# Structure Learning of Sparse GGMs over Multiple Access Networks

Mostafa Tavassolipour, Armin Karamzade, Reza Mirzaeifard, Seyed Abolfazl Motahari,
and Mohammad-Taghi Manzuri Shalmani

*Abstract*—A central machine is interested in estimating the underlying structure of a sparse Gaussian Graphical Model (GGM) from datasets distributed across multiple local machines. The local machines can communicate with the central machine through a wireless multiple access channel. In this paper, we are interested in designing effective strategies where reliable learning is feasible under power and bandwidth limitations. Two approaches are proposed: Signs and Uncoded methods. In Signs method, the local machines quantize their data into binary vectors and an optimal channel coding scheme is used to reliably send the vectors to the central machine where the structure is learned from the received data. In Uncoded method, data symbols are scaled and transmitted through the channel. The central machine uses the received noisy symbols to recover the structure. Theoretical results show that both methods can recover the structure with high probability for large enough sample size. Experimental results indicate the superiority of Signs method over Uncoded method under several circumstances.

*Index Terms*—Structure learning, Gaussian graphical model, distributed learning

## I. INTRODUCTION

IN recent years, by the explosion of the volume of training data, distributed machine learning has become more important than ever. Many modern big datasets are distributed over several hosting machines which are connected to each other via some communication links. In such systems, designing distributed learning algorithms which efficiently exploit the available resource demands a careful design where intensive computational workloads and the amount of data communications are taken into account.

This paper is focused on the problem of structure learning in Gaussian Graphical Models (GGMs) in distributed environments. A GGM for a $d$-dimensional random vector $\mathbf{x} = (x_1, \cdots, x_d)^T \in \mathbb{R}^d$ is specified by a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \cdots, d\}$ is the set of vertices and $\mathcal{E} \subseteq \mathcal{V}^2$ is the set of edges. The model comprises all $d$-dimensional normal distributions $\mathcal{N}(\mu, \Theta^{-1})$ where $\Theta$ is the precision matrix with $\Theta_{jk} \neq 0$ iff $(j, k) \in \mathcal{E}$. It worths mentioning that GGM is indeed a Markov Random Field (MRF).

In our problem setting, we assume that the data are distributed over multiple local machines so that each one contains a single dimension of the whole dataset. The local machines are connected to a central machine via a wireless medium. The central machine is responsible for inferring the conditional dependencies between the gathered data by the local machines. The system's block diagram is depicted in Figure 1. We also assume that the communication links between the local and the central machine are bandwidth limited implying that

transmission of the whole local datasets to the central machine is impossible. Due to this constraint, each local machine transmits some information from its local dataset to the central machine. Then, the central machine estimates the underlying graph structure using received information from the local machines.

In this paper, we consider a wireless multiple access channel between the local machines and the central node. Each local machine is equipped with a single antenna while the central node is equipped with multiple antennas. Hence, the overall channel between the local and central machines is modeled as a single-input multiple-output multiple access channel (SIMO-MAC).

We have proposed two communication schemes for transmitting information from the local machines to the central machine. In the first scheme, we have separate the source coding from the channel coding. In this scheme, we quantize the source samples into single bits and assume that there exists a channel coding such that the bits can be sent through the channel reliably. We refer to this scheme as *Signs method* in the paper.

In the second scheme, we do not use any source and channel coding. In this scheme, we put the source samples into the channel without any encoding. At the central machine, we directly estimate the underlying graph structure using received data from the channel. We refer to this scheme as *Uncoded method*.

We have shown through theoretical analysis and experiments that by transmitting only 1 bit per sample; the central machine can reliably recover the underlying graph structure. More precisely, we have shown theoretically that by consuming only 1 bit per sample, under some mild conditions, the central machine can perfectly recover the graph structure with high probability. Moreover, the true signs of the edges weights of the graph are obtained.

The paper is organized as follows. In Section II, we provide a brief review on structure learning of GGMs. Section III describes the detail of our modeling for the source and the communication channel. In sections IV and V, we describe Signs and Uncoded methods, respectively. Section VI provides experimental results to compare and evaluate the proposed methods. Finally, Section VII concludes the paper.

## II. RELATED WORK

The problem of structure learning of GGMs from data samples has applications in many fields including biology

Figure 1. The whole system block diagram.

and social networks. For example, it has been used for gene regulatory networks reconstruction in [1], [2] and analysis of users relationships in social networks in [3]. There are many studies addressing this problem from various perspectives [4], [5], [6].

The Chow-Liu algorithm obtains the maximum likelihood estimate of the structure if the underlying graph is a tree [7]. Although this algorithm is applicable for discrete random variables, it can be used for tree structured GGMs in a similar manner [8]. Tavassolipour et al. [9] proposed a distributed version of the Chow-Liu algorithm and proved it can recover the underlying tree structure with high probability. Tan et al. in [10] and [11] provided an analysis of the error exponent of the Chow-Liu algorithm on tree-structured GGMs.

GGMs have the property that the neighbors of each variable can be obtained by solving a linear regression problem for the corresponding variable on other variables. This approach is referred to as *neighborhood selection* in the literature. For the sparse structures, there are some methods which penalize the linear regression problem with $\ell_1$ of the coefficient vector [12], [13].

Among the proposed methods for sparse structure estimation of GGMs, the $\ell_1$-regularized maximum likelihood approaches are more popular [6], [12], [14]. This class of model is analyzed in several articles (see [4] and refs therein). For instance, Ravikumar et al. [4] analyzed the performance of the $\ell_1$-regularized maximum likelihood estimator (MLE) under high dimensional scaling. They showed that with probability converging to one, the estimated structure correctly specifies the zero pattern of the true precision matrix. A similar study is conducted by [15] which analyzed the consistency of the $\ell_1$-regularized MLE in the Frobenius norm.

Besides the lasso typed estimators, thresholding based estimators are proposed for sparse recovering of the precision matrix [5], [8]. For example, Sojoudi [5] proposed a simple thresholding method and showed, under certain conditions, the resulting structure is identical to the structure obtained by the lasso.

In the distributed setting, there are several studies which address the problem of covariance/precision matrix estimation [16], [17], [18]. Arroyo and Hou [18] studied the problem of sparse precision matrix estimation in the situation where the samples are distributed among several machines. Their work

differs from our setting in the sense that we assume the data are split across dimensions whereas they split the data across samples. Meng et al. [16] addressed estimation of the precision matrix in a distributed manner where the zero pattern of the precision matrix is known in advanced.

## III. PROBLEM FORMULATION

We are given $n$ i.i.d. random vectors drawn from a $d$-dimensional zero mean normal distribution $\mathcal{N}(\mathbf{0}, Q_x)$ with $(Q_x)_{jj} = 1$. The focus of this paper is the problem of estimating the zero pattern of the sparse precision matrix $\Theta_x = Q_x^{-1}$ in a situation where the data is stored in $d$ separate local machines such that each machine possesses one dimension of the sample vectors.

Denoting the whole gathered data by $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}\}$, the $j$th machine captures the $j$-th dimension of the sample vectors. We denote the $j$-th dimension of the $i$-th sample by $x_j^{(i)}$, i.e. $\mathbf{x}^{(i)} = [x_1^{(i)}, \cdots, x_d^{(i)}]^T$. Hence, the local data at the $j$-th local machine is $\{x_j^{(1)}, \cdots, x_j^{(n)}\}$.

The probability density function of the normal distribution is given by

$$f(\mathbf{x}; \Theta) = \frac{1}{\sqrt{\det(2\pi\Theta^{-1})}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\Theta\mathbf{x}\right\}. \tag{1}$$

The negative log-likelihood of $n$ i.i.d. samples is given by

$$g(\Theta) = \text{tr}(\Theta S_x) - \log\det(\Theta), \tag{2}$$

where $S_x = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^{(i)}\left(\mathbf{x}^{(i)}\right)^T$ is reffered to as sample covariance matrix. One of the well known methods for estimating the sparse precision matrix is to solve an $\ell_1$-regularized maximum likelihood function stated as

$$\widehat{\Theta} = \arg\min_{\Theta > 0} \left\{g(\Theta) + \lambda_n \|\Theta\|_{1,\text{off}}\right\}, \tag{3}$$

where $\|\Theta\|_{1,\text{off}} := \sum_{j \neq k} |\Theta_{jk}|$, and $\lambda_n$ is a user defined regularization parameter. There is an efficient algorithm known as *glasso* [19] for solving the above log-determinant program. In [4], it is stated that the Hessian matrix of (2) is given by

$$\Gamma = \nabla_\Theta^2 g(\Theta)\Big|_{\Theta=\Theta_x} = \Theta_x^{-1} \otimes \Theta_x^{-1}. \tag{4}$$

where $\otimes$ is the Kronecker matrix product. The entry $\Gamma_{(j,k),(l,m)}$ corresponds to the second derivative $\frac{\partial^2 g}{\partial\Theta_{jk}\partial\Theta_{lm}}$, evaluated at $\Theta = \Theta_x$ and $\Gamma_{(j,k),(l,m)} = \text{cov}\{x_j x_k, x_l x_m\}$.

We define the set of non-zero entries in the true precision matrix $\Theta_x$ as

$$S(\Theta_x) = \{(j,k) \in \mathcal{V}^2 | (\Theta_x)_{jk} \neq 0\}. \tag{5}$$

Recall that $\mathcal{V} = \{1, \cdots, d\}$. Note that $S(\Theta_x)$ includes the diagonal entries. Let $S^c(\Theta_x)$ be the complement set of $S(\Theta_x)$ which includes all pairs $(j,k)$ where $(\Theta_x)_{jk} = 0$. For any two subsets $T$ and $T'$ of $V \times V$, the notation $\Gamma_{TT'}$ denotes a $|T| \times |T'|$ sub-matrix of $\Gamma$ with rows and columns indexed by $T$ and $T'$, respectively.

We adopt the incoherence condition used by Ravikumar et al. [4] to obtain error bounds on consistency of the solutions. We define max-row-sum norm of a $d$ by $d$ matrix $X$ as

$$\|\|X\|\|_\infty = \max_{i=1,\cdots,d} \sum_{j=1}^{d} |X_{ij}|. \tag{6}$$

**Assumption 1.** *(Incoherence Condition [4])*
*There exists some $\alpha \in (0,1]$ such that*

$$\|\|\Gamma_{S^c S}(\Gamma_{SS})^{-1}\|\|_\infty \leq (1 - \alpha). \tag{7}$$

An implication of the above condition is that the non-edge pairs cannot have strong influence on the edges.

**Assumption 2.** *(Covariance Control [4])*
*There exist constants $\kappa_\Sigma, \kappa_\Gamma < \infty$ such that*

$$\|\|Q_x\|\|_\infty \leq \kappa_\Sigma, \tag{8}$$

$$\|\|\Gamma_{SS}^{-1}\|\|_\infty \leq \kappa_\Gamma. \tag{9}$$

The above assumptions imply that the covariance elements along any row of $Q_x$ and $\Gamma_{SS}^{-1}$ have bounded $\ell_1$ norms.

Having no access to the original data, the central machine finds the solution of the optimization problem (3) using received data from the local machines. The likelihood function $g(\Theta)$ in (2) depends on the samples via the sample covariance matrix $S_x$. Thus, obtaining an appropriate approximate of the sample covariance matrix at the central machine would result in a good estimate of the underlying structure.

### A. Channel Model

There is a wireless channel between the local machines and the central node. In this paper, we assume that the channel can be modeled as a single-input multiple-output Gaussian multiple access channel (SIMO-GMAC) with additive white noise. The number of antennas at the receiver is assumed to be $m$. All transmitters have equal transmit power which is denoted by $p$. This implies, for all $j$, the following constraint on the transmit symbols should be satisfied:

$$\frac{1}{n} \sum_{i=1}^{n} \left| s_j^{(i)} \right|^2 \leq p, \tag{10}$$

where $s_j^{(i)}$'s are the channel inputs at the local machine $j$. Denoting the transmit symbols by vector $\mathbf{s}$, the channel output is modeled by

$$\mathbf{y} = H\mathbf{s} + \mathbf{z}, \tag{11}$$

where $H \in C^{m \times d}$ is assumed to be an invertible complex matrix. In the fading environments, the channel gains are drawn from independent circularly symmetric complex Gaussian distribution. The additive noise $\mathbf{z}$ is an independent circularly symmetric Gaussian vector with covaraience matrix $\sigma_z^2 I_d$.

## IV. SIGNS METHOD

We assume that each local machine applies a sign function on its local dataset to obtain binary data. More precisely, given samples $\{x_j^{(1)}, \cdots, x_j^{(n)}\}$ at local machine $j$, it obtains the signs dataset $\{\hat{x}_j^{(1)}, \cdots, \hat{x}_j^{(n)}\}$ where $\hat{x}_j = \text{sign}(x_j)$. Then, it transmits the binary data to the central machine with the rate of 1 bit per sample using a channel encoder and decoder.

Denoting the bit rate of the channel at local machine $j$ by $R_j$, the achievable bit rates for all machines are characterized by [20]

$$\sum_{k \in S} R_k \leq \lg \det \left( \frac{p}{\sigma_z^2} H_S^H H_S + I_{|S|} \right), \qquad \forall\ S \subseteq \{1, \cdots, d\}, \tag{12}$$

where $H_S$ is the sub-matrix of $H$ that includes the rows and columns indexed by $S$, $H^H$ is the Hermitian of $H$, $I_{|S|}$ is the $|S| \times |S|$ identity matrix, and $\lg(\cdot)$ is the logarithm function in base 2. Throughout this section, we assume that $R_j \geq 1$, for all $j = 1, \cdots, d$. Therefore, there exists a channel encoding to transmit 1 bit per sample at each local machine.

At the central machine, our goal is to solve the optimization problem (3) on the received binary data from all local machines. In order to obtain a solution that is close to the solution obtained by the original data, we seek a suitable approximation for the sample covariance matrix $S_x$ in (3). Thus, our goal is to estimate the sample covariance matrix as accurate as possible using the received signs data. In this section, we propose an estimator for the sample covariance matrix and theoretically show its error decreases exponentially by the sample size $n$.

Let $x_j$ and $x_k$ be jointly normal with zero means, unit variances and correlation coefficient $\rho_{jk}$. If $\hat{x}_j$ and $\hat{x}_k$ be the corresponding sign variables, then the joint probability mass function (pmf) of $\hat{x}_j$ and $\hat{x}_k$ is given by [21]

$$\begin{array}{c|cc}
\hat{x}_j \backslash \hat{x}_k & -1 & +1 \\
\hline
-1 & \beta_{jk}/2 & (1-\beta_{jk})/2 \\
+1 & (1-\beta_{jk})/2 & \beta_{jk}/2
\end{array} \tag{13}$$

where $\beta_{jk} \in [0,1]$ and given by

$$\beta_{jk} = \frac{1}{2} + \frac{\arcsin(\rho_{jk})}{\pi}. \tag{14}$$

The equation (14) can be rewritten as

$$\rho_{jk} = \sin(\pi(\beta_{jk} - \frac{1}{2})) = -\cos(\pi\beta_{jk}). \tag{15}$$

Thus, by proposing an estimator for $\beta_{jk}$, using (15) we can obtain an estimator for $\rho_{jk}$. The following estimator for $\beta_{jk}$ is optimal in the sense that it is unbiased and has minimum variance (UMVE) [22],

$$\hat{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{x}_j^{(i)} \hat{x}_k^{(i)} = 1), \tag{16}$$

where $\mathbb{I}(.)$ is the indicator function. $\hat{\beta}_{jk}$ is indeed a binomial random variable with success probability of $\beta_{jk}$. By substituting $\hat{\beta}_{jk}$ in (15), we use

$$\hat{\rho}_{jk} = -\cos(\pi\hat{\beta}_{jk}), \qquad (17)$$

as an estimator for $\rho_{jk}$. Although $\hat{\rho}_{jk}$ is indeed biased, it is a consistent estimator for $\rho_{jk}$. Following lemma gives an error bound on this estimator.

**Lemma 1.** *Let $x_j$ and $x_k$ be two jointly normal variables with zero means, unit variances and correlation coefficient $\rho_{jk}$. Then, for the estimator (17), we have*

$$\Pr\left(\left|\hat{\rho}_{jk} - \rho_{jk}\right| \geq \delta\right) \leq 2\exp\left(-\frac{2}{\pi^2}n\delta^2\right), \qquad (18)$$

*where $\delta \geq 0$.*

*Proof.* Since the function $\cos(\cdot)$ is a 1-Lipschitz function, i.e. $|\cos(x) - \cos(y)| \leq |x - y|$, we have

$$\Pr\left(\left|\hat{\rho}_{jk} - \rho_{jk}\right| \geq \delta\right) = \Pr\left(\left| -\cos(\pi\hat{\beta}_{jk}) + \cos(\pi\beta_{jk})\right| \geq \delta\right)$$
$$\leq \Pr\left(\pi|\hat{\beta}_{jk} - \beta_{jk}| \geq \delta\right)$$
$$= \Pr\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \frac{\delta}{\pi}\right).$$

Since $\hat{\beta}_{jk}$ is sum of $n$ independent Bernoulli random variables, applying the Hoeffding inequality yields

$$\Pr\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \frac{\delta}{\pi}\right) \leq 2\exp\left(-\frac{2}{\pi^2}n\delta^2\right),$$

which completes the proof. $\square$

Lemma 1 shows that the error of proposed estimator $\hat{\rho}_{jk}$ is controlled by the number of samples exponentially. Using this estimator, we can obtain an estimator for the sample covariance matrix as follows

$$\widehat{S}_x = -\cos(\frac{\pi}{2}B), \qquad (19)$$

where $B_{jk} = \hat{\beta}_{jk}$, and $\cos(.)$ function is applied on the input matrix element wise, i.e. $(\widehat{S}_x)_{ij} = -\cos(\frac{\pi}{2}\beta_{ij})$. By substituting the above sample covariance matrix into (3), we can solve the regularized maximum likelihood problem.

Note that the sample covariance matrix $\widehat{S}_x$ defined in (19) is not necessarily positive semi-definite. But, it does not affect the convexity and uniqueness solution of (3). Ravikumar et al. [4] proved that the problem (3) is convex and has a unique solution for any sample covariance matrix with strictly positive diagonal elements which holds for $\widehat{S}_x$ in (19).

By incorporating Lemma 1 and the theorems 1 and 2 in [4], we can conclude that the precision matrix obtained by solving (3) with the sample covariance matrix $\widehat{S}_x$ in (28), recovers the true structure with high probability. Moreover, the proposed method correctly recovers the signs of the edges with high probability.

More precisely, the event $\mathcal{M}(\Theta_x; \widehat{\Theta}_x)$ indicates that $\Theta_x$ and $\widehat{\Theta}_x$ do agree on the zero entries and for the nonzero entries they have the same sign. Theorems 1 and 2 state that the event $\mathcal{M}(\Theta_x, \widehat{\Theta}_x)$ occurs with high probability. Before stating the theorems we should define some properties of the underlying GGM. We denote the maximum degree of the underlying graph structure by $\Delta$. The minimum absolute value of the edges weighs in the precision matrix is denoted by $\theta_{\min}$ which is

$$\theta_{\min} = \min_{(i,j) \in E(\Theta_x)} |(\Theta_x)_{ij}|. \qquad (20)$$

**Theorem 1.** *Consider a normal distribution satisfying the incoherence Assumption 1 and 2 with parameter $\alpha \in (0, 1]$. Let $\widehat{\Theta}_x$ be the solution of the log-determinant program (3) with sample covariance $\widehat{S}_x$ in (19) and regularization parameter $\lambda_n = (8\pi/\alpha)\sqrt{\frac{1}{2n}\ln\frac{2}{\epsilon}}$ for some $0 < \epsilon \leq d^{-2}$. Then,*

*(a) If the sample size is lower bounded as*

$$n > C_{\text{sign}}^2 \Delta^2 \left(1 + \frac{8}{\alpha}\right)^2 \ln\frac{2}{\epsilon}, \qquad (21)$$

*where*

$$C_{\text{sign}} = 3\sqrt{2}\pi \max\{\kappa_\Sigma\kappa_\Gamma, \kappa_\Sigma^3\kappa_\Gamma^2\},$$

*then with probability at least $1 - d^2\epsilon$, the edge set specified by $\widehat{\Theta}_x$ is a subset of the true edge set.*

*(b) If the sample size satisfies the lower bound*

$$n > T_{\text{sign}}^2 \left(1 + \frac{8}{\alpha}\right)^2 \ln\frac{2}{\epsilon}, \qquad (22)$$

*where*

$$T_{\text{sign}} = \sqrt{2}\pi \max\{\kappa_\Gamma\theta_{\min}^{-1}, 3\Delta \max\{\kappa_\Sigma\kappa_\Gamma, \kappa_\Sigma^3\kappa_\Gamma^2\}\},$$

*then,*

$$\Pr\left(\mathcal{M}(\widehat{\Theta}_x; \Theta_x)\right) \geq 1 - d^2\epsilon. \qquad (23)$$

**Remark 1.** *Note that the proposed sign method is applicable for any channel with capacity greater than or equal to 1 bit.*

## V. UNCODED METHOD

In this section, we assume that each local machine puts its local data into the channel without any source or channel coding. The central machine estimates the underlying graph structure using received data from the channel. At the central machine, no source or channel decoding is used. It infers the structure directly from the output samples of the channel.

As described in Section III-A, we consider a SIMO-GMAC. Each local machine can transmit two consequent samples by each channel use. More precisely, denoting the channel input symbol by $\mathbf{s} = \mathbf{s}_R + j\mathbf{s}_I$, each local machine can put two consequent samples as real and imaginary parts of the input symbol. Therefore, at the central machine, $n/2$ vectors are received that each one is $2d$-dimensional.

In this way, the equation (11) can be decomposed as

$$\mathbf{y} = (H_R + jH_I)(\mathbf{s}_R + j\mathbf{s}_I) + (\mathbf{z}_R + j\mathbf{z}_I), \qquad (24)$$

where $j^2 = -1$. Hence, it can be rewritten in a block-matrix form as

$$\begin{bmatrix} \mathbf{y}_R \\ \mathbf{y}_I \end{bmatrix} = \begin{bmatrix} H_R & -H_I \\ H_I & H_R \end{bmatrix} \begin{bmatrix} \mathbf{s}_R \\ \mathbf{s}_I \end{bmatrix} + \begin{bmatrix} \mathbf{z}_R \\ \mathbf{z}_I \end{bmatrix}, \qquad (25)$$

where $H_R, H_I$ are $d \times d$ real matrices, and all the real and imaginary part vectors are $d$-dimensional. In this way, two

source samples can be transmitted per channel use: a sample is put into the real part and the other is put into the imaginary part. In particular, $\mathbf{s}_R + j\mathbf{s}_I = \sqrt{\frac{p}{2}}(\mathbf{x}_R + j\mathbf{x}_I)$, where $\mathbf{x}_R$ and $\mathbf{x}_I$ are two independent samples from the source. In this way, the transmit power constraints are satisfied.

The central machine estimates the conditional dependencies of the vectors $\mathbf{x}$ using the received vectors $\mathbf{y}$.

### A. Approximating the Sample Covariance

Defining matrix $\widetilde{H} = \sqrt{\frac{p}{2}} \begin{bmatrix} H_R & -H_I \\ H_I & H_R \end{bmatrix}$, $\tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x}_R \\ \mathbf{x}_I \end{bmatrix}$, and $\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_R \\ \mathbf{y}_I \end{bmatrix}$. When transmitting samples through the channel, if we put two independent samples as real and imaginary parts of the vector $\tilde{\mathbf{x}}$, then the covariance matrix of $\tilde{\mathbf{x}}$ is

$$Q_{\tilde{x}} = \begin{bmatrix} Q_x & 0 \\ 0 & Q_x \end{bmatrix}. \quad (26)$$

On the other hand, according to equation (25), we have

$$Q_{\tilde{x}} = \widetilde{H}^{-1} Q_{\tilde{y}} \widetilde{H}^{-T} - \sigma_z^2 \widetilde{H}^{-1} \widetilde{H}^{-T}, \quad (27)$$

where $Q_{\tilde{x}}$ and $Q_{\tilde{y}}$ is the covariance matrix of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$, respectively. By substituting the sample covariance matrix of $\tilde{\mathbf{y}}$ into the above expression, we obtain an approximation for the sample covariance matrix of $\tilde{\mathbf{x}}$, as

$$S_{\tilde{x}} = \widetilde{H}^{-1} S_{\tilde{y}} \widetilde{H}^{-T} - \sigma_z^2 \widetilde{H}^{-1} \widetilde{H}^{-T}. \quad (28)$$

**Lemma 2.** *Given $n$ i.i.d. samples of the vector $\tilde{\mathbf{y}}$. Then, for the sample covariance matrix $S_{\tilde{x}}$ in (28) we have*

$$\Pr\left(|(S_{\tilde{x}})_{jk} - (Q_{\tilde{x}})_{jk}| \geq \delta\right) \leq 4 \exp\left(\frac{-n\delta^2}{c}\right), \quad (29)$$

*where*

$$c = 3200 \left(1 + \sigma_z^2 / \lambda_{\min}^2(\widetilde{H})\right)^2, \quad (30)$$

*and $\lambda_{\min}(\widetilde{H})$ is the minimum eigenvalue of $\widetilde{H}$.*

*Proof.* We define the random variable $\mathbf{w}$ as follows

$$\mathbf{w} = \widetilde{H}^{-1} \mathbf{y} = \mathbf{x} + \widetilde{H}^{-1} \mathbf{z}.$$

It is clear that $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, Q_w)$, where $Q_w = Q_x + \sigma_z^2 \widetilde{H}^{-1} \widetilde{H}^{-T}$. Denoting $\mathbf{w} = [w_1, \cdots, w_p]^T$, $w_j / \sqrt{(Q_w)_{jj}}$ is a standard normal variable which is sub-Gaussian with parameter 1. Thus, according to Lemma 1 in [4], we have

$$\Pr\left(|(S_w)_{jk} - (Q_w)_{jk}| \geq \delta\right) \leq \exp\left\{-\frac{n\delta^2}{3200 \max_i(Q_w)_{ii}^2}\right\},$$

where $S_w = \widetilde{H}^{-1} S_y \widetilde{H}^{-T}$ is the sample covariance over $\mathbf{w}^{(i)} = \widetilde{H}^{-1} \mathbf{y}^{(i)}$ samples. On the other hand, we have

$$\Pr\left(|(S_{\tilde{x}} - Q_{\tilde{x}})_{jk}| \geq \delta\right) = \Pr\left(\left|\left(\widetilde{H}^{-1}\left(S_{\tilde{y}} - Q_{\tilde{y}}\right)\widetilde{H}^{-T}\right)_{jk}\right| \geq \delta\right)$$

$$= \Pr\left(|(S_w - Q_w)_{jk}| \geq \delta\right)$$

$$\leq \exp\left\{-\frac{n\delta^2}{3200 \max_i(Q_w)_{ii}^2}\right\} \quad (31)$$

By finding an upper bound on $\max_i(Q_w)_{ii}^2$, we can obtain an upper bound on the above probability.

$$\max_i (Q_w)_{ii} = \max_i (Q_x + \sigma_z^2 \widetilde{H}^{-1} \widetilde{H}^{-T})_{ii}$$

$$\leq \max_i \left(|Q_x| + \left|\sigma_z^2 \widetilde{H}^{-1} \widetilde{H}^{-T}\right|\right)_{ii}$$

$$\leq \max_i (Q_x)_{ii} + \max_j (\sigma_z^2 \widetilde{H}^{-1} \widetilde{H}^{-T})_{jj}. \quad (32)$$

By decomposing $\widetilde{H}^{-1} \widetilde{H}^{-T}$ as $U\Lambda^2 U^T$, we have

$$\max_i (\sigma_z^2 \widetilde{H}^{-1} \widetilde{H}^{-T})_{ii} = \max_i(\sigma_z^2 U\Lambda^2 U^T)_{ii}$$

$$\leq \max_i(\sigma_z^2 \lambda_{\max}^2(\widetilde{H}^{-1})UU^T)_{ii}$$

$$= \frac{\sigma_z^2}{\lambda_{\min}^2(\widetilde{H})}.$$

By combining the above bound and (32), and substituting into (31) we obtain the claimed bound in the lemma. □

Next, we define the approximate sample covariance matrix of $\mathbf{x}$ as

$$\hat{S}_x = \frac{1}{2}\left(\begin{bmatrix} I_d & 0_d \end{bmatrix} S_{\tilde{x}} \begin{bmatrix} I_d \\ 0_d \end{bmatrix} + \begin{bmatrix} 0_d & I_d \end{bmatrix} S_{\tilde{x}} \begin{bmatrix} 0_d \\ I_d \end{bmatrix}\right), \quad (33)$$

where $0_d$ is a $d \times d$ zero matrix.

**Lemma 3.** *For $\hat{S}_x$ defined in (33), we have*

$$\Pr\left(|(S_x)_{jk} - (Q_x)_{jk}| \geq \delta\right) \leq 8 \exp\left(\frac{-n\delta^2}{2c}\right). \quad (34)$$

*Proof.* From (33), it is clear that

$$(\hat{S}_x)_{jk} = \frac{1}{2}\left((S_{\tilde{x}})_{jk} + (S_{\tilde{x}})_{(d+j)(d+k)}\right).$$

Thus, we have

$$\Pr\left(|(\hat{S}_x)_{jk} - (Q_x)_{jk}| \geq \delta\right)$$

$$= \Pr\left(|(S_{\tilde{x}})_{jk} + (S_{\tilde{x}})_{(d+j)(d+k)} - 2(Q_{\tilde{x}})_{jk}| \geq 2\delta\right)$$

$$\leq \Pr\left(|(S_{\tilde{x}})_{jk} - (Q_{\tilde{x}})_{jk}| + |(S_{\tilde{x}})_{(d+j)(d+k)} - (Q_{\tilde{x}})_{jk}| \geq 2\delta\right)$$

$$\leq \Pr\left(|(S_{\tilde{x}})_{jk} - (Q_{\tilde{x}})_{jk}| \geq \delta\right) +$$

$$\quad \Pr\left(|(S_{\tilde{x}})_{(d+j)(d+k)} - (Q_{\tilde{x}})_{jk}| \geq \delta\right)$$

$$\leq 8 \exp\left(\frac{-n\delta^2}{2c}\right),$$

where the last inequality is obtained by the error bound of Lemma 2 for the sample size $n/2$. □

Note that the matrix $\widehat{S}_x$ in (33) is not necessarily positive semi-definite. But this does not affect the convexity of the optimization problem (3).

By substituting $\widehat{S}_x$ from (33) into the (3), we can solve the $\ell_1$ regularized maximum likelihood problem and obtain a sparse solution for the precision matrix $\Theta_x$. Similar to Theorem 1, we can guarantee that Uncoded method can correctly recover the underlying graph structure with high probability.

**Theorem 2.** *Consider a normal distribution satisfying the incoherence Assumption 1 and 2 with parameter $\alpha \in (0, 1]$. Let $\widehat{\Theta}_x$ be the solution of the log-determinant program (3) with*

sample covariance $\widehat{S}_x$ in (33) and regularization parameter $\lambda_n = (8\pi/\alpha)\sqrt{\frac{1}{2n}\ln\frac{2}{\epsilon}}$ for some $0 < \epsilon \le d^{-2}$.

(a) If the sample size is lower bounded as

$$n > C_{\text{uncoded}}^2 \, \Delta^2 \left(1 + \frac{8}{\alpha}\right)^2 \ln\frac{8}{\epsilon}, \qquad (35)$$

where

$$C_{\text{uncoded}} = 6\sqrt{2c} \; \max\{\kappa_\Sigma\kappa_\Gamma, \kappa_\Sigma^3\kappa_\Gamma^2\}.$$

then with probability at least $1 - d^2\epsilon$, the edge set specified by $\widehat{\Theta}_x$ is a subset of the true edge set.

(b) If the sample size satisfies the lower bound

$$n > T_{\text{uncoded}}^2 \left(1 + \frac{8}{\alpha}\right)^2 \ln\frac{8}{\epsilon}, \qquad (36)$$

where

$$T_{\text{uncoded}} = 2\sqrt{2c}\max\{\kappa_\Gamma\theta_{\min}^{-1}, 3\Delta \; \max\{\kappa_\Sigma\kappa_\Gamma, \kappa_\Sigma^3\kappa_\Gamma^2\}\}$$

then,

$$\Pr\left(\mathcal{M}(\widehat{\Theta}_x; \Theta_x)\right) \ge 1 - d^2\epsilon. \qquad (37)$$

**Remark 2.** *Since the channel has real and imaginary parts, 2 samples can be transmitted by each channel access (i.e. each local machine can transmit n samples by n/2 channel uses). Therefore, if the sample generation rate at the source is less than or equal to twice of the channel's rate, the machines can transmit all samples without any sample loss.*

## VI. Experiments

In this section, the performance of our proposed methods is evaluated by performing several experiments. In our simulations[1], the *glasso* package [19] is used to solve $\ell_1$-regularized MLE of the precision matrix. This package is based on the block coordinate descent algorithm proposed by [6].

In our experiments, a sparse random precision matrix is generated as follows. First, we generate a random sparse graph with a fixed probability of the edge presence, say 0.1, and also set its maximum node degree to $\Delta = 5$. Then, we choose edge weights uniformly in $[-1, 1]$ for the symmetric precision matrix $\Theta$. Next, we make it positive definite matrix by adding a scaled identity matrix. Finally, we normalize the precision matrix to set the variances to 1. Also, we ensure that the generated matrix satisfies Assumption 1.

We employ the True Positive and False Positive Rates (TPR and FPR, respectively) as our performance measures. TPR is defined as the percentage of the predicted edges (non-zero off-diagonal entries in the precision matrix) that correctly detected. Similarly, FPR is the percentage of the predicted non-edges (zero entries in the precision matrix) that incorrectly detected.

We have experimentally observed that $\lambda_n^{*\text{signs}} \approx 4\lambda_n^{*\text{original}}$ and $\lambda_n^{*\text{uncoded}} \approx \frac{2}{3}\lambda_n^{*\text{original}}$, where $\lambda_n^{*\text{signs}}$, $\lambda_n^{*\text{uncoded}}$, and $\lambda_n^{*\text{original}}$ are the best regularization parameter for the signs, uncoded, and original data, respectively.

In order to have a fair comparison between Signs and Uncoded methods, we have used identical parameters for the

[1]The source code is available at https://github.com/ArminKaramzade/distributed-sparse-GGM.

channel. More precisely, we assume $H = I$ which ensures all the local machines have identical bit rates. We set the bit rate of each local machine, i.e. $R_j$, to 2 bits. Thus, in the proposed methods, each local machine can transmit $n$ bits by $n/2$ channel uses. According to (12), in order to achieve the bit-rate of 2, the signal to noise ratio (SNR) should set to 3 (i.e. $\frac{p}{\sigma_z^2} = 3$).

In the first experiment, we evaluate the performance of our methods with respect to the dimension $d$. Figure 2 shows TPR and FPR as a function of $d$ for sample sizes $n = 1000$ and $10,000$. In this experiment, the error curves are averaged over 20 different random graphs and for each graph the TPR and FPR are averaged over 10 different random samples. As can be seen from Figure 2, Signs method outperforms Uncoded method. However, all three methods have approximately the same FPR.

Figure 3 reflects the performance of the methods as a function of the sample size $n$ for $d = 50$ and $d = 100$. As can be seen, by increasing the sample size $n$, the performance of all methods increases. In this experiment again Signs method outperforms the uncoded scheme.

In Figure 4, the probability of perfect structure recovery for a star-shaped graph is depicted. In this experiment, the underlying star graph consists of $d = 70$ nodes. The precision matrix is generated as the inverse of a covariance matrix with $(Q_x)_{ij} = \frac{1}{4}$ for all $(i, j) \in \mathcal{E}$ which satisfies Assumption 1. The probability of perfect recovery is estimated by running the proposed methods 100 times and counting the number of times that the structure is recovered exactly. As can be seen from the figure, all methods recover the structure exactly for large enough sample sizes as claimed by theorems 1 and 2.

In Figure 5, we measure the TPR of Uncoded method for different values of the SNR. The experiment is performed on a random graph with $d = 40$ nodes with maximum degree of $\Delta = 5$ and $n = 10,000$. In this experiment, we have generated the channel matrix $H$ with entries drawn from i.i.d. standard normal samples. The TPR curve is averaged over 100 different channel matrices. As can be seen from Figure 5, for SNR greater than 5, the performance of Uncoded method is very close to the TPR of the original data.

The FPR curve is not plotted, since the error values were negligible even for small SNR.

## VII. Conclusion

In this paper, we have studied the sparse structure learning of GGMs where the data are distributed across multiple local machines. Two methods are proposed to send information from the local machines to the central machine, namely, Signs and Uncoded methods. We have analytically and experimentally shown that the central machine can recover the underlying graph if large enough sample sizes are transmitted to the central machine.

Our experiments show that, under the same conditions, Signs method outperforms the uncoded scheme. Both methods have small FPR which is close to the FPR obtained by the original data.

Figure 2. TPR and FPR as a function of dimension $d$ for sample sizes $n = 1000$ and $n = 10,000$

(a) $n = 1000$ (b) $n = 1000$ (c) $n = 10,000$ (d) $n = 10,000$



(a) $d = 50$ (b) $d = 50$ (c) $d = 100$ (d) $d = 100$

Figure 3. TPR and FPR as a function of the sample size $n$ for random graphs with 50 and 100 nodes.



Figure 4. Probability of perfect structure recovery for the star graph with $d = 70$.



Figure 5. TPR as a function of SNR in Uncoded method for a random graph with $d = 40$ and maximum degree $\Delta = 5$.

## VIII. Acknowledgment

## References

[1] 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.

[2] 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.

[3] 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.

[4] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu *et al.*, "High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[5] S. Sojoudi, "Equivalence of graphical lasso and thresholding for sparse graphs," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3943–3963, 2016.

[6] 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.

[7] 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.

[8] A. Anandkumar, V. Y. Tan, F. Huang, and A. S. Willsky, "High-dimensional gaussian graphical model selection: Walk summability and local separation criterion," *Journal of Machine Learning Research*, vol. 13, no. Aug, pp. 2293–2337, 2012.

[9] M. Tavassolipour, S. A. Motahari, and M. T.-M. Shalmani, "Learning of tree-structured gaussian graphical models on distributed data under communication constraints," *IEEE Transactions on Signal Processing*, 2018.

[10] 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.

[11] V. Y. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A large-deviation analysis of the maximum-likelihood learning of markov tree structures," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1714–1735, 2011.

[12] N. Meinshausen, P. Bühlmann *et al.*, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[13] S. Chen, D. M. Witten, and A. Shojaie, "Selection and estimation for mixed graphical models," *Biometrika*, vol. 102, no. 1, pp. 47–64, 2014.

[14] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "Quic: quadratic approximation for sparse inverse covariance estimation," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2911–2947, 2014.

[15] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu *et al.*, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[16] Z. Meng, D. Wei, A. Wiesel, and A. Hero III, "Distributed learning of gaussian graphical models via marginal likelihoods," in *Artificial Intelligence and Statistics*, 2013, pp. 39–47.

[17] A. Wiesel and A. O. Hero, "Distributed covariance estimation in gaussian graphical models," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 211–220, 2012.

[18] J. Arroyo and E. Hou, "Efficient distributed estimation of inverse covariance matrices," in *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. IEEE, 2016, pp. 1–5.

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

[20] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[21] R. H. Bacon, "Approximations to multivariate normal orthant probabilities," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 191–198, 1963.

[22] 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.

**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 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.

**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.

**Armin Karamzade** is M. Sc. student of Artificial Intelligence at the Sharif University of Technology, Tehran, Iran. He received his B. Sc. degree in Computer Engineering from Iran University of Science and Technology, Tehran, Iran, in 2017. His research interests include mostly machine learning, optimization, and statistic.

**Reza Mirzaeifard** received the B.Sc. degree of computer engineering from Shahid Beheshti University, Tehran, Iran, in 2017. Currently, He is M.Sc. student of Artificial Intelligence program at Computer Engineering Department of Sharif University of Technology. His research interests include machine learning, information theory, tensor decomposition, bioinformatics.