

Learning of Tree-Structured Gaussian Graphical Models on Distributed Data under Communication Constraints

Mostafa Tavassolipour, Seyed Abolfazl Motahari, and Mohammad-Taghi Manzuri Shalmani

Abstract—In this paper, learning of tree-structured Gaussian graphical models from distributed data is addressed. In our model, samples are stored in a set of distributed machines where each machine has access to only a subset of features. A central machine is then responsible for learning the structure based on received messages from the other nodes. We present a set of communication efficient strategies, which are theoretically proved to convey sufficient information for reliable learning of the structure. In particular, our analyses show that even if each machine sends only the signs of its local data samples to the central node, the tree structure can still be recovered with high accuracy. Our simulation results on both synthetic and real-world datasets show that our strategies achieve a desired accuracy in inferring the underlying structure, while spending a small budget on communication.

Index Terms—Structure learning, Chow-Liu algorithm, Gaussian Graphical Model.



1 INTRODUCTION

MANY modern systems acquire data at several repositories which are stored at different locations. In many situations, it is impossible to transfer the distributed data completely to a central machine due to communication constraints. Designing communication-efficient learning algorithms is desired to transfer enough information from repositories to the central machine and to reliably infer the learning model.

Many learning algorithms can be modified to run distributively at several machines to perform a learning task. There are many papers that propose distributed (parallel) version of various learning algorithms [1], [2], [3], [4]. However, some learning algorithms could not be efficiently parallelized on distributed data. For example, when each local machine has access to some attributes (dimensions) of data samples, many learning algorithms could not be run distributively. In such situations, typically, there exists a central machine which is responsible for running the learning algorithm. Due to communication constraints, local machines could not transmit their whole datasets to the central machine. In fact, the central machine may have access to a lossy compressed version or a subset of the original data. Thus, designing and analysis of the learning algorithms to make an appropriate trade-off between accuracy and the amount of communication is of great importance.

In general, three models can be considered for data in distributed settings: horizontal model, where the data are distributed across samples; vertical model, where data are distributed over dimensions; and hybrid model which is the combination of both horizontal and vertical models. Designing distributed learning systems for the horizontal model

has been addressed in many papers [3], [5]. In [6] and [7], the vertical model is studied from information theoretic point of view. This paper, in fact, addresses structure learning of tree-structured Gaussian Graphical Models (GGM) in the case of vertical model.

A GGM is a Markov Random Field (MRF) with normal variables which indeed form a joint normal distribution with mean μ and covariance matrix Q . If the ij -th component of Q^{-1} is zero, then the variables i and j are conditionally independent given the other variables. From this fact, one can construct the structure of the GGM by connecting node i to j iff the ij -th component of Q^{-1} is non-zero. GGMs are widely used in many applications such as gene regulatory networks [8], [9], [10], brain connectivity learning [11], etc.

In this paper, we analyze and propose structure learning methods based on the Chow-Liu algorithm over distributed data [12]. We assume that the data are split across dimensions (vertical model) among multiple machines. It worths mentioning that, unlike the horizontal model, the local machines are not capable of summarizing any statistics reflecting dependencies between dimensions in the vertical model. Hence, any inference requires some amount of communication between machines. This makes the problem nontrivial and challenging. We have assumed that the local machines are connected to a central machine over communication limited channels. Each machine compresses and transmits its local dataset to the central machine. Finally, the central machine by applying the Chow-Liu algorithm on the received distorted data, estimates the structure of the underlying GGM.

Interestingly, we convey an important message regarding accuracy of structure learning on distributed data for tree-structured GGMs: spending few bits per symbol is sufficient to transmit enough information to the central machine for the purpose of estimating the structure with

- M. Tavassolipour, S. A. Motahari, and M. T. Manzuri Shalmani are with the Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.

high accuracy. We justify this result by rigorous analysis and numerical experiments on synthetic and real datasets.

The paper is organized as follows. Section 2 provides a comprehensive literature review on structure learning and distributed statistical inference. In Section 3, we describe the problem in great detail and present our main contributions. Section 4 is devoted to structure learning using signs of the Gaussian data with theoretical analyses on its estimation error. In Section 5, a per-symbol quantization scheme is proposed and analyzed. We present the results of some experiments on real and synthetic datasets in Section 6. Section 7 concludes the paper.

2 RELATED WORK

2.1 Structure Learning

In the context of graphical models, inferring the underlying graph structure from data samples is of great importance. The structure learning is a *model selection* problem in which one estimates the underlying graph from i.i.d. samples drawn from some MRFs or Bayesian networks. The structure learning plays an important role in many applications such as reconstructing gene regulatory networks from gene expressions [13], [14], brain connectivity learning [11], relationship analysis in social networks [15], etc.

Structure learning of GGMs is equivalent to recovering the support (fill pattern) of the inverse covariance matrix (concentration matrix). The sparse structure estimation of the concentration matrix is discussed in many works [16], [17], [18], [19]. Among the sparse methods, maximizing the ℓ_1 -regularized likelihood is the most popular one. [17] and [18] proposed the graphical lasso (*glasso*), which finds the ML concentration matrix by an ℓ_1 regularization term. There are some coordinate-descent algorithms for solving the glasso problem [17], [20]. More recently, Hsieh et al. [21] proposed a new algorithm for solving the glasso problem enjoying super linear convergence rate. The consistency of the estimated graph using the glasso for high dimensional problems is studied in [22].

GGMs have an interesting property: one can obtain the neighborhood of each node by solving a linear regression problem for the corresponding variable on other variables. This approach of structure recovering is also known as *neighborhood selection* in the literature. For the sparse structures, combination of ℓ_1 regularization and the method of neighborhood selection is studied in [23]. In this method, an ℓ_1 regularized linear regression problem is solved separately for each node. Such a method may lead to inconsistencies between the inferred neighborhoods. Chen et al. [24] proposed some rules to resolve the inconsistencies.

Tan et al. [25] measured the complexity of the tree structures in view of the Chow-Liu algorithm. They also provided the analysis of the error exponent of the Chow-Liu algorithm on tree-structured GGMs in [26]. Moreover, they discussed the extreme structures which yield the best and worst error exponents. Resembling some features of this paper, the results in [26] differ from our results in two major aspects. First, our analysis is non-asymptotic while theirs is asymptotic in the number of samples. Second, we address the distributed version of the problem where only quantized data is available at the central machine. Having limited

access to the original data poses a significant challenge in the design and analysis.

2.2 Distributed Statistical Inference

The early works on distributed parameter estimation mainly focused on the asymptotic analysis of error exponents for given bit rates (see [6] and refs therein). More recently, studies are focused on characterizing the dependence between estimation performance and the communication constraint (see [27], [28], [29], [30], [31]). For example, Zhang et. al. [28] and Duchi et. al. [29] obtained some lower bounds on the minimax risks for distributed statistical parameter estimation under a given communication budget. They have studied the problem under single-round (non-interactive) and multiple-round (interactive) communication protocols between the local machines and the central one. A similar problem is addressed in some other papers such as [32] and [27]. Luo [32] showed that if each machine has a single one dimensional sample and transmits only one bit to the central machine, one can achieve the centralized minimax rate up to a constant factor for some specific problems. In a more recent work, Xu and Raginsky in [27] obtained lower bounds on Bayes risk in estimating parameters in a similar distributed setting. They studied the problem under both interactive and non-interactive communication protocols.

Some basic problems in machine learning such as classification, regression, hypothesis testing, etc. in distributed fashion are studied in [3], [33], [34]. Raginsky in [33] studied the classification and regression problem in distributed settings. He obtained an information-theoretic characterization of achievable predictor performance. He evaluated the results on non-parametric regression with Gaussian noise. The distributed hypothesis testing is studied by Amari [34] where a central machine makes decision on the correlation coefficient of two sequences stored in two different machines.

In this paper, we focus on the problem of distributed tree-structured GGM learning which is not studied previously. This work is similar to the problems studied by Ahlswede [35] and El-gamal [7] due to the fact that each local machine cannot estimate the parameters without any communication with other machines. This is in contrast to the most of the mentioned works where the local machines can have their own estimate of the underlying parameters. This fact makes the problem challenging as the local machines communicate with the central machine blindly.

3 PROBLEM STATEMENT AND PROPOSED METHODS

We are given n i.i.d. random vectors $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ drawn from a d -dimensional zero mean normal distribution $\mathcal{N}(\mathbf{0}, Q)$. Assume that the normal distribution can be factorized according to a tree model $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \dots, d\}$ is the set of nodes and \mathcal{E} is the set of edges. Factorization according to \mathcal{T} means that $(Q^{-1})_{jk} \neq 0$ if and only if $(j, k) \in \mathcal{E}$. Our goal is to find the structure of \mathcal{T} in a situation where data is stored in M machines such that each machine possesses some dimensions of the sample vectors. All machines are connected to a central machine via

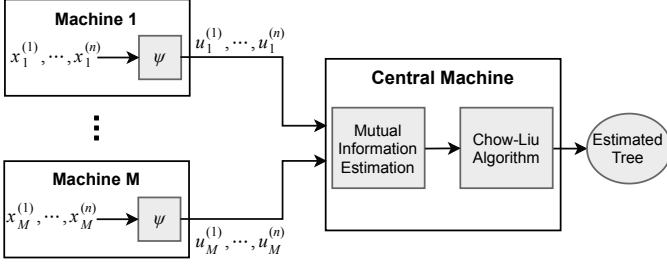


Fig. 1. The overall block diagram of our proposed model.

some communication limited links. This limitation makes it impossible for a machine to communicate its local dataset without any distortion to the central machine.

In this paper, we aim at proposing a communication efficient algorithm for estimating the underlying tree structure. In this setting, each machine transmits some information from its local dataset to the central machine which is responsible for estimating the structure from the received data.

Without loss of generality, we assume that the underlying normal distribution has zero mean and unit variance for all dimensions (i.e. $Q_{jj} = 1$). For convenience, we also assume that the machine \mathcal{M}_j contains the j -th dimension of the sample vectors. In this way, the number of machines is equal to the dimensionality of the normal distribution (i.e. $M = d$). We denote the j -th dimension of i -th sample by $x_j^{(i)}$, i.e. $\mathbf{x}^{(i)} = [x_1^{(i)}, \dots, x_d^{(i)}]^T$. Therefore, the local dataset on machine \mathcal{M}_j is $\{x_j^{(1)}, \dots, x_j^{(n)}\}$. Throughout the paper, we denote the quantized (compressed) version of $x_j^{(i)}$ by $u_j^{(i)}$.

We assume that the communication budget is R bits for each $x_j^{(i)}$. Thus, the overall communication cost for transmitting local datasets to the central machine is ndR bits.

The overall block diagram of our system is depicted in Fig. 1. In this system, \mathcal{M}_j encodes (quantizes) its local dataset using an R -bit encoder that can be represented by a function ψ_j which maps the samples to one of the predefined 2^{nR} reconstruction points denoted by $(u_j^{(1)}, \dots, u_j^{(n)})$, i.e. $(u_j^{(1)}, \dots, u_j^{(n)}) = \psi_j(x_j^{(1)}, \dots, x_j^{(n)})$. Since $(u_j^{(1)}, \dots, u_j^{(n)})$ can take only one of the 2^{nR} different reconstruction points, it can be transmitted to the central machine with nR bits. We assume that all machines incorporate the same encoding strategy. In this way, we have one encoder which is denoted by ψ .

Remark 1. The quantized datasets received by the central machine are not distributed according to the normal distribution. Moreover, in general, the tree structure is no longer a property of the new distribution. These facts make the recovery of the structure a rather challenging problem.

In case of having access to the original datasets, the central machine can run the Chow-Liu algorithm in [12] which gives the maximum likelihood (ML) tree structure [26]. In the Chow-Liu algorithm, the mutual information between any pair of vertices are estimated and used as the edges weights of a complete graph between vertices. The maximum weight spanning tree (MWST) gives the ML tree.

Therefore, the central part of the Chow-Liu algorithm is to estimate the mutual information efficiently.

In graph theory, there are two efficient algorithms for solving the MWST problem: Kruskal [36] and Prim [37] algorithms. Throughout this paper we incorporate the Kruskal algorithm for finding the MWST. In Kruskal algorithm, the edges weights are sorted in descending order and at each step an edge with the highest weight which does not form a cycle is added to the current forest. This algorithm continues until all vertices are covered. The output of this algorithm depends only on the order of edges weights.

In GGMs, the mutual information between any pair of variables, say x_j and x_k , is obtained by

$$I(x_j; x_k) = -\frac{1}{2} \ln(1 - \rho_{jk}^2), \quad (1)$$

where ρ_{jk} is the correlation coefficient between x_j and x_k . According to (1), one way to estimate the mutual information of two normal variables is to estimate their correlation coefficients first. In our problem setting where each variable is stored in a different machine, estimation of the correlation coefficients is a difficult task. This is due to the fact that one needs to calculate the statistic $\sum_i x_j^{(i)} x_k^{(i)}$. Computing such a statistic needs transmission of x_j and x_k samples to the central machine imposing high communication cost which is prohibitive in band-width limited networks.

3.1 Proposed Methods

Based on our system model, we need to design some efficient ways that machines communicate with the central machine and to implement an algorithm to recover the tree structure at the central machine. We propose two techniques to achieve these goals which are described as follow.

Sign Method

Each data point is quantized to a single bit by the sign function, i.e.

$$(\text{sign}(x_j^{(1)}), \dots, \text{sign}(x_j^{(n)})) = \psi(x_j^{(1)}, \dots, x_j^{(n)}).$$

Since the encoder maps each data point to a single reconstruction point independent of the other points, i.e. $u_j^{(i)} = \text{sign}(x_j^{(i)})$, the received data at the central machine is an i.i.d. sequence. However, as it mentioned earlier, the received data does not possess the tree structure of the original data.

Even though the Chow-Liu algorithm is only applicable for reconstructing tree structures, the central machine uses it to infer "a tree" structure embedded in the new distribution. This scheme is called *sign method*.

Through analysis presented in Section 4 we show that the tree reconstructed by the sign method is the true underlying structure with high probability. Our simulation results also support our analysis.

Per-symbol Quantization

In this scheme, each machine quantizes its data points to 2^R possible reconstruction points independent of the other points. Therefore, the received quantized points are i.i.d. and non-Gaussian. However, at the central machine, the

distribution is assumed to be normal and the correlations are estimated based on the received quantized datasets.

It worth mentioning that the encoder part of the sign method is a special case of the encoder used here with $R = 1$. In contrary, the tree reconstruction algorithms used at the central machine are different.

We provide an upper bound on the error of the correlation estimation in this case in Section 5. Simulation results also indicate by consuming a few bits for quantization, the estimated structure is often same as the structure obtained by the original data.

4 STRUCTURE LEARNING WITH SIGNS

In the sign method, the machine \mathcal{M}_j transmits $u_j^{(i)} = \text{sign}(x_j^{(i)})$ for $i = 1, \dots, n$ to the central machine. Note that since $x_j \sim \mathcal{N}(0, 1)$, u_j is a uniform Bernoulli variable over $\{-1, +1\}$.

The central machine receives the binary data from all machines to form the quantized dataset $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)}\}$ where $\mathbf{u}^{(i)} \in \{-1, +1\}^d$. Since the dimensions of the original normal vector \mathbf{x} are correlated, dimensions of \mathbf{u} are dependent as well. Although there is a simple map between the original normal vector and the signs, no closed form probability mass function (pmf) exists for $d \geq 4$. However, some approximations with desirable accuracies are proposed in [38].

Applying the Chow-Liu algorithm on $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)}\}$, the central machine obtains an estimate of the underlying structure. The remainder of this section is devoted to analyze the probability of incorrect structure recovery in a non-asymptotic regime.

4.1 Order Preserving of Mutual Information

Essential to the analysis, we show that the order of the true mutual information values between variables remains the same after applying the sign function. In this way, one can claim that by reliable estimating of the mutual information values of the signs, the true structure can be recovered using the Chow-Liu algorithm.

First, note that if $x_j \sim \mathcal{N}(0, 1)$ and $x_k \sim \mathcal{N}(0, 1)$ are jointly normal with the correlation coefficient ρ_{jk} , then the joint pmf of the corresponding signs u_j and u_k can be expressed as [38]

$$\begin{array}{c|cc} u_j \backslash u_k & -1 & +1 \\ \hline -1 & \theta_{jk}/2 & (1 - \theta_{jk})/2 \\ +1 & (1 - \theta_{jk})/2 & \theta_{jk}/2 \end{array} \quad (2)$$

where

$$\theta_{jk} = \frac{1}{2} + \frac{\arcsin(\rho_{jk})}{\pi}. \quad (3)$$

Therefore, the mutual information between u_j and u_k can be written as

$$I(u_j; u_k) = 1 - h(\theta_{jk}), \quad (4)$$

where $h(\cdot)$ is the binary entropy function given by

$$h(\theta) = -\theta \log(\theta) - (1 - \theta) \log(1 - \theta). \quad (5)$$

Lemma 1. *The sign function is an order preserving of the mutual information on Gaussian random variables.*

Proof. Consider two pairs of variables (x_j, x_k) and (x_r, x_s) from a GGM such that $I(x_j; x_k) > I(x_r; x_s)$. Let (u_j, u_k) and (u_r, u_s) be the corresponding sign variables. We need to show that $I(u_j; u_k) > I(u_r, u_s)$.

According to (1), if $I(x_j; x_k) > I(x_r; x_s)$ then $|\rho_{jk}| > |\rho_{rs}|$. Since the arcsin is a monotonic function, we have

$$\arcsin |\rho_{jk}| > \arcsin |\rho_{rs}|. \quad (6)$$

Since arcsin is odd, using (6) and (3) we have

$$|\theta_{jk} - \frac{1}{2}| > |\theta_{rs} - \frac{1}{2}|. \quad (7)$$

Consider the case where $\theta_{jk} > \frac{1}{2}$ and $\theta_{rs} > \frac{1}{2}$. Then, from (7) we have $\theta_{jk} > \theta_{rs}$. Since $h(\theta)$ is a descending function for $1/2 < \theta < 1$, we have $1 - h(\theta_{jk}) > 1 - h(\theta_{rs})$ which is the desired result.

For the case where $\theta_{jk} > \frac{1}{2}$ and $\theta_{rs} < \frac{1}{2}$, from (7) we have $\theta_{jk} > 1 - \theta_{rs}$. Since $h(\theta) = h(1 - \theta)$ and it is descending for $1/2 < \theta < 1$, we have $1 - h(\theta_{jk}) > 1 - h(1 - \theta_{rs})$ which is again the desired result. Similar arguments can be applied to the other two cases. \square

4.2 Probability of Incorrect Ordering

In the Chow-Liu algorithm on the signs, we rely on the estimates of the mutual information between all pairs of variables, $(j, k) \in \mathcal{V}^2$. It is shown in [7] that the following estimator for θ_{jk} is optimal in the sense that it is unbiased and has minimum variance (UMVE),

$$\hat{\theta}_{jk} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(u_j^{(i)} u_k^{(i)} = 1), \quad (8)$$

where $\mathbb{I}(\cdot)$ is the indicator function. By substituting $\hat{\theta}_{jk}$ into (4), an estimator for the mutual information of u_j and u_k is obtained which we denote it by $\hat{I}(u_j; u_k)$.

Let $e = (j, k) \in \mathcal{V}^2$ be a pair of nodes in the network. For simplicity in the notation, we use ρ_e , θ_e and I_e for ρ_{jk} , θ_{jk} and $I(u_j; u_k)$, respectively.

Definition 1 (Crossover Event). *Let e and e' be two pairs of nodes in the graph such that $I_e > I_{e'}$, the crossover event occurs when we estimate the mutual information in the reverse order, i.e. $\hat{I}_e \leq \hat{I}_{e'}$.*

The notion of *crossover event* was previously used by Tan et al. in [26]. The crossover event indicates a change in the ordering of the values of mutual information estimates. However, it does not necessarily lead to an incorrect tree recovery. On the other hand, if the estimated tree differs from the original tree \mathcal{T} , then at least one crossover event has occurred.

The mutual information in (4) is a monotonic function of $|\theta - 1/2|$ (see the proof of Lemma 1). Therefore, the probability of crossover event for pairs e and e' with $I_e > I_{e'}$ can be stated as

$$\Pr(\hat{I}_e \leq \hat{I}_{e'}) = \Pr\left(|\hat{\theta}_e - \frac{1}{2}| \leq |\hat{\theta}_{e'} - \frac{1}{2}|\right). \quad (9)$$

Lemma 2. *For the probability of the crossover event, it can be assumed $\theta_{jk} > 1/2$ for all $(j, k) \in \mathcal{V}^2$.*

Proof. Let $e = (j, k)$ and $e' = (r, s)$ be two arbitrary pairs of variables such that $I_e > I_{e'}$ or equivalently, $|\theta_e - 1/2| > |\theta_{e'} - 1/2|$. The probability of crossover event is

$$\Pr\left(\left|\hat{\theta}_e - \frac{1}{2}\right| \leq \left|\hat{\theta}_{e'} - \frac{1}{2}\right| \mid \left|\theta_e - \frac{1}{2}\right| > \left|\theta_{e'} - \frac{1}{2}\right|\right). \quad (10)$$

Let us assume $\theta_e < 1/2$ and $\theta_{e'} > 1/2$ (the other cases can be argued similarly). We define new variable $\tilde{u}_j = -u_j$. It is clear that the joint pmf of \tilde{u}_j and u_k is given by (2) with parameter $\tilde{\theta} = 1 - \theta_e$. Thus, $|\theta_e - 1/2| = |\tilde{\theta}_e - 1/2|$. Similarly, there exists the following relation between the estimators $\hat{\theta}_e$ and $\hat{\tilde{\theta}}$

$$\begin{aligned} \hat{\tilde{\theta}} &= \frac{1}{n} \sum_{i=1}^n \mathbb{I}(\tilde{u}_j^{(i)} u_k^{(i)} = 1) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(-u_j^{(i)} u_k^{(i)} = 1) \\ &= \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{I}(u_j^{(i)} u_k^{(i)} = 1)) = 1 - \hat{\theta}_e. \end{aligned}$$

Thus, $|\hat{\theta}_e - 1/2| = |\hat{\tilde{\theta}} - 1/2|$. Therefore, the crossover probability in (10) can be expressed as

$$\Pr\left(\left|\hat{\tilde{\theta}} - \frac{1}{2}\right| \leq \left|\hat{\theta}_{e'} - \frac{1}{2}\right| \mid \left|\tilde{\theta} - \frac{1}{2}\right| > \left|\theta_{e'} - \frac{1}{2}\right|\right),$$

where both $\tilde{\theta}$ and $\theta_{e'}$ are greater than $1/2$. \square

Lemma 2 shows that without loss of generality, we can assume all θ_{jk} s are greater than $1/2$ which is equivalent to assuming all correlations are positive (see equation (3)). Note that, the condition $\theta_{jk} > \frac{1}{2}$ does not imply that the estimator $\hat{\theta}_{jk}$ is also greater than $1/2$.

In the following lemma, to make the exposition of the ideas easier, we assume both $\hat{\theta}_e$ and $\hat{\theta}_{e'}$ are greater than $1/2$. However, in the supplementary material, we provide an upper bound on (9) for all cases.

Lemma 3. Let $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ be n i.i.d. samples drawn from a d -dimensional GGM. Assume that the variables have zero means and unit variances. Then, the probability of crossover event for two pairs $e = (j, k)$ and $e' = (r, s)$ with $\theta_e > \theta_{e'}$, is upper bounded by

$$\Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}) \leq e^{-nE}, \quad (11)$$

where $E = \ln(p_0 + 2\sqrt{p_1 p_2})$ and

$$p_0 = \Pr(u_j u_k = u_r u_s), \quad (12)$$

$$p_1 = \Pr(u_j u_k = -1, u_r u_s = 1), \quad (13)$$

$$p_2 = \Pr(u_j u_k = 1, u_r u_s = -1). \quad (14)$$

Moreover, the exponent E is the tightest possible, i.e.,

$$E = \lim_{n \rightarrow \infty} -\frac{1}{n} \ln \Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}). \quad (15)$$

Proof. Consider two pairs of nodes $e = (j, k)$ and $e' = (r, s)$. By defining a random variable T_i as

$$T_i = \mathbb{I}(u_r^{(i)} u_s^{(i)} = 1) - \mathbb{I}(u_j^{(i)} u_k^{(i)} = 1), \quad (16)$$

the probability of crossover event can be written as

$$\begin{aligned} \Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}) &= \Pr\left(\sum_{i=1}^n T_i \geq 0\right) \\ &= \Pr\left(e^{\lambda \sum_{i=1}^n T_i} \geq 1\right), \quad \lambda > 0 \\ &\stackrel{(a)}{\leq} \mathbb{E}\left[e^{\lambda \sum_{i=1}^n T_i}\right] \\ &= \left(\mathbb{E}\left[e^{\lambda T}\right]\right)^n \\ &= \left(p_0 + p_1 e^{\lambda} + p_2 e^{-\lambda}\right)^n, \end{aligned}$$

where the inequality (a) is by Markov's inequality. The random variable T can take values $[0, 1, -1]$ with probabilities $[p_0, p_1, p_2]$ defined in (12)-(14). By minimizing the last expression for $\lambda > 0$, we obtain the Chernoff bound as follows

$$\Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}) \leq (p_0 + 2\sqrt{p_1 p_2})^n = e^{-nE}, \quad (17)$$

where $E = \ln(p_0 + 2\sqrt{p_1 p_2})$. Incorporating Theorem 2.1.24 in [39], the exponent E is indeed tight. \square

Unfortunately, there is no closed-form solution for the probabilities in (12)-(14). However, when e and e' share a common variable, these probabilities can be obtained analytically. For example, if $u_k = u_s$ (i.e. u_k is the common variable between e and e'), then the probabilities are given by [38]

$$p_0 = \frac{1}{2} + \frac{\arcsin \rho_{jk} \rho_{ks}}{\pi}, \quad (18)$$

$$p_1 = \frac{1}{4} + \frac{-\arcsin \rho_{jk} + \arcsin \rho_{ks} - \arcsin \rho_{jk} \rho_{ks}}{2\pi}, \quad (19)$$

$$p_2 = \frac{1}{4} + \frac{\arcsin \rho_{jk} - \arcsin \rho_{ks} - \arcsin \rho_{jk} \rho_{ks}}{2\pi}. \quad (20)$$

In particular, for the star structure, where all the true edges share a common node, the bound of Lemma 3 can be calculated in a closed-form.

In the following lemma, we propose another upper bound on the probability of the crossover event using Hoeffding's bound. This bound is not tight, but it yields a closed form expression which can be used to obtain a closed form bound on the probability of incorrect tree estimation.

Lemma 4. The probability of the crossover event for two pairs e and e' with $\theta_e > \theta_{e'}$, is upper bounded by

$$\Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}) \leq e^{-\frac{1}{2} n \Delta \theta_{e,e'}^2}, \quad (21)$$

where $\Delta \theta_{e,e'} = \theta_e - \theta_{e'}$.

Proof. Since $\hat{\theta}_e$ and $\hat{\theta}_{e'}$ are unbiased estimators for θ_e and $\theta_{e'}$, we have

$$\Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}) = \Pr(\hat{\theta}_{e'} - \hat{\theta}_e - \mathbb{E}[\hat{\theta}_{e'} - \hat{\theta}_e] \geq \Delta \theta_{e,e'}).$$

Defining variable T_i as (16), we can write the above probability as

$$\Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}) = \Pr\left(\frac{1}{n} \sum_{i=1}^n (T_i - \mathbb{E}[T_i]) \geq \Delta \theta_{e,e'}\right).$$

It is clear that $T_i \in \{-1, 0, 1\}$, thus it is bounded in interval $[-1, 1]$. Using the Hoeffding's inequality we can obtain an upper bound on the probability of crossover event as

$$\Pr(\hat{\theta}_e \leq \hat{\theta}_{e'}) \leq e^{-\frac{1}{2}n\Delta\theta_{e,e'}^2}.$$

□

4.3 Probability of Incorrect Recovery

In this section, we are interested in bounding $\Pr(\hat{\mathcal{T}} \neq \mathcal{T})$ where $\hat{\mathcal{T}} = (\mathcal{V}, \hat{\mathcal{E}})$ refers to the estimated tree from our proposed sign method. If we assume that any error in the ordering of mutual information estimates may lead to incorrect recovery of the tree structure, then by the union bound we have

$$\Pr(\hat{\mathcal{T}} \neq \mathcal{T}) \leq \sum_{e, e' \in \mathcal{V}^2} e^{-\frac{1}{2}n\Delta\theta_{e,e'}^2} \quad (22)$$

In the following theorem, we improve on the preceding bound by removing some of the crossover events. Moreover, we obtain a more compact and suitable formula for the bound.

Theorem 1. *Let $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ be n i.i.d. samples drawn from a d -dimensional tree-structured GGM. We construct n binary vectors $\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)}\}$ where $\mathbf{u}^{(i)} \in \{-1, +1\}^d$ is the sign vector of $\mathbf{x}^{(i)}$. Assume for all $(j, k) \in \mathcal{E}$, $\alpha \leq \rho_{jk} \leq \beta$ where $0 < \alpha < \beta < 1$. Then the probability of incorrect tree recovering using the Chow-Liu algorithm via the binary vectors is upper bounded by*

$$\Pr(\hat{\mathcal{T}} \neq \mathcal{T}) \leq d^3 e^{-\frac{1}{2}nh^2(\alpha, \beta)}, \quad (23)$$

where $h(\alpha, \beta) = \frac{1}{\pi} (\arcsin \alpha - \arcsin \alpha\beta)$.

The rest of this section is devoted to the proof of the theorem. By removing any edge of the tree, \mathcal{T} splits into two separate sub-trees. Let us assume that in the procedure of tree reconstruction, a true edge $e = (j, k)$ does not appear in $\hat{\mathcal{T}}$. Removing e from the tree splits \mathcal{T} into $\mathcal{T}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{T}_2 = (\mathcal{V}_2, \mathcal{E}_2)$. Hence, there should be an edge $e' = (r, s) \in \hat{\mathcal{E}}$ which is not in \mathcal{E} and connects \mathcal{T}_1 and \mathcal{T}_2 . This is due the fact that we constrain the tree to be connected. The following lemma shows that e is in fact the strongest edge connecting \mathcal{T}_1 and \mathcal{T}_2 .

Lemma 5. *Consider a tree-structured GGM $\mathcal{T} = (\mathcal{V}, \mathcal{E})$. Let $e \in \mathcal{E}$ be an edge which connects two sub-trees $\mathcal{T}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{T}_2 = (\mathcal{V}_2, \mathcal{E}_2)$. Then for any node pair (r, s) where $r \in \mathcal{V}_1$ and $s \in \mathcal{V}_2$, we have*

$$\theta_e \geq \theta_{rs}.$$

Proof. In any tree-structured Gaussian distributions, we have [40]

$$\rho_{rs} = \prod_{e \in \text{Path}(r, s)} \rho_e, \quad (24)$$

where $\text{Path}(r, s)$ is the set of edges in the path connecting r and s in the tree. This means that the correlation of (r, s) is less than any edge in the path connecting them. Since $r \in \mathcal{V}_1$ and $s \in \mathcal{V}_2$, then the path connecting r and s must include

the edge e . On the other hand, according to (3), if $\rho_{rs} \leq \rho_e$ then $\theta_{rs} \leq \theta_e$ since the function \arcsin is monotonic. □

According to the Kruskal algorithm, the estimated $\hat{\theta}_{jks}$ for all $(j, k) \in \mathcal{V}^2$ are sorted in a descending order. Scanning from the top of the list, an edge is selected as a part of the tree if it does not create any cycle with the previous picked edges. In the case of error, a true edge like $e \in \mathcal{E}$ is replaced by another false edge $e' \notin \mathcal{E}$ if $\hat{\theta}_{e'} > \hat{\theta}_e$ and e' connects the two subtrees created by removing e . On the other hand, from Lemma 1, we know that $\theta_{e'} < \theta_e$. Thus, replacing e by e' implies a crossover event on them.

Let $e \in \mathcal{E}$ be an edge which connects two sub-trees $\mathcal{T}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{T}_2 = (\mathcal{V}_2, \mathcal{E}_2)$. Let

$$\mathcal{C}(e) = \{e' \mid e' \notin \mathcal{E} \text{ and connects } \mathcal{T}_1 \text{ and } \mathcal{T}_2\},$$

be the set of all candidate false edges which can substitute the edge e in the estimated tree. Then, from the union bound and incorporating Lemma 4 and Lemma 5, we have

$$\begin{aligned} \Pr(\mathcal{T} \neq \hat{\mathcal{T}}) &\leq \sum_{e \in \mathcal{E}} \Pr(e \notin \hat{\mathcal{E}}) \\ &\leq \sum_{e \in \mathcal{E}} \sum_{e' \in \mathcal{C}(e)} \Pr(\hat{\theta}_{e'} \geq \hat{\theta}_e) \\ &\leq \sum_{e \in \mathcal{E}} \sum_{e' \in \mathcal{C}(e)} e^{-\frac{1}{2}n\Delta\theta_{e,e'}^2}. \end{aligned} \quad (25)$$

If we obtain a lower bound on $\Delta\theta_{e,e'}$ which is independent of the tree structure, say $\Delta_0 \leq \Delta\theta_{e,e'}$, then we have

$$\Pr(\mathcal{T} \neq \hat{\mathcal{T}}) \leq d^3 e^{-\frac{1}{2}n\Delta_0^2}. \quad (26)$$

To this end, we need the following definition.

Definition 2 (Strongest Rival). *The strongest rival of an edge e is an edge $e^* \in \mathcal{C}(e)$ with the highest θ_{e^*} .*

To find the strongest rival for an edge $e = (j, k)$, we use the correlation decay property of the tree-structured GGMs. In fact, it is easy to see that e^* is either an edge connecting node j to one of the neighbors of node k or vice versa. In other words, the strongest rival of an edge e , is one of the its neighbor edges (two edges are neighbor if share a common node). Fig. 2 illustrates the strongest rival of the edge $e = (j, k)$ in a sample tree. The following lemma gives a lower bound on $\Delta\theta_{e,e^*}$.

Lemma 6. *Let e^* be the strongest rival of an edge $e \in \mathcal{E}$. If the correlation coefficients ρ_{jk} for all $(j, k) \in \mathcal{E}$ satisfy $\alpha \leq \rho_{jk} \leq \beta$ where $0 < \alpha < \beta < 1$, then $\Delta\theta_{e,e^*} \geq h(\alpha, \beta)$ where*

$$h(\alpha, \beta) = \frac{1}{\pi} (\arcsin \alpha - \arcsin \alpha\beta). \quad (27)$$

Proof. Let $e = (j, k)$. Since $e^* \notin \mathcal{E}$ and it connects a neighbor of j to k (or vice versa), using (24) we have

$$\alpha\rho_e \leq \rho_{e^*} \leq \beta\rho_e.$$

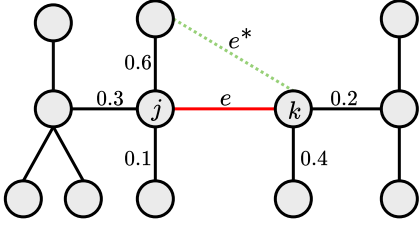


Fig. 2. Illustration of the strongest rival edge for $e = (j, k)$ in a sample tree. The weights are the correlation coefficients.

Therefore, in order to obtain a lower bound on $\Delta\theta_{e,e^*}$, we need to solve the following constrained optimization problem:

$$\begin{aligned} \min_{\rho_e, \rho_{e^*}} \quad & \arcsin \rho_e - \arcsin \rho_{e^*}, \\ \text{subject to:} \quad & \alpha \rho_e \leq \rho_{e^*} \leq \beta \rho_e, \\ & \alpha \leq \rho_e \leq \beta. \end{aligned} \quad (28)$$

Defining a new parameter $\eta = \rho_{e^*} / \rho_e$, the above optimization problem can be written as

$$\begin{aligned} \min_{\rho_e, \eta} \quad & \arcsin \rho_e - \arcsin \eta \rho_e, \\ \text{subject to:} \quad & \alpha \leq \rho_e \leq \beta \\ & \alpha \leq \eta \leq \beta. \end{aligned} \quad (29)$$

Taking derivative with respect to ρ_e and η we have,

$$\begin{aligned} \frac{\partial}{\partial \rho_e} &= \frac{1}{\sqrt{1-\rho_e^2}} - \frac{\eta}{\sqrt{1-\eta^2\rho_e^2}} > 0, \\ \frac{\partial}{\partial \eta} &= \frac{-\rho_e}{\sqrt{1-\eta^2\rho_e^2}} < 0. \end{aligned}$$

Hence, the minimum is attained at $(\rho_e, \eta) = (\alpha, \beta)$. Thus, $\rho_{e^*} = \eta \rho_e = \alpha \beta$. By substituting $(\rho_e, \rho_{e^*}) = (\alpha, \alpha \beta)$ into $\Delta\theta_{e,e^*}$ the lower bound in (27) is obtained. \square

With the above lemma, we obtain $\Delta_0 = h(\alpha, \beta)$ in (26). Hence, we complete the proof of Theorem 1.

Remark 2. Since we have not imposed any constraint on the tree structure, the upper bound in Theorem 1 is obtained for the worst case of the structure resulting the prefactor of d^3 which is tight for the chain structure. However, the prefactor can be reduced in many cases such as the star structure which requires d^2 as the prefactor. However, the prefactor has negligible effect if sufficiently large sample size n is available.

Remark 3. Replacing Δ_0 for all the rival edges is not optimal in general. However, in the special case of the star structure with equal edge weights, for instance, it is tight. Hence, some prior knowledge about the tree structure can lead to possibly better bounds.

5 STRUCTURE LEARNING WITH DIRECT CORRELATION ESTIMATION

The Chow-Liu algorithm requires a good estimate of the mutual information between pairs of variables. In GGMs,

according to (1), quality of the mutual information estimation depends on the accuracy of correlation estimation. In the sign method (Section 4), we do not estimate the correlations between normal variables based on the binary data. By contrast, in this section, we aim to quantize data with R bits and to estimate the correlation coefficients between the normal variables using the quantized data.

Instead of estimating the mutual information in (1) using the quantized data, we use an unbiased estimator for ρ^2 . This is due to the fact that obtaining an unbiased estimator for the mutual information is a hard problem. The following estimator for ρ^2 is unbiased:

$$\widehat{\rho^2} = \frac{n}{n+1} \left(\bar{\rho}^2 - \frac{1}{n} \right), \quad (30)$$

where $\bar{\rho}$ is the sample correlation coefficient, i.e.

$$\bar{\rho} = \frac{1}{n} \sum_{i=1}^n x_j^{(i)} x_k^{(i)}. \quad (31)$$

In our problem setting, we calculate $\bar{\rho}$ using the R -bit quantized data. We define

$$\bar{\rho}_q = \frac{1}{n} \sum_{i=1}^n u_j^{(i)} u_k^{(i)}, \quad (32)$$

which is the sample correlation coefficient of the quantized data. Obviously, by increasing the bit rate R , $\bar{\rho}_q$ approaches $\bar{\rho}$. To measure the performance of the quantization method, we define the relative error as

$$\text{err}_{\text{rel}} \triangleq \mathbb{E} [|\bar{\rho} - \bar{\rho}_q|], \quad (33)$$

where the expectation is taken over all original and quantized variables (x_j, x_k, u_j, u_k) . We call the error function in (33) as *relative correlation error*. It measures the expected difference between sample correlation coefficient on the original and quantized data. Our main objective is to obtain a good estimate for the true correlation coefficient. Thus, we define the estimation error of a quantization method as

$$\text{err}_{\text{est}} = \mathbb{E} [|\rho - \bar{\rho}_q|], \quad (34)$$

where the expectation is taken over the quantized variables (u_j, u_k) .

Although we aim at designing a quantization method which have a small estimation error, most of existing efficient source coding schemes are designed to minimize the *reconstruction error*. More precisely, denoting the original and quantized versions by x and u . These methods quantize the data with the following constraint

$$\mathbb{E} [(x - u)^2] \leq D, \quad (35)$$

where D is the maximum allowable reconstruction error. For example, there exists an optimal coding scheme for normal variables based on the rate-distortion theory [40]. In our problem setting, minimizing the reconstruction error is not the main objective. Our goal is to minimize the error functions defined in (33) and (34). However, the following theorem guarantees that upper bounding the reconstruction error results in upper bounding the relative correlation error.

Theorem 2. Let $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$ be two sets of i.i.d. samples from any joint distribution $P(x, y)$ such that x

and y have zero means and unit variances. If $\{u_1, \dots, u_n\}$ and $\{v_1, \dots, v_n\}$ are the corresponding encoded sequences by any quantizer with

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (x_i - u_i)^2 \right] \leq D_1,$$

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (y_i - v_i)^2 \right] \leq D_2,$$

then

$$\text{err}_{\text{rel}} \leq \sqrt{D_1} + \sqrt{D_2} + \sqrt{D_1 D_2}. \quad (36)$$

The proof of the theorem is presented in Appendix A. The above theorem suggests that designing a coding scheme with small reconstruction error yields a desirable relative correlation error.

Lemma 7. *The estimation error in (34) can be upper bounded as follow*

$$\text{err}_{\text{est}} \leq \sqrt{\frac{1 + \rho^2}{n}} + \text{err}_{\text{rel}}. \quad (37)$$

Proof. Using the triangle inequality, we have

$$\mathbb{E} [|\rho - \bar{\rho}_q|] \leq \mathbb{E} [|\rho - \bar{\rho}|] + \mathbb{E} [|\bar{\rho} - \bar{\rho}_q|]. \quad (38)$$

On the other hand, using the Jensen's inequality and the convexity of the square function, we have

$$\mathbb{E}^2 [|\rho - \bar{\rho}|] \leq \mathbb{E} [(\rho - \bar{\rho})^2] = \frac{1 + \rho^2}{n}. \quad (39)$$

By combining (38) and (39), the bound in (37) is obtained. \square

Next, we obtain upper bounds on err_{est} and err_{rel} for our per-symbol quantization method. In this method, we first quantize each sample to a discrete random variable $u \in \mathcal{U}$ where $|\mathcal{U}| = 2^R$. Thus, each sample u can be encoded by R bits. Assume we have a standard normal random variable $x \sim \mathcal{N}(0, 1)$ which we want to quantize by R bits. To this end, we construct 2^R equally probable bins over the real axis and use bins' centroids as the reconstruction points. These centroids are indeed the members of \mathcal{U} which is used for compression of x . To compress x , we send index of the bin that x belongs to it by R bits.

For the standard normal distribution, u can have a value from the codebook $\mathcal{U} = \{c_1, \dots, c_{2^R}\}$ where c_i is the centroid of i -th bin. Denoting the i -th bin interval by (a_i, a_{i+1}) , its centroid c_i is obtained by

$$c_i = \frac{2^R}{\sqrt{2\pi}} \left(e^{-a_i^2/2} - e^{-a_{i+1}^2/2} \right), \quad i = 1, \dots, 2^R. \quad (40)$$

Interval boundaries $\{a_i\}$ are obtained as follow. We first set $a_1 = -\infty$. Then, given a_{i-1} , we iteratively obtain a_i as the solution of the following equation:

$$\int_{a_{i-1}}^{a_i} \mathcal{N}(x; 0, 1) dx = 2^{-R}, \quad i = 2, \dots, 2^R + 1.$$

The expectation of reconstruction error for this coding scheme is obtained as follows,

$$\begin{aligned} \mathbb{E} [(x - u)^2] &= 1 + \sigma_u^2 - 2 \sum_{i=1}^{2^R} \int_{-\infty}^{+\infty} c_i x p(x) p(u|x) dx \\ &= 1 + \sigma_u^2 - 2 \sum_{i=1}^{2^R} \int_{a_i}^{a_{i+1}} c_i x p(x) dx \\ &= 1 + \sigma_u^2 - 2 \cdot 2^{-R} \sum_{i=1}^{2^R} c_i^2 \\ &= 1 - \sigma_u^2, \end{aligned} \quad (41)$$

where σ_u^2 is the variance of discrete variable u . Evidently, by increasing the bit rate R , σ_u^2 approaches 1. After encoding the normal variables using the above method, the central machine estimates the correlation between x_j and x_k for any $(j, k) \in \mathcal{V}^2$ using (32). Then, it estimates ρ^2 by substituting $\bar{\rho}_q$ in (30). Finally, by applying the Chow-Liu algorithm, the structure of underlying GGM is estimated.

According to Theorem 2 and the reconstruction error in (41), the relative correlation estimation error in (33) is upper bounded by

$$\text{err}_{\text{rel}} \leq 2\sqrt{1 - \sigma_u^2} + 1 - \sigma_u^2. \quad (42)$$

The variance σ_u^2 is a function of the bit rate R , but it does not have an explicit closed form expression. By incorporating Lemma 7, the estimation error of per-symbol quantizer is upper bounded by

$$\text{err}_{\text{est}} \leq 2\sqrt{1 - \sigma_u^2} + 1 - \sigma_u^2 + \sqrt{\frac{1 + \rho^2}{n}}. \quad (43)$$

In Section 6, we will evaluate the performance of this method for structure estimation on real and synthetic datasets.

6 EXPERIMENTS

In this section, we evaluate the performance of our proposed structure learning methods empirically on some synthetic and real-world datasets. The results show that quantizing samples to few bits is enough to estimate the underlying structure of the model. In all the experiments, double-precision floating-points (64 bit) are used for the original data.

6.1 Synthetic Data

Synthetic data are generated from a random tree with d nodes. Then, a random weight is assigned to each edge of the tree which corresponds to the correlation coefficient between endpoint variables of the edge. The correlation coefficient between any pair of variables which are not neighbor, can be obtained by (24). Thus, using this weighted tree, the covariance matrix of the GGM is obtained and n i.i.d. samples are drawn from the underlying normal distribution. Finally, dimensions of the data are distributed among d machines.

In Fig. 3, the performance of the sign method and per-symbol quantization for different values of n and R is plotted. In this experiment, the underlying GGM has 20

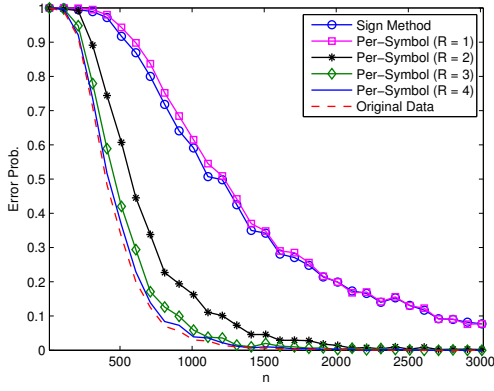


Fig. 3. The structure estimation error for different values of R and n . Here, the random GGM has 20 nodes.

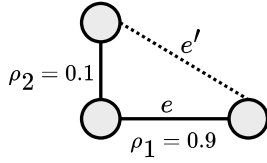


Fig. 4. A sample sub-tree of three nodes for evaluation of the crossover error bound.

variables. To approximate the probability of error, for each sample size n , we run the algorithms 1000 times and count the number of incorrect estimated trees. The experiment shows that recovering the structure from the sign method yields better performance than the per-symbol method for $R = 1$. Interestingly, the error of 4-bit per-symbol is very close to the non-quantized (original data) curve. This means that the accuracy of correlation estimation using 4-bit quantized data is sufficient to achieve the centralized performance in structure estimation.

6.1.1 Evaluation of the Error Bounds

We first focus on a simple tree with three nodes and correlations coefficients $\rho_1 = 0.9$ and $\rho_2 = 0.1$, as depicted in Fig. 4. A crossover event happens if the estimate of the mutual information associated to e' exceeds that of e . To evaluate the error bounds of this event obtained in Lemma 3 and Lemma 4, we also provide the exact error which can be calculated by a brute force summation of the tail probability of $\Pr(\hat{\theta}_e \leq \hat{\theta}_{e'})$. Fig. 5 depicts the probability of crossover event versus the number of samples. As can be seen, the upper bound of Lemma 3 converges to the exact error faster than the bound of Lemma 4.

In Fig. 6, the exponent of exact error, Chernoff bound (Lemma 3) and Hoeffding bound (Lemma 4) for the structure of Fig. 4 are compared. In the figure, the quantity $-\frac{1}{n} \ln \Pr(\hat{\theta}_e \leq \hat{\theta}_{e'})$ is plotted for various sample sizes. As stated in Lemma 3, the bound obtained based on Chernoff bound is tight in the exponent. However, Hoeffding bound is not tight in this case.

Fig. 7 shows the probability of incorrect recovery of the tree structure using the sign method as a function of sample size. We have used a star structured tree with 20 nodes

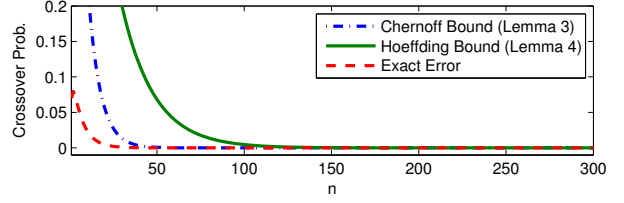


Fig. 5. Probability of the crossover event for e and e' in Fig. 4.

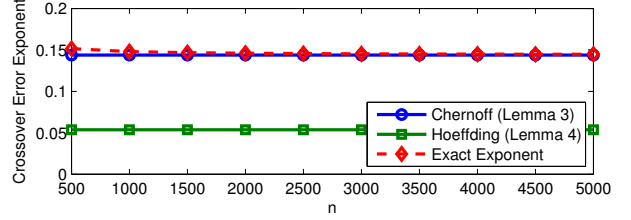


Fig. 6. The crossover error exponent for e and e' in Fig. 4.

and correlations of 0.5 which due to Remark 3 is the worst structure.

In Fig. 8, the exponent of the bound of Theorem 2 on err_{rel} is plotted as a function of bit rate R . In this experiment, the true correlation coefficient is $\rho = 0.5$ and the sample size is $n = 1000$. The empirical error curve is obtained by averaging over 1000 runs. In the figure, the y axis represents the quantity $-\frac{1}{R} \ln(\text{err}_{\text{rel}})$. As can be seen from the figure, the upper bound is not tight in the exponent for Gaussian data. Note that the error bound in Theorem 2 is valid for any distribution and any quantization method.

6.1.2 Quality versus Quantity

Generally, quantization of data samples decreases the accuracy of any parameter estimation. The quantization of all samples, which is considered so far, may not be the

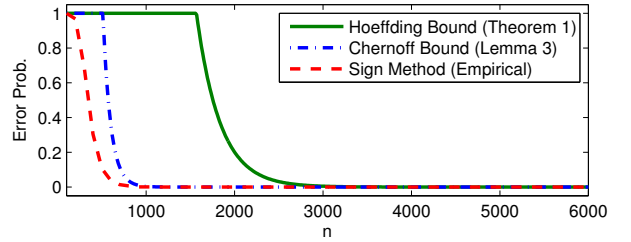


Fig. 7. Probability of incorrect tree estimation for the star structure.

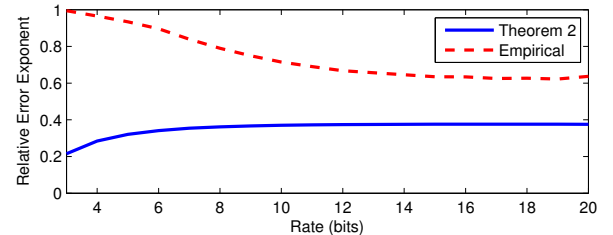


Fig. 8. Relative error exponent of per-symbol method.

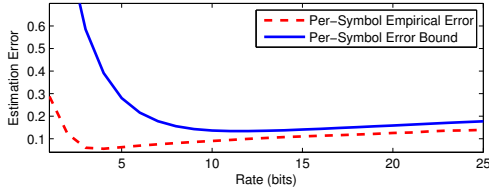


Fig. 9. The mean absolute error of correlation estimation. The total communication cost for each machine is limited to $K = 1000$ bits.

best strategy for decreasing the communication complexity. In fact, if the budget of total number of transmission bits is fixed, then one may want to sub-sample from the local datasets and allocate the available bits to these samples and discard the rest.

Based on the bound in (43) and through an experiment, we show that there is a trade-off between the quality of quantized samples and the size of sub-sampled data to achieve the best performance.

For example, assume that the budget of communication is $K = 1000$ bits and the number of local data samples is $n = 1000$. This means each local machine can, for instance, transmit 1000 samples which are quantized by 1 bit. However, machines can select the first 500 samples and quantize them to 2 bits. Which method is better in the sense of minimizing the estimation error err_{est} ?

To answer the question, we have simulated our proposed algorithm with $K = 1000$ and $n = 1000$. In Fig. 9, the estimation error in (34) for the estimator in (32) is plotted for various bit rates. In this experiment, the true value of correlation is 0.5. As can be seen from the figure, the error is minimized when $R = 4$ bits are used for quantization which is correspond to the sub-sampling size of 250. The figure also shows that the estimation with large number of highly distorted (quantized) samples is inefficient as well as estimating with low number of high quality samples. In fact, this experiment suggests that in situations with large local datasets and limited communication cost, the optimal strategy is to quantize some portion of the whole dataset with an acceptable distortion. The upper bound in (43) for the per-symbol scheme is also plotted for the sake of comparison.

6.2 Real-World Dataset: Skeleton Recovering

To assess our methods on real datasets, the MAD¹ dataset is used. This dataset is designed for human activity recognition and event detection in the computer vision area [41], [42]. The MAD dataset is generated by a Microsoft Kinect sensor in indoor environment. In the dataset, there are 20 sensors attached to 20 joints of the human body. Each sensor records 3D coordinate of its corresponding joint while the subject does an activity. The MAD dataset has three modalities includes RGB video, 3D depth, and skeleton (3D coordinate of the joints). In this experiment, we have used the skeleton modality.

In the dataset, 20 subjects perform 35 different actions (e.g. jumping, walking, running, etc.) and each subject re-

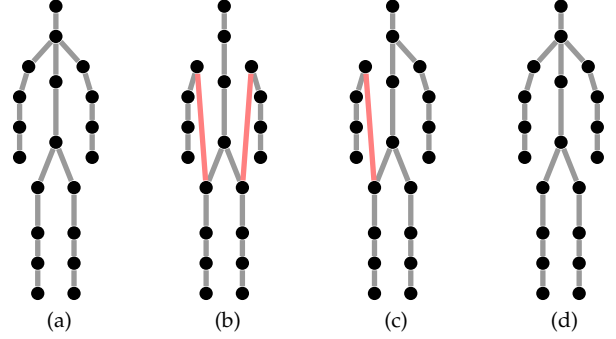


Fig. 10. Structure learning of the human body skeleton on x dimension of the MAD dataset. (a) The true human body skeleton. The estimated structures using quantized data with rates 1, 3, and 6 bits are shown in (b), (c), and (d), respectively. The skeleton of (b) is obtained by both signs and 1-bit per-symbol methods.

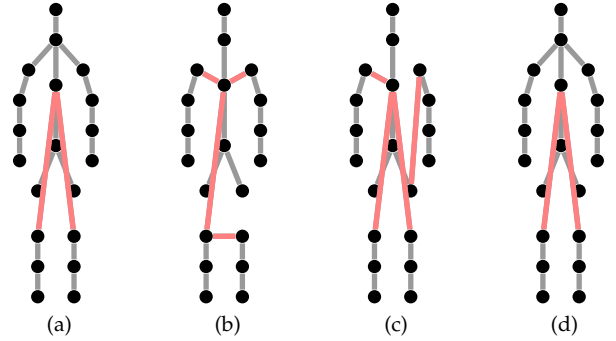


Fig. 11. Structure learning of the human body skeleton on z dimension of the MAD dataset. The recovered structure using the original data is shown in (a). The estimated structures using quantized data with 1-bit sign method, 1-bit per-symbol method, and 7-bit per-symbol are shown in (b), (c), and (d), respectively.

peats the actions twice. Finally, the skeleton dataset contains 243586 3D coordinates per joint.

We assume that the skeleton dataset follows from a tree-structured GGM which its structure is identical to the human body skeleton as depicted in Fig. 10-(a). This assumption intuitively makes sense for such a dataset. Gaussian assumption for similar datasets are proposed in [43] and [44].

In this experiment, the skeleton dataset is quantized to several bit rates using the proposed per-symbol quantizer. Fig. 10 shows the results for bit rates 1, 3, 5, and 6 bits. Applying the Chow-Liu algorithm on x dimension of the original (non-quantized) data, perfectly recovers the body skeleton. Quantizing the x dimension to 1 bit (using per-symbol and the sign method) results in merely two disagreement edges as showed in Fig. 10-(b). Quantizing to 3 bits has only one incorrect edge and quantizing to 6 bits recovers the body skeleton perfectly.

A similar experiment is performed on the z dimension. The results are depicted in Fig. 11. Interestingly, the z dimension does not follow a tree structured GGM even for the case where the original data is available. However, as the bit rate increases the original structure can be recovered reliably. We have not presented the experiment on the y dimension here. This is due to the fact that the structure inferred from the original data has no relation with the human skeleton.

1. Multi-modal Action Database (available at: <http://www.humansensing.cs.cmu.edu/mad>)

7 CONCLUSION

In this paper, we have studied the structure learning of tree-structured GGMs on distributed datasets. Due to communication constraints, we have jointly designed quantization and learning algorithms to achieve high accuracy in inferring the underlying tree structure. In particular, we have proposed two methods for compressing the local datasets: sign method and per-symbol quantizer.

Being rather simple and intuitive, the experimental results show that the per-symbol quantizer yields high accuracy for structure estimation by spending a few bits per sample. Pushing down the number of bits per sample to one, we have proved through experiments and analytical reasoning that even in this case, one can obtain exponentially decaying error probabilities in structure learning.

Our result can be extended in several directions. For instance, the tree structure can be generalized to sparse structures where sparse learning methods such as glasso over the quantized data might be crucial. As another extension, one can study and solve a similar problem on discrete variables with sparse MRFs. Removing the central machine and allowing communication between local machines change the problem significantly and it worths of investigations.

APPENDIX A PROOF OF THEOREM 2

Since variables x and y have unit variances, we have $0 \leq D_1, D_2 \leq 1$. The relative correlation error can be upper bounded as follows

$$\begin{aligned} \text{err}_{\text{rel}} &= \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n u_i v_i \right| \right] \\ &= \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n y_i (x_i - u_i) + \frac{1}{n} \sum_{i=1}^n u_i (y_i - v_i) \right| \right] \\ &\leq \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n y_i (x_i - u_i) \right| \right] + \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n u_i (y_i - v_i) \right| \right]. \end{aligned}$$

Using the Cauchy-Schwartz inequality, we obtain

$$\begin{aligned} \text{err}_{\text{rel}} &\stackrel{(a)}{\leq} \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n (x_i - u_i)^2 \right)^{1/2} \right] + \\ &\quad \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n u_i^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n (y_i - v_i)^2 \right)^{1/2} \right] \\ &\stackrel{(b)}{\leq} \left(\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n y_i^2 \right] \right)^{1/2} \left(\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (x_i - u_i)^2 \right] \right)^{1/2} + \\ &\quad \left(\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right] \right)^{1/2} \left(\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (y_i - v_i)^2 \right] \right)^{1/2} \\ &\leq \sqrt{\mathbb{E}[y^2] D_1} + \sqrt{D_2 \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right]}. \end{aligned} \quad (44)$$

On the other hand, we have

$$\begin{aligned} D_1 &\geq \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (x_i - u_i)^2 \right] \\ &= 1 + \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right] - 2\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n x_i u_i \right] \\ &\geq 1 + \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right] - 2\mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n u_i^2 \right)^{1/2} \right] \\ &\geq 1 + \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right] - 2 \left(\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n x_i^2 \right] \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right] \right)^{1/2} \\ &= \left(\sqrt{\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right]} - 1 \right)^2. \end{aligned}$$

Hence,

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n u_i^2 \right] \leq (\sqrt{D_1} + 1)^2. \quad (45)$$

By substituting the above bound into (44), we obtain

$$\text{err}_{\text{rel}} \leq \sqrt{D_1} + \sqrt{D_2} + \sqrt{D_1 D_2}, \quad (46)$$

which completes the proof of Theorem 2.

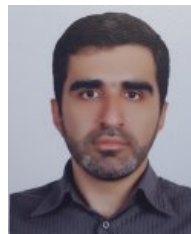
REFERENCES

- [1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [2] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1663–1707, 2010.
- [3] M. Tavassolipour, S. A. Motahari, and M.-T. M. Shalmani, "Learning of Gaussian processes in distributed and communication limited systems," *arXiv preprint arXiv:1705.02627*, 2017.
- [4] Z. Meng, D. Wei, A. Wiesel, and A. O. Hero, "Marginal likelihoods for distributed parameter estimation of gaussian graphical models," *IEEE Transactions on Signal Processing*, vol. 62, no. 20, pp. 5425–5438, 2014.
- [5] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *Journal of the American Statistical Association*, no. just-accepted, 2018.
- [6] T. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, 1998.
- [7] M. El Gamal and L. Lai, "On rate requirements for achieving the centralized performance in distributed estimation," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 2020–2032, 2017.
- [8] A. Irrthum, L. Wehenkel, P. Geurts *et al.*, "Inferring regulatory networks from expression data using tree-based methods," *PloS one*, vol. 5, no. 9, p. e12776, 2010.
- [9] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.
- [10] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, no. 5659, pp. 799–805, 2004.
- [11] S. Huang, J. Li, L. Sun, J. Ye, A. Fleisher, T. Wu, K. Chen, E. Reiman, A. D. N. Initiative *et al.*, "Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation," *NeuroImage*, vol. 50, no. 3, pp. 935–949, 2010.
- [12] 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.
- [13] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria, "A review on the computational approaches for gene regulatory network construction," *Computers in biology and medicine*, vol. 48, pp. 55–65, 2014.

- [14] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models a review," *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.
- [15] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 981–990.
- [16] S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, E. Reiman, A. D. N. Initiative *et al.*, "A sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1328–1342, 2013.
- [17] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
- [18] 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.
- [19] 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.
- [20] R. Mazumder and T. Hastie, "The graphical lasso: New insights and alternatives," *Electronic journal of statistics*, vol. 6, pp. 2125–2149, 2012.
- [21] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "Quic: quadratic approximation for sparse inverse covariance estimation," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2911–2947, 2014.
- [22] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu *et al.*, "High-dimensional covariance estimation by minimizing ℓ_1 -penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.
- [23] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, pp. 1436–1462, 2006.
- [24] S. Chen, D. M. Witten, and A. Shojaie, "Selection and estimation for mixed graphical models," *Biometrika*, vol. 102, no. 1, pp. 47–64, 2014.
- [25] 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.
- [26] —, "Learning Gaussian tree models: Analysis of error exponents and extremal structures," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2701–2714, 2010.
- [27] A. Xu and M. Raginsky, "Information-theoretic lower bounds on Bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, 2017.
- [28] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems*, 2013, pp. 2328–2336.
- [29] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang, "Optimality guarantees for distributed statistical estimation," *arXiv preprint arXiv:1405.0782*, 2014.
- [30] A. Garg, T. Ma, and H. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances in Neural Information Processing Systems*, 2014, pp. 2726–2734.
- [31] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, 2016, pp. 1011–1020.
- [32] Z.-Q. Luo, "Universal decentralized estimation in a bandwidth constrained sensor network," *IEEE Transactions on information theory*, vol. 51, no. 6, pp. 2210–2219, 2005.
- [33] M. Raginsky, "Learning from compressed observations," in *Information Theory Workshop, 2007. ITW'07. IEEE*. IEEE, 2007, pp. 420–425.
- [34] S.-i. Amari, "On optimal data compression in multiterminal statistical inference," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 5577–5587, 2011.
- [35] R. Ahlswede and M. Burnashev, "On minimax estimation in the presence of side information about remote data," *The Annals of Statistics*, pp. 141–171, 1990.
- [36] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.
- [37] R. C. Prim, "Shortest connection networks and some generalizations," *Bell Labs Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.
- [38] R. H. Bacon, "Approximations to multivariate normal orthant probabilities," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 191–198, 1963.
- [39] A. Dembo and O. Zeitouni, "Large deviations techniques and applications. corrected reprint of the second (1998) edition. stochastic modelling and applied probability, 38," 2010.
- [40] T. M. Cover and J. A. Thomas, *Elements of information theory 2nd edition*. John Wiley & Sons, 2006.
- [41] D. Huang, S. Yao, Y. Wang, and F. De La Torre, "Sequential max-margin event detectors," in *European conference on computer vision*. Springer, 2014, pp. 410–424.
- [42] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: A review," *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.
- [43] A. Damianou and N. Lawrence, "Deep Gaussian processes," in *Artificial Intelligence and Statistics*, 2013, pp. 207–215.
- [44] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "A non-parametric bayesian network prior of human pose," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1281–1288.



Mostafa Tavassolipour received the B.Sc. degree from Shahed University, Tehran, Iran, in 2009, and the M.Sc. degree from Computer Engineering department of Sharif University of Technology (SUT), Tehran, Iran, in 2011. Currently, He is a Ph.D. student of Artificial Intelligence program at Computer Engineering Department of Sharif University of Technology. His research interests include machine learning, image processing, information theory, content based video analysis, and bioinformatics.



Seyed Abolfazl Motahari is an assistant professor at Computer Engineering Department of Sharif University of Technology (SUT). He received his B.Sc. degree from the Iran University of Science and Technology (IUST), Tehran, in 1999, the M.Sc. degree from Sharif University of Technology, Tehran, in 2001, and the Ph.D. degree from University of Waterloo, Waterloo, Canada, in 2009, all in electrical engineering. From August 2000 to August 2001, he was a Research Scientist with the Advanced Communication Science Research Laboratory, Iran Telecommunication Research Center (ITRC), Tehran. From October 2009 to September 2010, he was a Postdoctoral Fellow with the University of Waterloo, Waterloo. From September 2010 to July 2013, he was a Postdoctoral Fellow with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley. His research interests include multiuser information theory and Bioinformatics. He received several awards including Natural Science and Engineering Research Council of Canada (NSERC) Post-Doctoral Fellowship.



Mohammad Taghi Manzuri Shalmani received the B.Sc. and M.Sc. in electrical engineering from Sharif University of Technology (SUT), Iran, in 1984 and 1988, respectively. He received the Ph.D. degree in electrical and computer engineering from the Vienna University of Technology, Austria, in 1995. Currently, he is an associate professor in the Computer Engineering Department, Sharif University of Technology, Tehran, Iran. His main research interests include digital signal processing, stochastic modeling,

and Multi-resolution signal processing.