Overlapping Decomposition for Causal Graphical Modeling

Lei Han[†], Guojie Song[†,*], Gao Cong[‡], Kunqing Xie[†]
[†]Key Laboratory of Machine Perception (Ministry of Education), EECS, Peking University, China
{hanlei, gjsong, xkq}@cis.pku.edu.cn
[‡]School of Computer Engineering, Nanyang Technological University, Singapore
gaocong@ntu.edu.sg

ABSTRACT

Causal graphical models are developed to detect the dependence relationships between random variables and provide intuitive explanations for the relationships in complex systems. Most of existing work focuses on learning a single graphical model for all the variables. However, a single graphical model cannot accurately characterize the complicated causal relationships for a relatively large graph. In this paper, we propose the problem of estimating an overlapping decomposition for Gaussian graphical models of a large scale to generate overlapping sub-graphical models. Specifically, we formulate an objective function for the overlapping decomposition problem and propose an approximate algorithm for it. A key theory of the algorithm is that the problem of solving a $k+1$ node graphical model can be reduced to the problem of solving a one-step regularization based on a solved k node graphical model. Based on this theory, a greedy expansion algorithm is proposed to generate the overlapping subgraphs. We evaluate the effectiveness of our model on both synthetic datasets and real traffic dataset, and the experimental results show the superiority of our method.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics

Keywords

Causality, Graphical Model, Overlapping Decomposition

1. INTRODUCTION

Causal graphical models are established to meaningfully characterize causal or statistical relationships that exist among variables of interest and quantify them. The problem of characterizing the causal relationships between variables in complex systems, such as economics, biological systems, traffic systems, climate change, etc., is important and fundamental. For example, economists want to know whether *burning natural gas* is a causal factor for the *global warming*.

*Corresponding author. Email: gjsong@cis.pku.edu.cn

KDD'12, August 12–16, 2012, Beijing, China.

The Gaussian graphical model [7], which learns the causalities between variables through their covariance, is one of the most promising causal modeling methods. It has been successfully employed in many applications, such as mining the causalities of climate attributes [2], gene regulatory network discovery [5], etc. In addition, several causal models on temporal evolving graphs have been proposed with applications in cross-species gene expression analysis [11], oil-production equipment stage capture [10] and climate research [12].

These methods construct a single graphical model to capture all the causalities between variables, treating all the variables together. They are typically developed for a small number of variables (usually in the order of tens). We will discuss theoretically in Section 4.2 that learning causalities through a single Gaussian graphical model by the variable covariance will be inaccurate, when large number of variables are considered and small number of observations are available. As a matter of fact, a causality model with only 20 variables can be overwhelming and difficult to interpret at a global level [1, 15]. Worse still, it is much more challenging to understand and construct causal relationships using causal graphical models for a relatively large graph (e.g., with hundreds of variables), although many applications need to deal with large graphs with heterogenous and complicated relationships. For instance, a highway network in traffic systems often contains hundreds of sample nodes, whose observations are counts of passing vehicles collected by sensors. In such traffic networks, complicated causalities exist between the vehicle counts.

Therefore, it is essential to develop techniques to discover such causalities in a large network. To cope with the challenging problem, we propose to decompose a large graphical model into multiple overlapping sub-graphical models. For decomposing a graphical model, it is important to consider both the heterogeneity and homogeneity, where heterogeneity means the local causalities and homogeneity refers to the overlaps between sub-graphical models. For example in traffic systems, some crucial traffic nodes may highly correlate with several different local regions, and thus these important nodes should be considered as overlap (homogeneity) by these local regions; meanwhile, we also need to find the causalities within a region (heterogeneity).

Unfortunately, decomposing graphical models is NP-hard [15] even if overlaps are not allowed. When we allow overlaps, the decomposition problem becomes more challenging because the search space becomes larger, which is due to more combinations of subgraph structures than those in the non-overlapping case.

In this paper, we address the challenging problem of estimating an overlapping decomposition for Gaussian graphical models of a large scale. We propose a novel approximation algorithm with performance guarantee, which is based on a local subgraph expansion

strategy. Specifically, we first formulate the optimization problem with an objective function comprising the negative log-likelihood of the observations of Gaussian sub-graphical models and some penalized items to constrain the structure of the subgraphs. Unfortunately, the penalized log-likelihood methods [4, 19] in Gaussian graphical model cannot be used to solve the approximation problem. Instead, we propose an algorithm that starts with the initial small subgraphs and incrementally computes the new Gaussian graphical model when a new node is involved.

One key technique is that we prove that the problem of solving a $k+1$ node graphical model can be reduced to the problem of solving a one-step regularization based on a solved k node graphical model, referred to as additive expanding property. We study the correctness and accuracy of this technique with detailed analysis. Based on the technique, we propose a greedy expansion algorithm for generating the overlapping sub-graphical models.

We evaluate the proposed method with two sets of experiments: First we empirically verify the properties of the proposed overlapping decomposition method on synthetic networks, and compare with the single graphical model [4, 19] and the non-overlapping decomposition method (which is a special case of the proposed overlapping decomposition method). The experimental results demonstrate the advantages of our techniques. Second, we evaluate the proposed techniques on real-life traffic data by learning the causalities between traffic observation points (e.g., a on-ramp) and detecting the traffic regularity in large traffic networks, which is essential for traffic analysis.

In summary, our main contributions are four-fold:

1. We formulate an objective function for the problem of overlapping decomposition of graphical model, and reduce the problem to a local subgraph expansion problem.
2. We extend the penalized log-likelihood in Gaussian graphical model to satisfy an additive expanding property and demonstrate its correctness and accuracy with detailed asymptotic analysis.
3. We propose a constrained greedy subgraph expansion algorithm for generating the overlapping subgraphs.
4. We evaluate our method on both synthetic and real-life traffic data. Experimental results show the effectiveness and superiority of our overlapping decomposition theory.

The rest of this paper is organized as follows. In Section 2, we briefly review closely related work. Section 3 presents the preliminaries and the problem statement. In Section 4, we present the proposed method including two core techniques and a demonstration. Experimental studies are reported in Section 5. We conclude this paper and present future directions in Section 6.

2. RELATED WORK

Most of existing work on causal graphical models builds a single graphical model. This renders them impractical to relatively large graphs (with more than hundreds of nodes). To cope with larger set of time series variables, Ruan et al. [15] propose to cluster time series variables into groups such that strong causal relations appear only between time series within a group while the causal relation between inter-group variables is weak. The clustering problem is formulated as a regression coefficient sparsification problem for graphical model decomposition. However, the approach [15] only considers non-overlapping decompositions while ignoring the overlap between subgraphs, which exists in many real-world applications. Moreover, the approach [15] is developed for time series variables, and is based on the Vector Autoregressive model, a type of temporal graphical model, rather than Gaussian graphical model as we use.

Our work is closely related to the joint estimation methods for multiple graphical models that share common structures [3,6]. The joint estimation methods [3,6] are proposed to learn multiple graphical models on the data from different categories but with the same set of features (variables), considering both the underlying homogeneity and heterogeneity of networks. They estimate multiple graphical models for different categories of the features, but not decomposing the features themselves. We proceed to use the example application scenario [3] to illustrate these methods. Consider a set of webpages collected from computer science departments of universities, and we want to find the causalities between selected keywords (e.g., "book", "model", "problem", etc.) appearing in the collection. These keywords can be treated as features, and appear in webpages of different categories, such as "student", "faculty", "project", etc. These features may display different dependence structures for different categories while sharing some common causalities across categories. The joint estimation methods cannot be applied to solve our problem, and they cannot be employed to discover the complicated causal relationships in large feature networks. First, these methods do not consider the decomposition on features of a large graph. Second, these methods are developed for graphs with a small number of features (in the order of tens).

Our proposed approach is also related to detecting overlapped community structures [9]. Community structure detection aims to group similar nodes together based on known distance measurements of nodes or correlations among nodes themselves. In contrast, in our problem we aim to uncover the causalities among correlated nodes, and furthermore find subgraphs based on the discovered causalities but not the known properties of nodes themselves. Thus, our problem is essentially different from community structure detection.

3. PRELIMINARY AND PROBLEM STATEMENT

3.1 Preliminary: Gaussian graphical model

As a member of the causal graphical model, Gaussian graphical model(GGM) assumes the joint distribution of the variables to be Gaussian. In GGM, the dependence structure (or causality) is determined from the covariance matrix of the variables, and a natural way to evaluate the causalities is to estimate the inverse of the covariance matrix [7, 8, 18]. Consider p random variables $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$, each variable $\boldsymbol{x}_i$ having n observations $\boldsymbol{x}_i = (x_i^1, ..., x_i^n)^T$, where we usually have $n \gg p$. Without loss of generality, we assume $\boldsymbol{X}$ follows a multivariate Gaussian distribution $N(\boldsymbol{\mu}, \Sigma)$, where the mean vector $\boldsymbol{\mu}$ is p-dimensional and each element in covariance matrix Σ is the expected value $\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$. The causality matrix Ω is the inverse of the covariance matrix, i.e., $\Omega = \Sigma^{-1}$. There exists a causal relationship between variables x_i and x_j iff $\Omega_{ij} \neq 0$ [7,18]. Therefore, the key problem is to calculate Ω. The estimation of Ω can be obtained by minimizing the penalized log-likelihood criterion [4, 19],

$$\widehat{\Omega} = \arg\min_{\Omega} \quad tr(\widehat{\Sigma}\Omega) - \log|\Omega| + \lambda \sum_{i \neq j} |\theta_{ij}| \qquad (1)$$

where θ_{ij} is the element in Ω; $\widehat{\Sigma}$ is the sample covariance matrix estimated on input X; $|\cdot|$ and $tr(\cdot)$ are the determinant and the trace in matrix calculus, respectively; λ is a tuning parameter.

The part $tr(\widehat{\Sigma}\Omega) - \log|\Omega|$ of Equation 1 corresponds to the negative log-likelihood of the observations of a Gaussian graphical model. The part $\lambda \sum_{i \neq j} |\theta_{ij}|$ is called a ℓ_1 penalty, which is to shrink some of the off-diagonal elements in $\widehat{\Omega}$ to zero. The tuning parameter λ controls the sparsity of $\widehat{\Omega}$. This minimization problem can be solved efficiently by the graphical lasso algorithms proposed in [4, 19].

3.2 Problem Definition

Problem Definition: Given p random variables $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$, where p is large and each variable $\boldsymbol{x}_i$ has n observations, $\boldsymbol{x}_i = (x_i^1, ..., x_i^n)^T$, we aim to learn the causal relationships between these variables.

In other words, we aim to encode the structure of $\boldsymbol{X}$ with an undirected graph $G = (V, E)$, where each node v in $V = \{v_1, \cdots, v_p\}$ corresponds to a variable in $\boldsymbol{X}$. The edge set E indicates the causalities between any two variables. More precisely, if $\boldsymbol{x}_i$ is correlated to $\boldsymbol{x}_j$, then edge e_{ij} is included in E. Thus, our objective is to obtain E. As introduced in Section 3.1, E can be constructed by estimating the causality matrix Ω of $\boldsymbol{X}$. We add an edge e_{ij} to E iff $\Omega_{ij} \neq 0$.

Instead of creating a single Gaussian graphical model for G, we propose to construct K Gaussian sub-graphical models with overlaps to discover the causalities between variables. Each subgraph, corresponding to a Gaussian sub-graphical model, is denoted as $g_i = (SV_i, SE_i), 1 \leq i \leq K$. The causal relationships reflected in E, $E = \bigcup_i SE_i$, are the output.

The challenge is to generate the K sub-graphical models and estimate Ω_i for each SE_i. To achieve this, we proceed to define a new objective function, which helps to formulate the decomposition problem clearly. We first introduce some notations: K vectors $\{\Gamma_1, \cdots, \Gamma_K\}$, where each $1 \times p$ vector Γ_i represents the component of a subgraph g_i. Element γ_{ij} in Γ_i is 1 if node j appears in subgraph g_i, and 0 otherwise. In addition, we set $\sum_{i=1}^{K} \sum_{j=1}^{p} \gamma_{ij} \geq p$, because we allow overlaps and we do not restrict that every node has to be included in at least one subgraph. This is reasonable because if a node is independent from all others, it should be left alone.

Objective function: We formulate the overlapping decomposition for a graphical model into the problem of estimating a set of Ω_i by minimizing

$$\{\widehat{\Omega}_i\}_{i=1}^{K} = \arg \min_{\{\Omega_1, \cdots, \Omega_K\}} \sum_{i=1}^{K} \{tr(\widehat{\Sigma}_i \Omega_i) - \log|\Omega_i|\} + \lambda_1 \sum_{i=1}^{K} \sum_{j \neq k} |\theta_{i,jk}| + \lambda_2 \sum_{i=1}^{K} \|\Gamma_i\|_1^2 + \lambda_3 \sum_{i<i'} \|\Gamma_i \circ \Gamma_{i'}\|_1^2, \quad s.t. \sum_{i=1}^{K} \|\Gamma_i\|_1 \geq p, \tag{2}$$

where $\theta_{i,jk}$ is the element in Ω_i; $\|\cdot\|_1$ is the ℓ_1 norm; $\circ$ means the Hadamard product; λ_1, λ_2 and λ_3 are tuning parameters.

In Equation 2, $tr(\widehat{\Sigma}_i \Omega_i) - \log|\Omega_i|$ represents the negative log-likelihood of the observations of Gaussian sub-graphic model g_i. The equation includes three penalized items:

- penalty $\lambda_1 \sum_{i=1}^{K} \sum_{j \neq k} |\theta_{i,jk}|$ controls the sparsity of the causalities in each subgraph;
- penalty $\lambda_2 \sum_{i=1}^{K} \|\Gamma_i\|_1^2$ is a constraint on the size of each subgraph and balances them, because the sum of the square is small when the subgraphs have the similar size;
- penalty $\lambda_3 \sum_{i<i'} \|\Gamma_i \circ \Gamma_{i'}\|_1^2$ gives a constraint on the sizes of overlaps and balances the sizes of overlaps, because the Hadamard product between any two Γ_i and $\Gamma_{i'}$ denotes the common nodes they share.

Given p random variables $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$, and assume $\boldsymbol{X}$ is encoded with a large graph G, our problem of discovering the causality structures in G is formulated as finding a set of overlapped subgraphs based on Equation 2.

Challenge and Solution Overview: The overlapping decomposition problem is a combinational optimization problem and it is NP-hard even if the penalty factors are not considered [15] . The problem involves $Kp + Kp^2$ unknown variables in the worst case, where $\{\Gamma_1, \cdots, \Gamma_K\}$ needs Kp variables, and $\{\Omega_1, \cdots, \Omega_K\}$ needs at most Kp^2 variables. Note that $\{\Omega_1, \cdots, \Omega_K\}$ needs fewer variables if we do not allow overlaps. Such a large number of unknown variables makes this problem computationally challenging. Moreover, we even do not know how to select a best K for this problem.

Hence, instead of finding the optimal solution to the complicated function, we propose a novel approximation algorithm for solving the overlapping decomposition problem, called *local subgraph expansion*. Our algorithms adopt a bottom-up strategy that expand the initial subgraphs by adding selected nodes gradually until the structure of overlapped subgraphs will reach convergence.

During this process, a key operation is to choose whether to include a new node in a subgraph. This operation is invoked many times, and calls for efficient techniques. Specifically, assume that there is a k-node subgraph whose inner causal relationships have been detected. We want to know whether a node v_{k+1} should be added to it. A straightforward method is creating a new Gaussian graphical model on all the $k+1$ nodes. However, this ignores the known causal relationships in the k-node subgraph and is computationally expensive. Thus, a natural question is whether we can reuse the known causal relationships in a subgraph to detect the causal relationships between a new node and the subgraph. In the next section, we present the proposed approximation method with performance guarantees for the operation.

4. PROPOSED METHOD

In this section, we propose two techniques. The first technique is generalized by a theorem in Section 4.1. This technique is used to check whether a new variable (node) should be included in a subgraph. The technique extends the penalized log-likelihood criterion in Equation 1 so that it can be incrementally expanded to accommodate new nodes. We call it Additive Penalized Log-likelihood Expansion (APLE). In Section 4.2, we also show the correctness and accuracy for APLE, which also motivate the necessity to decompose a large graphical model into sub-graphical models from a theoretical view.

The second technique is a local greedy approach presented in Section 4.3. We define a fitness function based on APLE approach. Moreover, taking into account the constraints (penalties in Equation 2) on the subgraph structures, we develop the Constraint Greedy Subgraph Expansion (CGSE) algorithm, which can achieve the *local subgraph expansion* process.

4.1 Additive Penalized Log-likelihood Expansion

Suppose that $\ell(\Omega^{(k+1)})$ is a new penalized log-likelihood criterion computed by Equation 1, which is from adding a new variable x_{k+1} into a solved penalized log-likelihood criterion $\ell(\Omega^{(k)})$,

where

$$\Omega^{(k+1)} = \begin{bmatrix} \Omega^{(k)} & \boldsymbol{\theta} \\ \boldsymbol{\theta}^T & \theta_{k+1} \end{bmatrix}$$

and $\boldsymbol{\theta}$ is a $k \times 1$ causality vector between x_{k+1} and $\{x_1, ..., x_k\}$ which we want to get. Since $\Omega^{(k)}$ is already solved (we assume it is solved by graphical lasso), it is positive definite [4, 19]. Thus, θ_{k+1} controls whether $\Omega^{(k+1)}$ is positive definite, and determines the $k+1\ th$ eigenvalue of $\Omega^{(k+1)}$. Also we select θ_{k+1} to guarantee $\Omega^{(k+1)}$ positive definite.

***Theorem 1*:** Suppose $\widehat{\Omega}^{(k)}$ is a local minimizer of $\ell(\Omega^{(k)})$, then there exists a local minimizer of $\ell(\Omega^{(k+1)})$, $\widehat{\Omega}^{(k+1)}$, such that $\widehat{\Omega}^{(k+1)} = (\widehat{\Omega}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)})$, where $\widehat{\boldsymbol{\theta}}^{(k)}$ is a local minimizer of

$$\begin{aligned} min \quad & 2\widehat{\boldsymbol{\varepsilon}}^T\boldsymbol{\theta} + \widehat{\varepsilon}_{k+1}\theta_{k+1} \\ -log(\theta_{k+1} - \boldsymbol{\theta}^T \cdot \widehat{\boldsymbol{\Sigma}}^{(k)} \cdot \boldsymbol{\theta}) & + 2\lambda_\theta \|\boldsymbol{\theta}\|_1 \end{aligned} \tag{3}$$

where $(\widehat{\boldsymbol{\varepsilon}}, \widehat{\varepsilon}_{k+1})$ is the sample covariance vector between x_{k+1} and $\{x_1, ..., x_{k+1}\}$.

***Proof*:** The new penalized log-likelihood criterion $\ell(\Omega^{(k+1)})$ is

$$\ell(\Omega^{(k+1)}) = tr(\widehat{\Sigma}^{(k+1)}\Omega^{(k+1)}) - \log|\Omega^{(k+1)}| + \lambda \sum_{i \neq j}^{k+1} |\theta_{ij}|$$

Compactly, we introduce three symbols to denote the items

$$I_1^{(k+1)} = tr(\widehat{\Sigma}^{(k+1)}\Omega^{(k+1)})$$

$$I_2^{(k+1)} = \log|\Omega^{(k+1)}|$$

$$I_3^{(k+1)} = \lambda \textstyle\sum_{i \neq j}^{k+1} |\theta_{ij}|$$

We unfold $\widehat{\Sigma}^{(k+1)}$ and $\Omega^{(k+1)}$ into block matrices to detect the relationship between $\ell(\Omega^{(k+1)})$ and $\ell(\Omega^{(k)})$. The unfolded matrices are derived as

$$\begin{aligned} I_1^{(k+1)} &= tr(\begin{bmatrix} \widehat{\Sigma}^{(k)} & \widehat{\boldsymbol{\varepsilon}} \\ \widehat{\boldsymbol{\varepsilon}}^T & \widehat{\varepsilon}_{k+1} \end{bmatrix} \cdot \begin{bmatrix} \Omega^{(k)} & \boldsymbol{\theta} \\ \boldsymbol{\theta}^T & \theta_{k+1} \end{bmatrix}) \\ &= tr(\begin{bmatrix} \widehat{\Sigma}^{(k)}\Omega^{(k)} + \widehat{\boldsymbol{\varepsilon}}\boldsymbol{\theta}^T & \widehat{\Sigma}^{(k)}\boldsymbol{\theta} + \theta_{k+1}\widehat{\boldsymbol{\varepsilon}} \\ \widehat{\boldsymbol{\varepsilon}}^T\Omega^{(k)} + \widehat{\varepsilon}_{k+1}\boldsymbol{\theta}^T & \widehat{\boldsymbol{\varepsilon}}^T\boldsymbol{\theta} + \widehat{\varepsilon}_{k+1}\theta_{k+1} \end{bmatrix}) \\ &= tr(\widehat{\Sigma}^{(k)}\Omega^{(k)}) + tr(\widehat{\boldsymbol{\varepsilon}}\boldsymbol{\theta}^T) + \widehat{\boldsymbol{\varepsilon}}^T\boldsymbol{\theta} + \widehat{\varepsilon}_{k+1}\theta_{k+1} \\ &= I_1^{(k)} + 2\widehat{\boldsymbol{\varepsilon}}^T\boldsymbol{\theta} + \widehat{\varepsilon}_{k+1}\theta_{k+1} \end{aligned}$$

$$\begin{aligned} I_2^{(k+1)} &= \log\begin{vmatrix} \Omega^{(k)} & \boldsymbol{\theta} \\ \boldsymbol{\theta}^T & \theta_{k+1} \end{vmatrix} \\ &= \log\left(|\Omega^{(k)}| \cdot |\theta_{k+1} - \boldsymbol{\theta}^T(\Omega^{(k)})^{-1}\boldsymbol{\theta}|\right) \\ &= I_2^{(k)} + \log\left(\theta_{k+1} - \boldsymbol{\theta}^T\widehat{\Sigma}^{(k)}\boldsymbol{\theta}\right) \end{aligned}$$

$$I_3^{(k+1)} = \lambda \sum_{i \neq j}^{k+1} |\theta_{ij}| = I_3^{(k)} + 2\lambda\|\boldsymbol{\theta}\|_1$$

where $\widehat{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\theta}$ are $k \times 1$ vectors; the derivation of $I_2^{(k+1)}$ is obtained by Leibniz formula. Finally, we can get

$$\ell(\Omega^{(k+1)}) = \ell(\Omega^{(k)}) +$$

$$2\widehat{\boldsymbol{\varepsilon}}^T\boldsymbol{\theta} + \widehat{\varepsilon}_{k+1}\theta_{k+1} - \log\left(\theta_{k+1} - \boldsymbol{\theta}^T\widehat{\Sigma}^{(k)}\boldsymbol{\theta}\right) + 2\lambda\|\boldsymbol{\theta}\|_1 \tag{4}$$

□

Similarly, we use $\ell(\boldsymbol{\theta}^{(k)})$ to represent Equation 4, then we have

$$\ell(\Omega^{(k+1)}) = \ell(\Omega^{(k)}) + \ell(\boldsymbol{\theta}^{(k)}) \tag{5}$$

Thus, it is understandable that to solve $\ell(\Omega^{(k+1)})$ based on a solved $\ell(\Omega^{(k)})$, we just need to solve $\ell(\boldsymbol{\theta}^{(k)})$ for an additional problem. Note that Equation 3 is exactly a ℓ_1 regularization problem which can be solved efficiently using the algorithms in [16,17].

One should be noted that the correctness and accuracy of $\widehat{\Omega}^{(k+1)} = (\widehat{\Omega}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)})$ has not been completely guaranteed by the above proof, because parameter λ_θ must be selected appropriately to make sure $(\widehat{\Omega}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)})$ equals the local minimizer $\widehat{\Omega}^{(k+1)}$. In other words, λ_θ has to be adapted corresponding to k as the expansion continues (until subgraphs converge). Since λ_θ controls the sparsity of $\widehat{\boldsymbol{\theta}}^{(k)}$, if λ_θ stays unchangeable, when the subgraph expands, constraint on $\|\widehat{\boldsymbol{\theta}}^k\|_1$ will become inappropriate which leads to the uncertainty of $\widehat{\Omega}^{(k+1)} = (\widehat{\Omega}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)})$.

Thus, a new question raises that how to adapt λ_θ to make the evaluation accurate when the subgraph expands. In next subsection, we will discuss the asymptotic behavior of APLE and show how to select λ_θ to ensure the correctness and accuracy of *Theorem 1*.

4.2 Asymptotic Analysis of APLE

Asymptotic property is crucial for APLE which makes sure that $\widehat{\Omega}^{(k+1)}$ can be obtained by $\widehat{\Omega}^{(k)}$ and $\widehat{\boldsymbol{\theta}}^{(k)}$ separately under an appropriate λ_θ. A detailed asymptotic analysis of Equation 1 has been discussed in [14]. Inspired by it, we establish the analysis to our APLE approach as follows.

Let the true causality matrix of $\ell(\Omega)$ be Ω_0, the true covariance matrix be Σ_0, $\Omega_0 = (\Sigma_0)^{-1}$, as well as true causality vector $\boldsymbol{\theta}_0$ and true covariance vector $\boldsymbol{\varepsilon}_0$. Let $\|\cdot\|_F$ be the Frobenius norm. We make the following assumptions.

A1: There exists a constant η such that $0 < \varphi_{max}(\Omega_0^{(k)}) \leq \eta$, where $\varphi_{max}(\cdot)$ denotes the maximum eigenvalue.

A2: There exist constants σ_1 and σ_2 such that $\sigma_1 \leq \theta_{k+1} \leq \sigma_2$ will guarantee $\Omega^{(k+1)}$ positive definite and $\varphi_{max}(\Omega_0^{(k+1)}) \leq \eta$. (Note that θ_{k+1} determines the $k+1\ th$ eigenvalue of $\Omega^{(k+1)}$).

Theorem 2: Let $\widehat{\boldsymbol{\theta}}^{(k)}$ be the local minimizer in Equation 3. Under A1 and A2, if $\lambda_\theta = C_0\sqrt{\frac{\log k}{n}}$, C_0 is a positive constant, then

$$\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}_0^{(k)}\|_F = O_P\left(\sqrt{\frac{k\log k}{n}}\right) \tag{6}$$

where $O_P(\cdot)$ is the order in probability.

Proof Let $G(\Delta_\theta) = \ell(\boldsymbol{\theta}_0 + \Delta_\theta) - \ell(\boldsymbol{\theta}_0)$. Assume that there exists a bounded convex set

$$\mathcal{G} = \{\Delta_\theta : \|\Delta_\theta\|_F \leq Mr_n\},$$

where M is a positive constant and

$$r_n = \sqrt{\frac{k\log k}{n}} \to 0 \quad (n \gg k)$$

Note that $G(\Delta_\theta)$ is a convex function, if we demonstrate that G is strictly positive everywhere on the boundary $\partial\mathcal{G}$ ($\|\Delta_\theta\|_F = Mr_n$), then G has a local minimum inside $\mathcal{G}$. Actually,

$$G(\Delta_\theta) = \ell(\boldsymbol{\theta_0} + \Delta_\theta) - \ell(\boldsymbol{\theta_0})$$
$$= 2\widehat{\boldsymbol{\varepsilon}}^T\Delta_\theta - (\log(\theta_{k+1} - (\boldsymbol{\theta}_0^T + \Delta_\theta)\widehat{\Sigma}^{(k)}(\boldsymbol{\theta}_0 + \Delta_\theta)) \quad (7)$$
$$-\log(\theta_{k+1} - \boldsymbol{\theta}_0^T\widehat{\Sigma}^{(k)}\boldsymbol{\theta}_0)) + 2\lambda_\theta(\|\boldsymbol{\theta}_0 + \Delta_\theta\|_1 - \|\boldsymbol{\theta}_0\|_1)$$

For the subtraction of the logarithm terms in Equation 7, assume that $f(\boldsymbol{\theta}) = \log(\theta_{k+1} - \boldsymbol{\theta}^T\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\theta})$, we can get from the derivation in proof of *Theorem 1* that $f(\boldsymbol{\theta}) = I_2^{(k+1)} - I_2^{(k)}$, therefore

$$f(\boldsymbol{\theta}_0 + \Delta_\theta) - f(\boldsymbol{\theta}_0) = (\log|\Omega_0^{(k+1)} + \Delta^{(k+1)}| - \log|\Omega_0^{(k+1)}|) -$$
$$(\log|\Omega_0^{(k)} + \Delta^{(k)}| - \log|\Omega_0^{(k)}|)$$

As has been proved by [14], for Ω_0 we have

$$\log|\Omega_0 + \Delta| - \log|\Omega_0|$$
$$= tr(\Sigma_0\Delta) - \widetilde{\Delta}^T[\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1}dv]\widetilde{\Delta}$$

where

$$F = \widetilde{\Delta}^T[\int_0^1 (1-v)(\Omega_0 + v\Delta)^{-1} \otimes (\Omega_0 + v\Delta)^{-1}dv]\widetilde{\Delta}$$
$$\geq \frac{1}{4\eta^2}\|\Delta\|_F^2$$

Thus we have

$$f(\boldsymbol{\theta}_0 + \Delta_\theta) - f(\boldsymbol{\theta}_0)$$
$$= (tr(\Sigma_0^{(k+1)}\Delta^{(k+1)}) - tr(\Sigma_0^{(k)}\Delta^{(k)})) - (F^{(k+1)} - F^{(k)})$$
$$\approx 2\boldsymbol{\varepsilon}_0^T\Delta_\theta - (F^{(k+1)} - F^{(k)})$$

Thus, we get

$$G(\Delta_\theta) = 2(\widehat{\boldsymbol{\varepsilon}}^T - \boldsymbol{\varepsilon}_0^T)\Delta_\theta + (F^{(k+1)} - F^{(k)})$$
$$+2\lambda_\theta(\|\boldsymbol{\theta}_0 + \Delta_\theta\|_1 - \|\boldsymbol{\theta}_0\|_1)$$

For each item in $G(\Delta_\theta)$ we have the following boundaries

$$B1 : |(\widehat{\boldsymbol{\varepsilon}}^T - \boldsymbol{\varepsilon}_0^T)\Delta_\theta| \leq C_1\sqrt{\frac{\log k}{n}}\|\Delta_\theta\|_1 \leq C_1\sqrt{\frac{k\log k}{n}}\|\Delta_\theta\|_F$$

$$B2 : F^{(k+1)} - F^{(k)} \geq \frac{1}{4\eta^2}(\|\Delta^{(k+1)}\|_F^2 - \|\Delta^{(k)}\|_F^2) \geq \frac{1}{2\eta^2}\|\Delta_\theta\|_F^2$$

$$B3 : \lambda_\theta(\|\boldsymbol{\theta}_0 + \Delta_\theta\|_1 - \|\boldsymbol{\theta}_0\|_1) \leq \lambda_\theta\|\Delta_\theta\|_1 \leq \lambda_\theta\sqrt{k}\|\Delta_\theta\|_F$$

where B1 is a boundary from [14], and B3 can be obtained by mean inequalities. Combine all the above items and finally we can get

$$G(\Delta_\theta) \geq \frac{1}{2\eta^2}\|\Delta_\theta\|_F^2 - 2C_1\sqrt{\frac{k\log k}{n}}\|\Delta_\theta\|_F - 2\lambda_\theta\sqrt{k}\|\Delta_\theta\|_F$$
$$= \|\Delta_\theta\|_F^2(\frac{1}{2\eta^2} - (2C_1\sqrt{\frac{k\log k}{n}} + 2\sqrt{k}\lambda_\theta)\|\Delta_\theta\|_F^{-1})$$

Take $\lambda_\theta = C_0\sqrt{\frac{\log k}{n}}$,

$$G(\Delta_\theta) \geq \|\Delta_\theta\|_F^2(\frac{1}{2\eta^2} - \frac{2C_1 + 2C_0}{M})$$

for M sufficiently large we can get $G(\Delta_\theta) > 0$.

□

According to the theorem in [3, 14], under certain assumptions, local minimizer $\widehat{\Omega}^{(k)}$ of Equation 1 satisfies

$$\|\widehat{\Omega}^{(k)} - \Omega_0^{(k)}\|_F = O_P(\sqrt{\frac{(k+s_k)\log k}{n}}) \quad (8)$$

where s_k is the count of non-zero off-diagonal elements in $\Omega_0^{(k)}$ ($s_k = \frac{k(k-1)}{2}$ would give a full matrix). For a local minimizer $\widehat{\Omega}^{(k+1)}$, we have

$$\|\widehat{\Omega}^{(k+1)} - \Omega_0^{(k+1)}\|_F^2 \approx \|\widehat{\Omega}^{(k)} - \Omega_0^{(k)}\|_F^2 + 2\|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}_0^{(k)}\|_F^2$$

Thus, combined with our *Theorem 2*, it can be easily found

$$O_P(\frac{(k+s_k)\log k}{n}) + O_P(\frac{2k\log k}{n})$$
$$\approx O_P(\frac{(k+1+s_{k+1})\log(k+1)}{n})$$

since $0 \leq s_{k+1} - s_k \leq 2k$.

So, if $\widehat{\boldsymbol{\theta}}^{(k)}$ is a local minimizer of Equation 3, we know that there exists a local minimizer $\widehat{\Omega}^{(k+1)}$ of $\ell(\Omega^{(k+1)})$ that ensures $\widehat{\Omega}^{(k+1)} = (\widehat{\Omega}^{(k)}, \widehat{\boldsymbol{\theta}}^{(k)})$. So far, the correctness and accuracy of *Theorem 1* has been guaranteed completely under *Theorem 2*.

It is worth mentioning that both Equation 8 and our *Theorem 2* are in line with the motivation of the decomposition on large scale graphical model. Note that both of them show that when the number of variables k is relatively large or exceeds the number of observations n of each variable, the error on the estimation will increase dramatically.

This suggests that, when we consider only a single graphical model on a large network, the result will be inaccurate especially when there are not enough observations to establish such a large graphical model. An example is that traffic systems often contain hundreds of ramps (variables), and the number of the observations for each ramp is limited by the sampling quantity. The periodicity of traffic behaviors is often measured by days. Thus, if we want to know the causalities between observations at a specific time period in a day, we can just get one value for each ramp one day. Therefore, the decomposition of a large graphical model is necessary.

4.3 Constraint Greedy Subgraph Expansion

We present the algorithm for the *local subgraph expansion* process based on our APLE approach. We consider the constraints corresponding to the penalized items in Equation 2, and apply them to the *local subgraph expansion* process in this subsection.

Add Constraints: When there is a new node (variable) x_{new} joint in a solved k-node subgraph g to do expansion, based on APLE we can obtain a new causality vector $\widehat{\boldsymbol{\theta}}_{new} = APLE(g, x_{new}, \lambda_\theta)$. Let $E_{\widehat{\theta}} = \{i : \widehat{\theta}_i \neq 0, 1 \leq i \leq k\}$, we define the fitness as

$$Fitness(\widehat{\boldsymbol{\theta}}_{new}) = e^{-\gamma o}\frac{|E_{\widehat{\theta}}|}{k} \quad (9)$$

where o is the number of subgraphs to which node v_{new} has been mapped, and thus γ controls the degree of overlaps, which can be regarded as a constraint. With γ, we have that the causality contributions of v_{new} to other subgraphs reduce as the number of subgraphs to which v_{new} has already been mapped increases.

Because the fitness in Equation 9 is always nonnegative, a threshold ϵ_f should be given as the minimum accepted fitness, which is actually a constraint on the size of each subgraph.

After each iteration of expansion, we check whether there are

near-duplicated subgraphs based on the following equation.

$$max\{\frac{|SV_i \cap SV_j|}{|SV_i|}, \frac{|SV_i \cap SV_j|}{|SV_j|}\} > \epsilon_o, \quad (10)$$

where SV_i is the set of edges in subgraph g_i and SV_j is for g_j; ϵ_o is the combination threshold.

We combine subgraphs g_i and g_j into a new subgraph if the above equation is satisfied. Here ϵ_o balances the sizes of overlaps.

Adaption of λ_θ: It has been mentioned above that as the subgraph expands, λ_θ has to be adapted to make APLE correct and accurate. According to *Theorem 2*, we know that $\lambda_\theta^{(k)} = C_0\sqrt{\frac{\log k}{n}}$, and thus when k expands to $k+1$, we have

$$\lambda_\theta^{(k+1)} = C_0\sqrt{\frac{\log(k+1)}{n}} = \sqrt{\log_k(k+1)}\lambda_\theta^{(k)} \quad (11)$$

In the following algorithm we will update λ_θ based on Equation 11.

The proposed Constraint Greedy Subgraph Expansion(CGSE) algorithm is outlined in *Algorithm 1*.

Algorithm 1 CGSE Algorithm

Input: (1)p random variables $X = \{x_1, ..., x_p\}$ where x_i contains n observations; (2)$\mathcal{K}$ initial seeds $S = \{S_1, ..., S_\mathcal{K}\}$ where $|S_1| = ... = |S_\mathcal{K}|$;
Parameters: (1)fitness threshold ϵ_f; (2)combination threshold ϵ_0; (3)initial tuning parameter λ_0;
Output: Correlations among p variables and the overlapping subgraphs;

1: $g = S$;
2: $K = \mathcal{K}$;
3: **for** $i = 1$ to K **do**
4: $\lambda_i = \lambda_0$;
5: **end for**
6: **repeat**
7: **for** $i = 1$ to K **do**
8: Find an unvisited variable x_j from the nodes that are not in subgraph g_i;
9: $k = |g_i|$;
10: $\lambda_i = (\log_{(k-1)} k)^{1/2}\lambda_i$;
11: $\boldsymbol{\theta}_{new} = APLE(g_i, x_j, \lambda_i)$;
12: **if** $Fitness(\boldsymbol{\theta}_{new}) > \epsilon_f$ **then**
13: Add x_j into g_i;
14: **end if**
15: **end for**
16: **for** each g_i in g **do**
17: **for** each g_j in g $(j \neq i)$ **do**
18: **if** $|g_i \cap g_j|/|g_i| > \epsilon_o$ **or** $|g_i \cap g_j|/|g_j| > \epsilon_o$ **then**
19: Combine g_j into g_i;
20: **end if**
21: **end for**
22: **end for**
23: $K = |g|$;
24: **until** Each subgraph stays unchangeable
25: Output g;

Algorithm Explanation: Without loss of generality, S can be selected randomly as long as the seeds in S are disjoint with each other. Lines 3–5 initialize the tuning parameter λ_0 for each seed. Lines 7–15 give one step expansion for each subgraph. We expand all the subgraphs together, which can achieve a balance for the size of each subgraph. Lines 16–22 check if two subgraphs should be combined.

Complexity Analysis: Assume that the final average subgraph size is R, lines 7–15 can be computed in $O(\mathcal{K} \cdot L(R))$ time. Lines 16–22 take at most $O(\mathcal{K}^2 p)$ time with auxiliary $O(p)$ space. The number of iteration of line 6 reaches p at most. Thus our CGSE algorithm takes $O(\mathcal{K}^2 p^2 + \mathcal{K} p \cdot L(R))$ time in the worst case. Where $L(R)$ is the complexity of ℓ_1 regularization method in [16], which is logarithmic complexity with R [16].

5. EXPERIMENTAL STUDY

We evaluate the proposed overlapping decomposition of graphical model (ODGM). We compare with the single graphical model(SGM), which is solved by the graphical lasso [4]. To further study the advantage of the overlapping decomposition, we adapt the proposed CGSE algorithm to support the non-overlapping decomposition of graphical model (NODGM) by setting $\gamma = +\infty$ in Equation 9 to forbid overlaps.

We report results on synthetic datasets in Section 5.1. In Section 5.2, we report the performance study on real-life traffic dataset, and show the usefulness of the results for traffic analysis.

5.1 Synthetic Data

5.1.1 Setting

Since we focus on graphical models of a relatively large scale, we generate a set of networks whose number of nodes, p, ranges from 100 to 900. Note that previous work normally uses networks containing tens of nodes. We set the number of observations $n = 800$. We follow the approach [6] to generate the synthetic data. To simulate the heterogeneity in large networks, we generate local centered network by K local Erdös-Rényi random graphs $\{g_1, g_2, ..., g_K\}$, $g_i = (SV_i, SE_i)$; for homogeneity we add edges between any g_i and g_j randomly. Specifically,

1. We generate K Erdös-Rényi graphs, each with a random size in $[20, 80]$, such that $\sum_i^K |SV_i| = p$. Let E_{cross} be the set of cross links between the K graphs, and let $E_{inner} = \bigcup_i^K SE_i$ be the set of total inner links. Let $\rho = |E_{cross}|/|E_{inner}|$ be a factor to control the homogeneity. We randomly add $\rho|E_{inner}|$ cross edges. Finally, we can get a network $G = (V, E)$, where $V = \bigcup_i^K SV_i$ and $E = E_{inner} \bigcup E_{cross}$.

2. Based on the above network, we create a covariance matrix following [13]. Define a $p \times p$ matrix A as

$$A_{ij} = \begin{cases} 1, & i = j \\ U([-1, -0.5] \cup [0.5, 1]), & (i,j) \in E \\ 0, & Else \end{cases}$$

where $U(\cdot)$ represents uniform distribution. We scale the diagonal elements to ensure positive definiteness and average the matrix with its transpose to get a symmetric A. Then the covariance matrix Σ is calculated as

$$\Sigma_{ij} = (A^{-1})_{ij}/\sqrt{(A^{-1})_{ii}(A^{-1})_{jj}}$$

3. We generate p dimensional samples from N(0,Σ).

We define Precision, Recall and F1-score to measure the effectiveness of different models in finding the causal relationships. Note that the true causal relationships in E are known in the generated data. Given an estimated $\widehat{E}$ returned by a method, we define these metrics as follows.

$$Pre = \frac{|\{(i,j) : (i,j) \in E, (i,j) \in \widehat{E}\}|}{|\{(i,j) : (i,j) \in \widehat{E}\}|}$$

$$Rec = \frac{|\{(i,j) : (i,j) \in E, (i,j) \in \widehat{E}\}|}{|\{(i,j) : (i,j) \in E\}|}$$

$$F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec}$$

Moreover, we set the fitness threshold $\epsilon_f = 0.1$, combination threshold $\epsilon_o = 0.6$, $\gamma = 0.1$ and $\lambda_0 = 0.2$. We set the number of seeds at $|S| = K$ and each S_i is selected randomly from the K Erdös-Rényi graphs with size $|S_i| = 3$.

5.1.2 Results

Varying p To evaluate these methods on networks of various sizes, we vary p from 100 to 900. We set $\rho = 0.3$. The performances of all the methods are shown in Figure 1. We can see that when p is small, SGM performs as well as ODGM, because a single graphical model can work well. However, as p increases, the accuracy of SGM falls dramatically. As explained in Section 4.2, such a large p makes it infeasible to derive a single graphical model from n observations. However, both decomposition methods still work well with the increase of p. ODGM achieves a high accuracy and outperforms NODGM consistently, because non-overlapping decomposition cannot capture the overlap information.

Varying ρ The parameter ρ plays an important role on controlling the homogeneity of the network. When $\rho = 0$, it means the network is essentially heterogeneous and is actually composed of several separate sub-networks, while a large ρ indicates that the edges in the network tend to distribute homogeneously. Figure 2 shows the F1-score of ODGM and NODGM while ρ is varied, where we set $p = 500$. As expected, when ρ approaches zero, NODGM performs as well as ODGM because the network can be divided completely into sub-networks. But as ρ increases, the disparity between ODGM and NODGM becomes larger since ODGM can discover the overlaps while NODGM losses more information.

We do not report the results for varying the parameters ϵ_f, ϵ_o and γ due to space limitation. Instead, we will study and visualize the effect of these parameters on a real-life traffic network in an intuitive way in the next section.

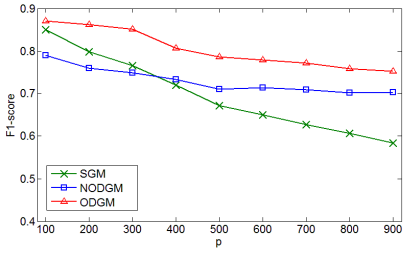


Figure 1: F1-score with p varying.

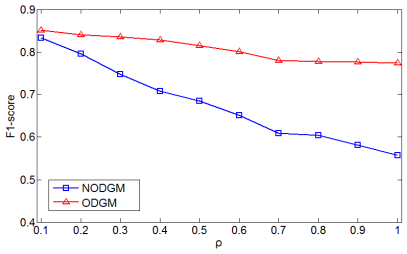
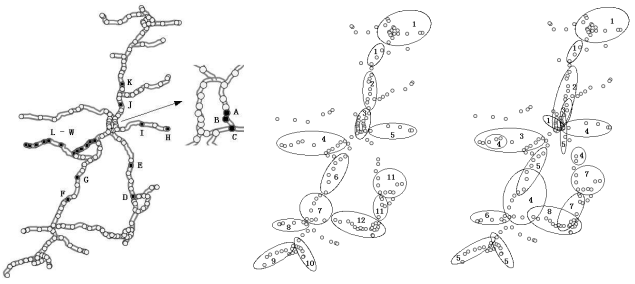


Figure 2: F1-score with ρ varying.

5.2 Traffic Data

5.2.1 Description and setting

We evaluate our methods on real-life traffic data. The features in this traffic dataset are observations collected from sensors located on ramps in a highway traffic network. Each observation is the vehicle count during a time interval. Figure 3(a) gives the structure of the highway traffic network from a province in China, in which each circle represents a traffic station consisting of an on-ramp and an off-ramp, and the line between any connected traffic stations is the bidirectional highway. There is an important ring in the network which is amplified on the right hand—the city in the center of this ring is a big city and plays a central role in the entire traffic network.



(a) (b) (c)

Figure 3: (a)Real traffic network; (b)the non-overlapping decomposition structure(NODGM); (c)the overlapping decomposition structure(ODGM).

There are total 180 traffic stations(circles), which correspond to 360 ramps, i.e., $p = 360$. The observations are collected at time interval 9:00-9:15 AM from 2011/1/1 to 2011/2/28 (59 days). Therefore, $n = 59$ for each feature. Due to the stability and periodicity of traffic behaviors, the observations follow a Gaussian distribution.

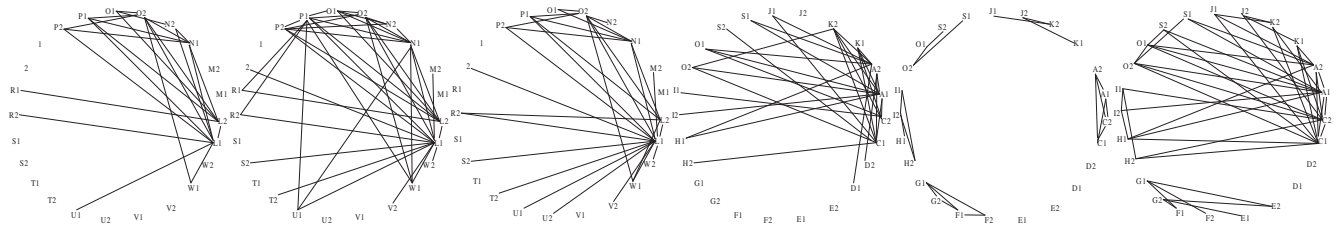
We set $\epsilon_f = 0.1$, $\epsilon_o = 0.6$, $\gamma = 0.1$ and $\lambda_0 = 50$. We set the number of seeds $|S| = 12$, and each $|S_i| = 6$.

Because there is no ground truth for causality matrix in real traffic data, F1-score cannot be measured. Nevertheless, the causal information detected is the most important for traffic research, and our domain experts can help with their knowledge on the causal relationship in the traffic network we use. Next, we compare the results returned from the different methods and discuss how the parameters in CGSE influence the causality structures.

5.2.2 Results and Analysis

Figure 3(b) and Figure 3(c) give the subgraph structures returned by NODGM and ODGM, respectively. For clear representation, we draw the results based on the initial traffic network with 180 traffic stations instead of 360 features, and a subgraph contains a traffic station node iff at least one feature (ramp) of this traffic station belongs to it. In the figures, the ellipses with the same label denote a subgraph. Since non-overlapping subgraphs have no intersections, the subgraphs cannot be combined together, and thus the number of final subgraphs equals to the number of seeds in Figure 3(b). For overlapping structure, when two subgraphs overlap at a certain rate ϵ_o, they are combined together. Thus, we end up with 8 subgraphs in Figure 3(c).

From the two figures, we observe: (1)Both NODGM and ODGM show that the causalities between vehicle flows follow the spatial distribution in general— the nearer are two features spatially in traffic network, the more correlated they tend to be; (2)ODGM highlights some crucial traffic nodes with highly overlaps, such as the nodes on the central ring. As mentioned earlier, the central ring is around the central city and plays an important role in the the entire traffic network. Additionally, traffic station C on the ring is the passageway connecting the unique airport of the entire network; and traffic stations A and B are the top 2 highest vehicle flow traffic stations on both on-ramp and off-ramp. These domain information



(a) SGM (b) NODGM (c) ODGM (d) SGM (e) NODGM (f) ODGM

Figure 4: Detail causalities among the selected features: (a), (b) and (c) are causalities by SGM, NODGM and ODGM under local concentrated features; (d), (e) and (f) are causalities by SGM, NODGM and ODGM under scattered features.

matches well with our ODGM result and gives an reasonable explanation. (3)ODGM is able to detect long distance causalities in addition to the local causalities within distances. For example, the components(ellipses) of subgraph 1, 3, 4 and 5 are distributed spatially, but they are highly correlated within the vehicle flow. In other words, there also exists long distance origin-destination demand in the traffic network. However, NODGM cannot mine such information described in both (2) and (3). (4) The sparsely located traffic stations are not included in any subgraphs in both the figures. We find that the vehicle flows in most of these traffic station ramps are nearly 0 during the observation periods, and almost 80% of the ramps and their located highways are newly built. Thus they are seldom used and have no causal relationship with other ramps.

Figure 4 gives the detailed causalities among a set of selected features. For each selected traffic station i, $i1$ and $i2$ denote the on-ramp and off-ramp features, respectively. In Figures 4(a), 4(b) and 4(c), the features are selected from traffic station L-W in Figure 3(a), and these traffic stations are selected locally concentrated. We can see that SGM detects fewer causal information than do N-ODGM and ODGM because a single graphical model treats the entire network globally, and can only detect the causalities from a global view. In this setting, NODGM detects more causal relationships than do ODGM, which also demonstrates that NODGM focuses more on a local view while ODGM is a nice compromise of SGM and NODGM.

While Figures 4(d), 4(e) and 4(f) provide the detailed causalities among the features selected from A, C-K, O and S, which are scattered in the network. As can be observed from the results, ODGM discovers more meaningful causal information than do SGM and NODGM, e.g., the causalities among $\{E1, E2, F1, F2, G1, G2\}$. Both ODGM and SGM are able to discover the important long distance causalities for the key traffic station A and C. However, NODGM is restrained by its non-overlapping structure and only detects the inner relationship within subgraphs, even if A and C are highly correlated with others.

These results obtained by ODGM are essentially important for the analysis of traffic systems for the following reasons. First, the traffic stations in the same subgraph are highly correlated and should be considered together by traffic systems. For example, it is possible that vehicle flows rush into each other within the same subgraph. Second, the causalities are very helpful for traffic flow prediction and anomaly detection which are hot concerns of traffic operators and managers. Third, it is important to find the highly overlapped traffic stations. These crucial traffic stations are correlated with a number of regions, based on which the regions with heavy traffic can be detected. On the other hand, the regions with light traffic can also be reflected by independent traffic stations. These information can be used by highway construction planners to design new roads.

5.2.3 *Varying parameters*

We study the effect of parameters γ, ϵ_f, ϵ_o and λ_0 for CGSE. Figure 5 shows the decomposition results of varying parameters. When varying each parameter, we use the aforementioned default values for the other parameters.

Parameter ϵ_f controls the minimum fitness, and restricts the size of each subgraph. As shown in Figure 5(a), when ϵ_f decreases, more features are added into subgraphs and the size of each subgraph becomes larger. Figure 5(b) shows the effect of ϵ_o. When ϵ_o is reduced, the subgraphs are more likely to be combined together under ϵ_o. Figure 5(c) shows the effect of γ, which controls overlaps, on the number of features with different overlap degrees. We observe that when γ increases, fewer overlaps exist in the decomposition structure, thus the number of overlapped features.

Figure 6 visualizes the generated subgraphs for selected parameter values to show the details. From these figures, we can see that the property of each parameters is in line with the results in Figure 5, and these figures give a more intuitive and understandable description for our method.

Due to space limitation and as λ_0 works similarly as ϵ_f, we do not give the results of varying λ_0, which controls the sparsity of the causalities in the penalize estimation problem in Equation 3. When λ_0 increases with ϵ_f fixed, the constraint on the sparsity becomes tighter and fewer causalities are detected, and thus smaller subgraphs generated. Conversely, when λ_0 decreases, more causalities will be discovered and the size of each subgraph will increase.

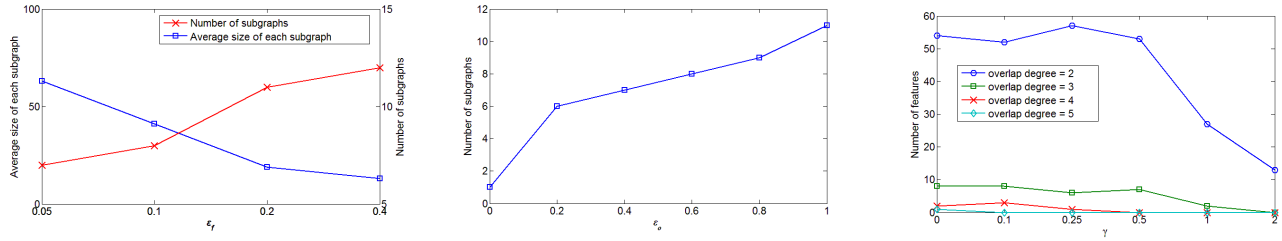
6. CONCLUSION AND FUTURE WORK

In this paper, we propose an overlapping decomposition technique for large scale graphical models. The techniques enables the penalized log-likelihood in Gaussian Graphical Model to satisfy an additive expanding property. We demonstrate its asymptotic stability. Based on this property, we develop a constraint greedy subgraph expansion algorithm to generate overlapped subgraphs. We demonstrate on both synthetic data and real-life traffic data that the overlapping decomposition method is more powerful than the single graphical model and its non-overlapping decomposition counterpart. In the application of traffic data analysis, the meaningful results show that our model can provide rich information for traffic analysis.

For future work, it is interesting to extend the static overlapping decomposition technique to deal with time-varying observations so that we can follow the evolvement of the causalities in a network.

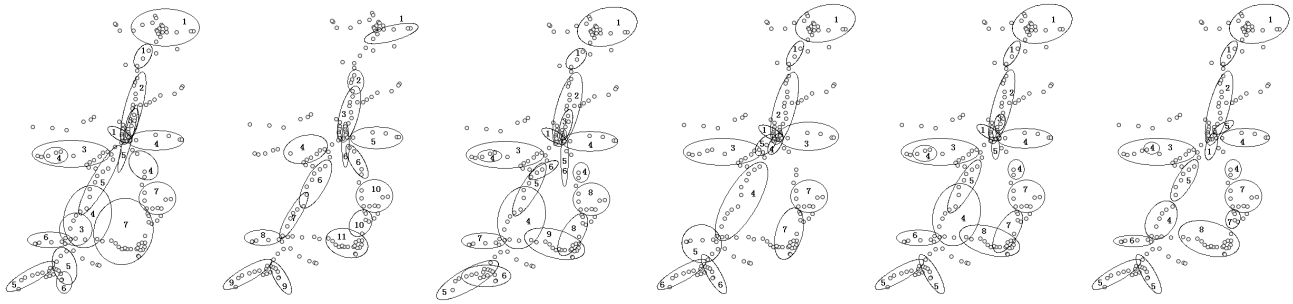
7. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (60703066, 60874082), and Beijing municipal natural science foundation (4102026).



(a) (b) (c)

Figure 5: (a)relationship among average subgraph size, the number of subgraphs and ϵ_f; (b) relationship between the number of subgraphs and ϵ_o; (c)relationship between overlap degree and γ.



(a) $\epsilon_f = 0.05$ (b) $\epsilon_f = 0.2$ (c) $\epsilon_o = 0.8$ (d) $\epsilon_o = 0.4$ (e) $\gamma = 0.5$ (f) $\gamma = 2$

Figure 6: Decomposition structures when varying each parameter while keeps others stable. Default setting: $\epsilon_f = 0.1$, $\epsilon_o = 0.6$ and $\gamma = 0.1$.

8. REFERENCES

[1] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75. ACM, 2007.

[2] X. Chen, Y. Liu, H. Liu, and J. Carbonell. Learning spatial-temporal varying graphs with applications to climate data analysis. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[3] P. Danaher, P. Wang, and D. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Arxiv preprint arXiv:1111.0324*, 2011.

[4] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[5] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science's STKE*, 303(5659):799, 2004.

[6] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[7] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[8] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[9] C. Lee, F. Reid, A. McDaid, and N. Hurley. Detecting highly overlapping community structure by greedy clique expansion. *Arxiv preprint arXiv:1002.1827*, 2010.

[10] Y. Liu, J. Kalagnanam, and O. Johnsen. Learning dynamic temporal graphs for oil-production equipment monitoring system. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234. ACM, 2009.

[11] Y. Liu, A. Niculescu-Mizil, A. Lozano, and Y. Lu. Temporal graphical models for cross-species gene regulatory network discovery. In *Proceedings of the 9th annual international conference on Computational Systems Bioinformatics*, 2010.

[12] A. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596. ACM, 2009.

[13] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[14] A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[15] N. Ruan, R. Jin, V. Lee, and K. Huang. A sparsification approach for temporal graphical model decomposition. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 447–456. IEEE, 2009.

[16] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. *Machine Learning: ECML 2007*, pages 286–297, 2007.

[17] M. Schmidt, G. Fung, and R. Rosales. Optimization methods for l1-regularization. *University of British Columbia, Technical Report TR-2009-19*, 2009.

[18] R. Thompson. Graphical models in applied multivariate statistics. *Journal of Classification*, 9(1):159–160, 1992.

[19] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.