysis, we use normalized samples, instead of samples directly from $\tilde{\mathbf{Y}}$, which correspond to the variable $\mathbf{Y}$ with distribution

$$\mathbf{Y} \sim \mathcal{N}\left(0, \frac{\Sigma + \sigma^2 \mathcal{I}}{1+\sigma^2}\right). \quad (5)$$

Thus, $\mathbb{E}[Y_i Y_j] = \rho_{i,j}^\dagger$. For the rest of the paper we use the notation $\mu_{i,j}^\dagger$, where $\mu_{i,j}^\dagger \equiv \mathbb{E}[Y_i Y_j]$ and its corresponding estimated value is denoted as $\hat{\mu}_{i,j}^\dagger$.

### 2.5 Chow-Liu Algorithm

The algorithm we study throughout this paper is the classical *Chow-Liu algorithm* (Algorithm 1), which requires as input the set of *noisy observations* $\{\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(n_\dagger)\}$, computes the estimates $\hat{\mu}_{i,j}^\dagger$, and returns a tree structure $\mathrm{T}_\dagger^{\mathrm{CL}}$, that is an estimate of T. The estimates $\hat{\mu}_{i,j}^\dagger$ are consistent with the Gaussian model as well because of (5). Following the noiseless case [28][Lemma A.2],

$$\mathrm{T}_\dagger^{\mathrm{CL}} = \mathrm{argmax}_{\mathrm{T}\in\mathcal{T}} \sum_{(i,j)\in\in\mathcal{E}_{\mathrm{T}}} \left|\hat{\mu}_{i,j}^\dagger\right|. \quad (6)$$

The difference between and Algorithm 1 in our setting and the noiseless version [28] is that we use the observed noisy variables $\mathbf{Y}$ rather than $\mathbf{X}$. This idea is immediately theoretically justified; the tree structure estimate $\mathrm{T}_\dagger^{\mathrm{CL}}$ converges to T when $n_\dagger \to \infty$ since

$$\lim_{n\to\infty} \frac{\hat{\mu}_{i,j}^\dagger}{(1-2q)^2} \overset{\text{a.s.}}{=} \mu_{i,j}, \quad (7)$$

for the hidden Ising model, whereas

$$\lim_{n\to\infty} (1+\sigma^2)\hat{\mu}_{i,j}^\dagger \overset{\text{a.s.}}{=} \mu_{i,j}, \quad (8)$$

for the case of a hidden Gaussian model. From (6), (7) and (8), it is true that, for both cases,

$$\lim_{n\to\infty} \mathrm{T}_\dagger^{\mathrm{CL}} \overset{\text{a.s.}}{=} \mathrm{T}. \quad (9)$$

**Algorithm 1** Chow – Liu

**Require:** Data set $\mathcal{D} = \{\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(n_\dagger)\}$

1: $\hat{\mu}^\dagger_{i,j} \leftarrow \frac{1}{n_\dagger} \sum_k y_i(k) y_j(k)$, for all $i, j \in \mathcal{V}$

2: $\mathrm{T}^{\mathrm{CL}}_\dagger \leftarrow \mathrm{MaximumSpanningTree}\Big(\cup_{i \neq j} \{\hat{\mu}^\dagger_{i,j}\}\Big)$

3: **return** $\mathrm{T}^{\mathrm{CL}}_\dagger$

### 2.6 Problem Statement

Under the models outlined above, we are interested in exact recovery the underlying tree-structures from noisy observations, as well as explicitly comparing the sample complexity of the hidden model $n_\dagger$ with $n$, denoting the number of samples required for exact structure learning when observations directly from the hidden variable $\mathbf{X}$ are available [28]. Likewise, we call $\mathrm{T}^{\mathrm{CL}}$ (compare with $\mathrm{T}^{\mathrm{CL}}_\dagger$) the structure that is learned by using purely noiseless observations of $\mathbf{X}$. From our discussion above, it readily follows that both $\mathrm{T}^{\mathrm{CL}}$ and $\mathrm{T}^{\mathrm{CL}}_\dagger$ converge to T. However, we expect that more samples should be required for a hidden model, compared to the noiseless case. The goal of this paper is to *quantitatively characterize this gap.*

## 3 Exact Recovery of Hidden Structures

For our structure learning task, we consider the *zero-one loss error measure*, which has been used to denote quantify structure recovery by [28, 59, 60, 72, 73] and it is defined as

$$\mathcal{L}^{0-1}\left(\mathrm{T}, \mathrm{T}^{\mathrm{CL}}_\dagger\right) \triangleq \mathbf{1}_{\mathrm{T} \neq \mathrm{T}^{\mathrm{CL}}_\dagger}. \tag{10}$$

Thus, the probability of incorrect reconstruction may be expressed as

$$\mathbb{P}\left(\mathrm{T}^{\mathrm{CL}}_\dagger \neq \mathrm{T}\right) = \mathbb{E}\left[\mathcal{L}^{0-1}\left(\mathrm{T}, \mathrm{T}^{\mathrm{CL}}_\dagger\right)\right]. \tag{11}$$

Further, in our analysis, the notation $\boldsymbol{S}_{\mathrm{KL}}(P||Q)$ is used for the *symmetric Kullback-Leibler (KL) divergence* between probability measures $P, Q$, defined as

$$\boldsymbol{S}_{\mathrm{KL}}(P||Q) \triangleq \boldsymbol{D}_{\mathrm{KL}}(P||Q) + \boldsymbol{D}_{\mathrm{KL}}(Q||P). \tag{12}$$

Specifically, for Ising model distributions $P, Q$ as in (1) with corresponding parameters $\boldsymbol{\theta}, \boldsymbol{\theta}'$, it is true that

$$\begin{aligned} \boldsymbol{S}_{\mathrm{KL}}\left(\boldsymbol{\theta}||\boldsymbol{\theta}'\right) &\triangleq \boldsymbol{S}_{\mathrm{KL}}(P||Q) \\ &= \sum_{s,t \in \mathcal{E}} \left(\theta_{st} - \theta'_{st}\right)\left(\mu_{st} - \mu'_{st}\right). \end{aligned} \tag{13}$$

### 3.1 Main Results

The first main result of the paper is presented below, providing a computable bound on the sufficient number of samples guaranteeing exact structure recovery for the hidden Ising model under consideration.

**Theorem 1** (Sufficient number of samples)**.** *Let* $\mathbf{Y}$ *be the output of a* $\mathrm{BSC}(q)^p$*, with input variable* $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(\alpha, \beta)$*. Fix a number* $\delta \in (0, 1)$*. If the number of samples of* $\mathbf{Y}$ *satisfies the inequality*

$$n_\dagger \geq \frac{32\left[1 - (1-2q)^4 \tanh\beta\right]}{(1-2q)^4 \left(1 - \tanh\beta\right)^2 \tanh^2\alpha} \log\frac{2p^2}{\delta}, \tag{14}$$

*then Algorithm 1 returns* $\mathrm{T}^{CL}_\dagger = \mathrm{T}$ *with probability at least* $1 - \delta$*.*

Complementary to Theorem 1, our next result characterizes the necessary number samples required for exact structure recovery whatsoever.

**Theorem 2** (Necessary number of samples)**.** *Let* $\mathbf{Y}$ *be the output of a* $\mathrm{BSC}(q)^p$*, with input variable* $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(\alpha, \beta)$*. If the given number of samples satisfies*

$$n_\dagger < \frac{\left[1 - (4q(1-q))^p\right]^{-1}}{16\alpha \tanh(\alpha)} e^{2\beta} \log(p), \tag{15}$$

*then for any (measurable) estimator* $\psi$*, it is true that*

$$\inf_{\psi} \sup_{\substack{\mathrm{T} \in \mathcal{T} \\ P \in \mathcal{P}_{\mathrm{T}}(\alpha,\beta)}} \mathbb{P}\left(\psi\left(\mathbf{Y}_{1:n_\dagger}\right) \neq \mathrm{T}\right) > \frac{1}{2}. \tag{16}$$

Theorems 1 and 2 provide sufficient and necessary conditions for exact structure recovery, when only noisy observations are available. It can be shown that the right hand-side of (14) is greater than the right-hand side of (15) for any $q$ in $[0, 1/2)$ (and for all possible values of $p, \beta, \alpha$), by comparing the two terms. For $q = 0$ the bounds reduce to the noiseless setting, the sufficient number of samples is an increasing function of $q$ and structure learning is always feasible as long as $q \neq 1/2$.

To better understand the effect of the noise, note that

$$\frac{\left[1 - (1-2q)^4 \tanh(\beta)\right]}{\left[(1-2q)^4 (1 - \tanh(\beta))\right]} \geq 1, \quad \forall q \in \left[0, \frac{1}{2}\right), \tag{17}$$

with $\beta \in \mathbb{R}$, and

$$\frac{1}{1 - (4q(1-q))^p} \geq 1, \quad \forall q \in \left[0, \frac{1}{2}\right), \, p \in \mathbb{N}, \tag{18}$$

which shows that the sample complexity in a hidden model is greater than the noiseless case ($q = 0$), for any (measurable) estimator (Theorem 2). When $q$ approaches $1/2$, the sample complexity goes to infinity, $n_\dagger \to \infty$, which makes structure learning impossible.

Theorem 2 is an extension of Theorem 3.1 by Bresler and Karzand [28] to our hidden model. Our results

combines Bresler and Karzand's method and a strong data processing inequality by Polyanskiy and Wu [29]. Upper bounds on the symmetric KL divergence for the output distribution $\mathrm{p}_\dagger(\cdot)$ can not be found in a closed form. However, by using the SDPI, we manage to capture the dependence of the bound on the parameters $\alpha, \beta, q$ and derive a non-trivial result. When $p$ goes to $\infty$, the bound of Theorem 2 becomes trivial and loose; $\lim_{p\to\infty} 1/\left[1-(4q(1-q))^p\right] \to 1$, which gives the classical data processing inequality (contraction of KL divergence for finite alphabets, [29, 74]). To recover a tighter bound for Theorem 2, one would have to derive a strong data processing inequality problem under Assumption 1. This is technically challenging problem, and constitutes a subject for future research.

**Remark.** *Bresler and Karzand [28] prove additional sample complexity results on upper-bounding the small-set total variation (ssTV) metric by a positive number $\eta$. A key property of tree-structured Ising models is independence of the product variables $X_iX_j$, for any $(i,j) \in \mathcal{E}$, (see [28, Lemma 8.6]). This simplifies the analysis in the noiseless case significantly [28, Lemma E.1, page 32]. Analyzing our hidden model is strictly more complicated, since the underlying graph of the observable layer is a complete graph, and the variables $Y_iY_j$ are not jointly independent for any $(i,j) \in \mathcal{E}$. We will examine possible solutions to this problem in future work to characterize how bounded correlations affect cascades in hidden models.*

Our third and final result characterizes the sample complexity for achieving exact structure recovery, when only noisy observations from a Gaussian hidden model (as defined above) are available.

**Theorem 3.** *Let $\mathbf{Y}$ be the output of a Gaussian channel, $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{X} \sim \mathrm{p}(\cdot) \in \boldsymbol{\mathcal{M}}_{\mathrm{T}}^{m,M}$ and $\mathbf{N} \sim \mathcal{N}(0,\sigma^2)$. Fix a number $\delta \in (0,1)$. The Chow-Liu algorithm recovers the structure, $\mathrm{T} = \mathrm{T}_\dagger^{CL}$ with probability at least $1-\delta$, if the number of samples satisfies the inequality*

$$n \geq \frac{R^2\left[7(1+\sigma^2)^2+\rho_M\right]\log^4\left(\frac{e^2p^3}{\delta}\right)}{\rho_m^2\left(1-\rho_M\right)^2}, \tag{19}$$

*and $R$ is a positive constant.*

Theorem 3 gives a lower bound on the sufficient number of samples needed for exact structure recovery, for the case of the Gaussian model. The required amount of observations increases as the power of noise increases. For $\sigma = 0$, the bound provides the sample complexity of the corresponding noiseless setting, while for $\sigma \to \infty$ the structure learning task becomes impossible.

### 3.2 Comparison with the noiseless setting

Theorems 1 and 2 *strictly generalize* noiseless tree-structure recovery [28, Theorem 3.1, Theorem 3.2] for our hidden model; the noiseless results correspond to $q = 0$. In particular, it very interesting to observe that, in the presence of noise, the dependence of our complexity bounds on $p$ is *still logarithmic*, that is, of the order of $\mathcal{O}(\log(p/\delta))$. To make the connection between sufficient conditions more explicit, it is true that, in the noiseless case, if the weakest edge satisfies the inequality

$$\tanh\alpha \geq \frac{4\epsilon}{\sqrt{1-\tanh\beta}}, \tag{20}$$

and $\epsilon$ is defined as $\epsilon \triangleq \sqrt{2\log\left(2p^2/\delta\right)/n}$, yielding

$$n \geq \frac{32}{\tanh^2\alpha\left(1-\tanh\beta\right)}\log\left(\frac{2p^2}{\delta}\right), \tag{21}$$

then the structure is recovered exactly with probability $1-\delta$ by the Chow-Liu algorithm. For our hidden model, the respective condition for the weakest edge is

$$\tanh\alpha \geq \frac{4\epsilon_\dagger\sqrt{1-(1-2q)^4\tanh\beta}}{(1-2q)^2\left(1-\tanh\beta\right)}, \tag{22}$$

and $\epsilon_\dagger$ is similarly defined as $\epsilon_\dagger \triangleq \sqrt{2\log\left(2p^2/\delta\right)/n_\dagger}$.

Note that, for $q = 1/2$, the mutual information of the hidden and observable variables is zero, that is, $\mathbf{X}$ and $\mathbf{Y}$ are independent, so structure recovery is impossible.

To make the relevant connection between necessary conditions, it holds that, in the noiseless case, if the number of samples satisfies

$$n < \frac{1}{16}e^{2\beta}\alpha^{-2}\log\left(p\right), \tag{23}$$

then for any (measurable) algorithmic mapping $\psi$, it is true that

$$\inf_{\psi}\sup_{\substack{\mathrm{T}\in\mathcal{T}\\ P\in\mathcal{P}_{\mathrm{T}}(\alpha,\beta)}}\mathbb{P}\left(\psi\left(X_{1:n}\right)\neq\mathrm{T}\right) > \frac{1}{2}. \tag{24}$$

When $q = 0$, we retrieve the noiseless result, while for any $q \in (0, \frac{1}{2})$ the sample complexity increases since $[1-(4q(1-q))^p]^{-1} > 1$ in (15) and for $q \to 1/2$ the required number of samples $n_\dagger \to \infty$ which makes structure learning impossible. The ratio between the noiseless and noisy necessary conditions indicates the gap between the hidden model and the noiseless one

$$\frac{n_\dagger}{n} \leq \left[1-(4q(1-q))^p\right]^{-1} \leq \frac{1}{\eta_{\mathrm{KL}}}, \tag{25}$$

where the right hand-side inequality is the strong data processing inequality for the binary symmetric channel by Polyanskiy and Wu [29, Equation (39)].

As far as Theorem 3 is concerned, this reduces to the noiseless setting for $\sigma = 0$. Recently, the performance of the Chow-Liu algorithm for the *noiseless* Gaussian case was studied by Tavassolipour [70]. In this work, a lower complexity bound is derived, closely resembling the noiseless Ising model. The approach of [70] might potentially drive further improvement of our hidden Gaussian model bound (Theorem 3), and is the subject of our future work.

## 4 Analysis: Proof Sketches

Due to space limitation a sketch of the proof for each theorem is given. The definition of necessary events and the complete proofs can be found in Section A (Appendix).

**Theorem 1.** To analyze the Chow-Liu algorithm, we consider the error event [28], at least one edge to be missed; if an edge $f = (w, \bar{w}) \in \mathrm{T}$ and $f \notin \mathrm{T}_{\dagger}^{\mathrm{CL}}$ (i.e. the edge is incorrectly not inferred), then there exists an edge $g \in \mathrm{T}_{\dagger}^{\mathrm{CL}}$ and $g \notin \mathrm{T}$ such that $f \in \mathrm{path}_{\mathrm{T}}(u, \bar{u})$, $g \in \mathrm{path}_{\mathrm{T}_{\dagger}^{\mathrm{CL}}}(w, \bar{w})$, and

$$\left(\sum_{i=1}^{n_\dagger} Z_{f,u,\bar{u}}^{(i)}\right)\left(\sum_{i=1}^{n_\dagger} M_{f,u,\bar{u}}^{(i)}\right) < 0, \tag{26}$$

where $Z_{f,u,\bar{u}} \triangleq Y_w Y_{\bar{w}} - Y_u Y_{\bar{u}}$ and $M_{f,u,\bar{u}} \triangleq Y_w Y_{\bar{w}} + Y_u Y_{\bar{u}}$. Thus, to show that the reconstruction is successful, we need to show that this event does *not* happen, with high probability.

By using Bresler and Karzand's method [75, Lemmas 9.6, 9.7] under the error event (at least one incorrect edge in the estimated tree structure $\mathrm{T}_{\dagger}^{\mathrm{CL}}$) we have $\left|\hat{\mu}_f^\dagger\right| \leq \left|\hat{\mu}_g^\dagger\right|$, which gives

$$\begin{aligned}
0 &\geq \left|\hat{\mu}_f^\dagger\right|^2 - \left|\hat{\mu}_g^\dagger\right|^2 \\
&= \left(\hat{\mu}_f^\dagger - \hat{\mu}_g^\dagger\right)\left(\hat{\mu}_f^\dagger + \hat{\mu}_g^\dagger\right) \\
&= \frac{1}{n_\dagger^2}\left(\sum_{i=1}^{n_\dagger} N_w^{(i)} X_w^{(i)} N_{\bar{w}}^{(i)} X_{\bar{w}}^{(i)} - N_u^{(i)} X_u^{(i)} N_{\bar{u}}^{(i)} X_{\bar{u}}^{(i)}\right) \\
&\qquad \times \left(\sum_{i=1}^{n_\dagger} N_w^{(i)} X_w^{(i)} N_{\bar{w}}^{(i)} X_{\bar{w}}^{(i)} + N_w^{(i)} X_u^{(i)} N_{\bar{u}}^{(i)} X_{\bar{u}}^{(i)}\right) \\
&= \frac{1}{n_\dagger^2}\left(\sum_{i=1}^{n_\dagger} Z_{f,u,\bar{u}}^{(i)}\right)\left(\sum_{i=1}^{n_\dagger} M_{f,u,\bar{u}}^{(i)}\right).
\end{aligned} \tag{27}$$

Notice that the random variables $Z_{f,u,\bar{u}}^{(i)}$, $M_{f,u,\bar{u}}^{(i)}$ are functions of noisy observables. To understand how these quantities behave we use Bernstein's inequality, which produces a factor of $(1-2q)^2$ to account for the variance in the noisy samples. The concentration of measure results we need are for $Z_{f,u,\bar{u}}^{(i)}$, $M_{f,u,\bar{u}}^{(i)}$. Defining events $E_Z$ and $E_M$ as

$$E_Z \triangleq \bigcap_{(w,\bar{w})\in\mathcal{E}, u,\bar{u}\in\mathcal{V}} E_Z^{(w,\bar{w}),u,\bar{u}}, \tag{28}$$

$$E_M \triangleq \bigcap_{(w,\bar{w})\in\mathcal{E}, u,\bar{u}\in\mathcal{V}} E_M^{(w,\bar{w}),u,\bar{u}}, \tag{29}$$

and

$$\begin{aligned}
E_Z^{(w,\bar{w}),u,\bar{u}} \triangleq \Bigg\{ &\left|\frac{1}{n_\dagger}\sum_{i=1}^{n_\dagger} Z_{e,u,\bar{u}}^{(i)} - \mathbb{E}\left[Z_{e,u,\bar{u}}\right]\right| \\
&\leq \max\left\{8\epsilon_\dagger^2, 4\epsilon_\dagger\sqrt{1-\mu_A^\dagger}\right\}\Bigg\},
\end{aligned} \tag{30}$$

$$\begin{aligned}
E_M^{(w,\bar{w}),u,\bar{u}} \triangleq \Bigg\{ &\left|\frac{1}{n_\dagger}\sum_{i=1}^{n_\dagger} M_{e,u,\bar{u}}^{(i)} - \mathbb{E}\left[M_{e,u,\bar{u}}\right]\right| \\
&\leq \max\left\{8\epsilon_\dagger^2, 4\epsilon_\dagger\sqrt{1+\mu_A^\dagger}\right\}\Bigg\},
\end{aligned} \tag{31}$$

it is possible to show that each occurs with probability at least $1-\delta'/2$ and $1-\delta''/2$ respectively, where $\epsilon_\dagger = \sqrt{2/n_\dagger \log(2p^2/\delta)}$ and $A = \mathrm{path}_{\mathrm{T}}(u,\bar{u}) \setminus \{e\}$. A union bound over all pairs $w, \bar{w}, u, \bar{u}$ and finally for the events $E_Z, E_M$ shows that the event $E_Z \cup E_M$ happens with probability at least $1-\delta$, where $\delta'/2 + \delta''/2 \leq 2\max\{\delta'/2, \delta''/2\} \triangleq \delta$.

**Theorem 2.** To show a (minimax) lower bound on tree structure estimation we follow the standard information-theoretic recipe using Fano's inequality [76, Corollary 2.6]. As for noiseless models [75, Section 8.1], we consider difficult instances of the problem correspond to graphs which are nearly chains, see also Section A.2 (Appendix). First, we define $P_{\theta^0}$ to be an Ising model distribution with underlying structure a chain with $p$ nodes and parameters $\theta_{j,j+1}^0 = \alpha$ when $j$ is odd and $\theta_{j,j+1}^0 = \beta$, when $j$ is even. The rest of family is constructed as follows: the elements of each $\theta^i$ are equal to the elements of $\theta^0$ apart from two elements $\theta_{i,i+1}^i = 0$ and $\theta_{i,i+2}^i = \alpha$ for each odd value of $i$. There are $(p+1)/2$ distributions in the constructed family. To find a non-trivial upper bound for the quantity $\boldsymbol{S}_{\mathrm{KL}}(P_{\theta^0}^\dagger || P_{\theta^i}^\dagger)$, (where $P_{\theta^i}^\dagger$ is the distribution of the observable variables of the $i^{\mathrm{th}}$ model) we require new techniques, namely the strong data processing inequality (SDPI) for the binary symmetric channel [29].

Consider our hidden model in which $\mathbf{X}$ is drawn from $\mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(\alpha,\beta)$, corrupted by multiplicative $\mathsf{Rademacher}(q)$ noise, $N_i$. We have [29]:

$$\begin{aligned}
\eta_{\mathrm{KL}} &\triangleq \sup_{Q}\sup_{P:0<\boldsymbol{D}_{\mathrm{KL}}(P||Q)<\infty} \frac{\boldsymbol{D}_{\mathrm{KL}}\left(P_{\mathbf{Y}|\mathbf{X}}\circ P || P_{\mathbf{Y}|\mathbf{X}}\circ Q\right)}{\boldsymbol{D}_{\mathrm{KL}}\left(P||Q\right)} \\
&\leq 1 - (4q(1-q))^p.
\end{aligned} \tag{32}$$

We combine this with an upper bound on the symmetric KL divergence between any pair of noiseless models in a specially constructed set of $M+1$ trees:

$$\boldsymbol{S}_{\mathrm{KL}}(P_{\theta^0}||P_{\theta^i}) \leq 4\alpha^2 e^{-2\beta}, \quad \forall i \in [M]. \tag{33}$$

This in turn yields the factor $1-(4q(1-q))^p$ in the final bound.

**Theorem 3.** The proof of Theorem 3 differs from that of Theorem 1 at two points. First, the correlation decay property holds as a correlation coefficient decay property, which makes the normalization of $\tilde{Y}$ an essential step for the analysis (see (5)). The correlation coefficient decay property [63] can be stated as follows. If $X \sim \mathrm{p}(\cdot) \in \boldsymbol{\mathcal{N}}_{\mathrm{T}}^{m,M}$, then

$$\rho_{i,j} = \prod_{e \in \mathrm{path}_{\mathrm{T}}(i,j)} \rho_e, \quad \forall (i,j) \in \mathcal{V}, \tag{34}$$

where $\rho_{i,j} = \mathbb{E}[X_i X_j]/\sqrt{\mathbb{E}[X_i^2]\mathbb{E}[X_j^2]}$. Furthermore, to bound the probability of the events analogous to (28) and (29) in the Gaussian case from above, we need concentration of measure inequalities for polynomials of dependent continuous random variables. For that purpose, we use a recent concentration result by Shudi and Sviridenko [77, Theorem 1.10]. This results in the polylogarithmic sample complexity in the Gaussian Case.

## 5 Experiments

For the experimental part we consider synthetic data. To demonstrate the performance of the Chow-Liu algorithm experimentally, we present the decay of the probability of incorrect recovery (Fig. 1), while the number of samples increases, for fixed values of the parameters $\alpha, \beta, p, \rho_M, \rho_m$. These results illustrate how noisy observations can degrade performance, unless we increase the sample size. Based on the experiments, more observations are required for the Gaussian model than the Ising model to provide an estimate $\mathrm{T}_{\dagger}^{\mathrm{CL}}$ with small $\mathbb{P}\left(\mathrm{T}_{\dagger}^{\mathrm{CL}} \neq \mathrm{T}\right)$ (compared under equivalent noise levels). The probability (Fig. 1) approaches zero with less observations in the case of the Ising model than the Gaussian model, for instance compare the lines for $q=0$, which corresponds to $\mathrm{SNR}=\infty$. Further experiments are provided in Section B (Appendix).

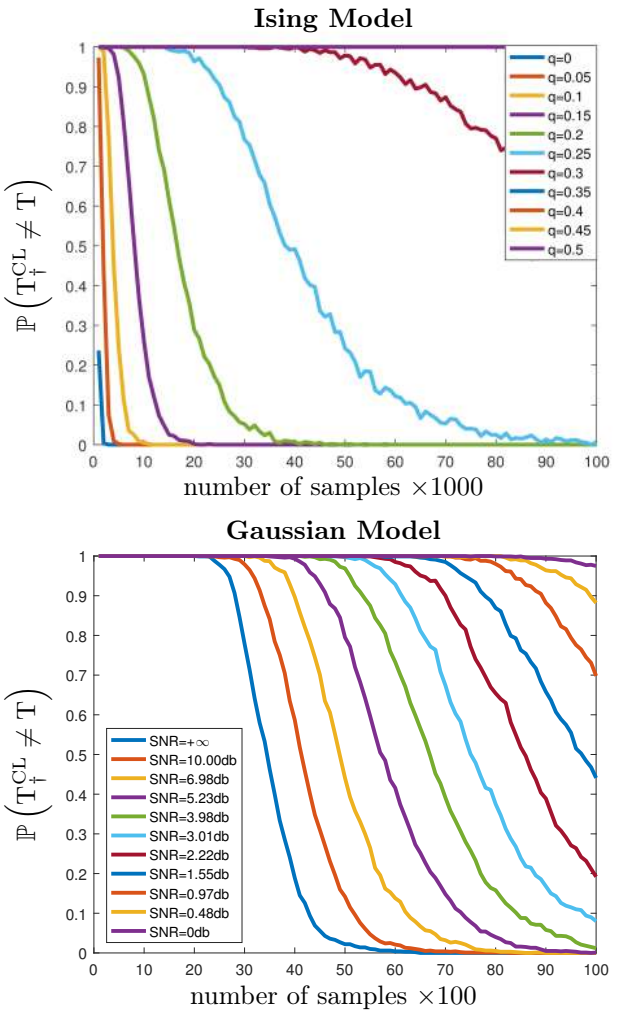


Figure 1: Estimating the probability of the error event through 1000 iterations.

## 6 Conclusion

We have analyzed the problem of perfect reconstruction of hidden tree-structures from noisy observations, using the well-known Chow-Liu algorithm. In particular, we have focused on two distinct cases, namely, hidden Ising models observed in multiplicative $\pm 1$ binary noise, and hidden Gaussian graphical models observed in additive Gaussian noise. For the case of a hidden Ising model, our results (Theorem 1 and Theorem 2) give lower and upper bounds on the sample complexity of accurately inferring the latent tree structure, strictly generalizing the previously-studied noiseless case. In particular, the lower bound shows that the number of samples needed to estimate the tree grows only as $\mathcal{O}(\log(p/\delta))$, where $\delta > 0$ is the probability of incorrect recovery. For hidden Gaussian models, we provide an extension of our method, and derive a lower bound on the number of sufficient samples for exact structure recovery, which is polylogarithmic in $p/\delta$ (Theorem 3). Experiments illustrate the impact of the noise and how properly accounting for noisy samples can lead to more accurate structure inference.

## References


[1] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[2] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," *arXiv preprint arXiv:1503.02351*, 2015.

[3] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3194–3203.

[4] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] A. Morningstar and R. G. Melko, "Deep learning the ising model near criticality," *arXiv preprint arXiv:1708.04622*, 2017.

[6] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Deep learning markov random field for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[7] B. Wu, B.-G. Hu, and Q. Ji, "A coupled hidden markov random field model for simultaneous face clustering and tracking in videos," *Pattern Recognition*, vol. 64, pp. 361–373, 2017.

[8] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching." in *AISTATS*, 2003.

[9] B. Wang, Z. Ou, and Z. Tan, "Learning trans-dimensional random fields with applications to language modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[10] S. Wisdom, J. Hershey, J. Le Roux, and S. Watanabe, "Deep unfolding for multichannel source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 121–125.

[11] M. Kim and P. Smaragdis, "Single channel source separation using smooth nonnegative matrix factorization with markov random fields," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*. IEEE, 2013, pp. 1–6.

[12] Y. Zuo, Y. Cui, G. Yu, R. Li, and H. W. Ressom, "Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical lasso," *BMC bioinformatics*, vol. 18, no. 1, p. 99, 2017.

[13] M. Banf and S. Y. Rhee, "Enhancing gene regulatory network inference through data integration with markov random fields," *Scientific Reports*, vol. 7, 2017.

[14] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 174–188, 2002.

[15] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[16] R. Van Handel, "Observability and nonlinear filtering," *Probability theory and related fields*, vol. 145, no. 1-2, pp. 35–74, 2009.

[17] R. Douc, E. Moulines, J. Olsson, R. Van Handel *et al.*, "Consistency of the maximum likelihood estimator for general hidden markov models," *the Annals of Statistics*, vol. 39, no. 1, pp. 474–513, 2011.

[18] D. S. Kalogerias and A. P. Petropulu, "Grid based nonlinear filtering revisited: recursive estimation & asymptotic optimality," *IEEE Transactions on Signal Processing*, vol. 64, no. 16, pp. 4244–4259, 2016.

[19] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[20] H. Matsuda, N. Ogita, A. Sasaki, and K. Satō, "Statistical mechanics of population: the lattice lotka-volterra model," *Progress of theoretical Physics*, vol. 88, no. 6, pp. 1035–1049, 1992.

[21] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of modern physics*, vol. 81, no. 2, p. 591, 2009.

[22] E. Y. Erten, J. T. Lizier, M. Piraveenan, and M. Prokopenko, "Criticality and information dynamics in epidemiological models," *Entropy*, vol. 19, no. 5, p. 194, 2017.

[23] L. Barnett, J. T. Lizier, M. Harré, A. K. Seth, and T. Bossomaier, "Information flow in a kinetic ising model peaks in the disordered phase," *Physical review letters*, vol. 111, no. 17, p. 177203, 2013.

[24] S. Torquato, "Toward an ising model of cancer and beyond," *Physical biology*, vol. 8, no. 1, p. 015017, 2011.

[25] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[26] T. Takaishi, "Multiple time series ising model for financial market simulations," in *Journal of Physics: Conference Series*, vol. 574, no. 1. IOP Publishing, 2015, p. 012149.

[27] W.-X. Zhou and D. Sornette, "Self-organizing ising model of financial markets," *The European Physical Journal B*, vol. 55, no. 2, pp. 175–181, 2007.

[28] G. Bresler and M. Karzand, "Learning a tree-structured ising model in order to make predictions," *arXiv preprint arXiv:1604.06749*, 2018.

[29] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*. Springer, 2017, pp. 211–249.

[30] M. Drton and M. H. Maathuis, "Structure learning in graphical modeling," *Annual Review of Statistics and Its Application*, vol. 4, pp. 365–393, 2017.

[31] D. Karger and N. Srebro, "Learning markov networks: Maximum bounded tree-width graphs," in *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2001, pp. 392–401.

[32] S. Højsgaard, D. Edwards, and S. Lauritzen, *Graphical models with R*. Springer Science & Business Media, 2012.

[33] G. Bresler, "Efficiently learning ising models on arbitrary graphs," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015, pp. 771–782.

[34] A. Ray, S. Sanghavi, and S. Shakkottai, "Improved greedy algorithms for learning graphical models," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3457–3468, 2015.

[35] A. Jalali, C. C. Johnson, and P. K. Ravikumar, "On learning discrete graphical models using greedy methods," in *Advances in Neural Information Processing Systems*, 2011, pp. 1935–1943.

[36] P. Ravikumar, M. J. Wainwright, J. D. Lafferty *et al.*, "High-dimensional ising model selection using l1-regularized logistic regression," *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

[37] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[38] O. Banerjee, L. E. Ghaoui, and A. d Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *Journal of Machine learning research*, vol. 9, no. Mar, pp. 485–516, 2008.

[39] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[40] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 695–709, 2005.

[41] ——, "Some extensions of score matching," *Computational statistics & data analysis*, vol. 51, no. 5, pp. 2499–2512, 2007.

[42] P. Nandy, A. Hauser, and M. H. Maathuis, "High-dimensional consistency in score-based and hybrid structure learning," *arXiv preprint arXiv:1507.02608*, 2015.

[43] L. Lin, M. Drton, A. Shojaie *et al.*, "Estimation of high-dimensional graphical models using regularized score matching," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 806–854, 2016.

[44] R. Foygel and M. Drton, "Extended bayesian information criteria for gaussian graphical models," in *Advances in neural information processing systems*, 2010, pp. 604–612.

[45] X. Gao, D. Q. Pu, Y. Wu, and H. Xu, "Tuning parameter selection for penalized likelihood estimation of gaussian graphical model," *Statistica Sinica*, pp. 1123–1146, 2012.

[46] R. F. Barber, M. Drton *et al.*, "High-dimensional ising model selection with bayesian information criteria," *Electronic Journal of Statistics*, vol. 9, no. 1, pp. 567–607, 2015.

[47] D. Vogel and R. Fried, "Elliptical graphical modelling," *Biometrika*, vol. 98, no. 4, pp. 935–951, 2011.

[48] D. Vogel and D. E. Tyler, "Robust estimators for nondecomposable elliptical graphical models," *Biometrika*, vol. 101, no. 4, pp. 865–882, 2014.

[49] M. Bilodeau, "Graphical lassos for meta-elliptical distributions," *Canadian Journal of Statistics*, vol. 42, no. 2, pp. 185–203, 2014.

[50] M. Finegold and M. Drton, "Robust graphical modeling of gene networks using classical and alternative t-distributions," *The Annals of Applied Statistics*, pp. 1057–1080, 2011.

[51] J. Fan, H. Liu, Y. Ning, and H. Zou, "High dimensional semiparametric latent graphical model for mixed data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 2, pp. 405–421, 2017.

[52] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 1610–1613.

[53] M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning latent tree graphical models," *Journal of Machine Learning Research*, vol. 12, no. May, pp. 1771–1812, 2011.

[54] A. Anandkumar and R. Valluvan, "Learning loopy graphical models with latent variables: Efficient methods and guarantees," *The Annals of Statistics*, pp. 401–435, 2013.

[55] S. Ma, L. Xue, and H. Zou, "Alternating direction methods for latent variable gaussian graphical model selection," *Neural computation*, vol. 25, no. 8, pp. 2172–2198, 2013.

[56] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014.

[57] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[58] D. Edwards, G. C. De Abreu, and R. Labouriau, "Selecting high-dimensional mixed graphical models using minimal aic or bic forests," *BMC bioinformatics*, vol. 11, no. 1, p. 18, 2010.

[59] V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning high-dimensional markov forest distributions: Analysis of error rates," *Journal of Machine Learning Research*, vol. 12, no. May, pp. 1617–1653, 2011.

[60] H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman, "Forest density estimation," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 907–951, 2011.

[61] C. Daskalakis, N. Dikkala, and G. Kamath, "Testing ising models," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2018, pp. 1989–2007.

[62] A. T. Chaganty and P. Liang, "Estimating latent-variable graphical models using moments and likelihoods," in *International Conference on Machine Learning*, 2014, pp. 1872–1880.

[63] V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning gaussian tree models: Analysis of error exponents and extremal structures," *arXiv preprint arXiv:0909.5216*, 2009.

[64] J. Friedman, T. Hastie, and R. Tibshirani, "Applications of the lasso and grouped lasso to the estimation of sparse graphical models," Technical report, Stanford University, Tech. Rep., 2010.

[65] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electronic journal of statistics*, vol. 6, p. 2125, 2012.

[66] H. Wang *et al.*, "Bayesian graphical lasso models and efficient posterior computation," *Bayesian Analysis*, vol. 7, no. 4, pp. 867–886, 2012.

[67] R. Mazumder and T. Hastie, "Exact covariance thresholding into connected components for large-scale graphical lasso," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 781–794, 2012.

[68] T. Sun and C.-H. Zhang, "Sparse matrix inversion with scaled lasso," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3385–3418, 2013.

[69] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[70] M. Tavassolipour, S. A. Motahari, and M.-T. M. Shalmani, "Learning of tree-structured gaussian graphical models on distributed data under communication constraints," *arXiv preprint arXiv:1809.08067*, 2018.

[71] S. L. Lauritzen, "Graphical models, volume 17 of oxford statistical science series," 1996.

[72] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of markov random fields from samples: Some observations and algorithms," *Lecture Notes in Computer Science*, vol. 5171, pp. 343–356, 2008.

[73] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4117–4134, 2012.

[74] M. Raginsky, "Strong data processing inequalities and *phi*-sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[75] G. Bresler and M. Karzand, "Learning a tree-structured ising model in order to make predictions," *arXiv preprint arXiv:1604.06749*, 2016.

[76] A. B. Tsybakov, "Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats," 2009.

[77] W. Schudy and M. Sviridenko, "Concentration and moment inequalities for polynomials of independent random variables," in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2012, pp. 437–446.

[78] G. Bennett, "Probability inequalities for the sum of independent random variables," *Journal of the American Statistical Association*, vol. 57, no. 297, pp. 33–45, 1962.

[79] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrika*, vol. 12, no. 1/2, pp. 134–139, 1918.

[80] H. Liu, J. Lafferty, and L. Wasserman, "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *Journal of Machine Learning Research*, vol. 10, no. Oct, pp. 2295–2328, 2009.

[81] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, ser. Lecture Notes in Computer Science, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer, March 4–7, pp. 265–284.