Common Substructure Learning of Multiple Graphical Gaussian Models

Satoshi Hara and Takashi Washio

The Institute of Scientific and Industrial Research (ISIR), Osaka University, Japan
{hara,washio}@ar.sanken.osaka-u.ac.jp

Abstract. Learning underlying mechanisms of data generation is of great interest in the scientific and engineering fields amongst others. Finding dependency structures among variables in the data is one possible approach for the purpose, and is an important task in data mining. In this paper, we focus on learning dependency substructures shared by multiple datasets. In many scenarios, the nature of data varies due to a change in the surrounding conditions or non-stationary mechanisms over the multiple datasets. However, we can also assume that the change occurs only partially and some relations between variables remain unchanged. Moreover, we can expect that such commonness over the multiple datasets is closely related to the invariance of the underlying mechanism. For example, errors in engineering systems are usually caused by faults in the sub-systems with the other parts remaining healthy. In such situations, though anomalies are observed in sensor values, the underlying invariance of the healthy sub-systems is still captured by some steady dependency structures before and after the onset of the error. We propose a structure learning algorithm to find such invariances in the case of Graphical Gaussian Models (GGM). The proposed method is based on a block coordinate descent optimization, where subproblems can be solved efficiently by existing algorithms for *Lasso* and the *continuous quadratic knapsack problem*. We confirm the validity of our approach through numerical simulations and also in applications with real world datasets extracted from the analysis of city-cycle fuel consumption and anomaly detection in car sensors.

Keywords: Graphical Gaussian Model, common substructure, block coordinate descent.

1 Introduction

In the real world, it is common that multivariate data, such as the stock market [1], gene regulatory networks [2], or biomedical measurements [3], to have a complex dependency structure among variables. Such a structure is closely tied to the intrinsic data generating mechanism, which one aims to reveal. For example, we can expect the interaction of brain sub-regions to be reflected by the dependency structures between fMRI signals [3].

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part II, LNAI 6912, pp. 1–16, 2011.

The dependency structure among variables also plays an important role in the analysis of multiple datasets. It is frequently seen that datasets collected under different conditions have different dependency structures, which is caused by a change in the underlying mechanism [2,4]. On the other hand, if some relations are common to several conditions, we can expect that background mechanism to have a certain invariance against the change. An illustrative example is an engineering system where system errors are observed as dependency anomalies of sensor values [5]. These are usually caused by a fault in a sub-system. The invariance, which in this example is the remaining healthy sub-systems, is captured by a steady dependency over the multiple datasets sampled before and after the error onset.

Motivated by the example above, we propose a method for finding common dependency structures from multiple datasets. In this paper, we consider the case of the Graphical Gaussian Model (GGM) [6]. GGM is one of the most basic models representing linear dependencies among continuous variables. Identification of the structure was firstly studied by Dempster [7] where it was referred to as *covariance selection*. Though classical approaches have encountered several difficulties, there is a recent development on the use of ℓ_1-regularization [8,9,10], that enables the design of an efficient Graphical Lasso (GLasso) algorithm [11]. Since this breakthrough, several extensions have been proposed [3,12,13,14,15]. For example, Zhang et al. [12] used a Fused Lasso type formulation [16] to extract structural changes in a two-sample situation. In the multi-task learning literature [17], joint estimation algorithms for a set of GGMs with the same topological structures [3,13,14] have been studied based on a group-Lasso [18], while Guo et al. [15] proposed iterative re-weighting of GLasso for estimating multiple GGMs.

Though several GGM learning methods have been proposed, to the best of our knowledge, there are no general techniques for finding a common substructure from multiple datasets[1]. In many practical situations, such as sensor data, the data is highly noisy and the estimated structures tend to have a high variance, which masks the invariance we wish to detect. The scarcity of available data is also a crucial factor in this problem. In our work, we penalized the variation in resulting structures and formulated the common substructure learning problem as an extension of the two approaches presented in Zhang et al. [12] and Honorio et al. [13], respectively. The problem is convex and the solution is obtained by adopting a block coordinate descent procedure [19]. We further show that the solution to the subproblem in the coordinate descent can be classified into three types each of which can be derived efficiently using existing methods. We confirm the validity of our approach through numerical simulations and also in an application with real world datasets derived from the analysis of city-cycle fuel consumption and anomaly detection in car sensors.

[1] Zhang et al. [12] considered a similar problem and although their approach provides a common substructure, it is limited to only two-sample situations. The approach by Chiquet et al. [14] adopted commonness only with respect to its signs.

The remainder of this paper is organized as follows; We first review existing methods for GGM learning and its extensions to joint estimation settings in Section 2. We formulate the common substructure learning problem in Section 3 and then, in Section 4, we present the block coordinate descent algorithm. Section 5 contains numerical simulations to show the validity of the proposed method using synthetic and real world data. Finally, we conclude the paper in Section 6.

2 Structure Learning of Graphical Gaussian Model

In this section, we review the GGM estimation problem [8,9,10,11] and extensions to joint estimation of multiple GGMs [12,13].

2.1 Graphical Gaussian Model

In multivariate analysis, covariance or correlation are commonly used as an indicator of the relationship between two variables. However, in general, the covariance between two variables x_j and $x_{j'}$ is affected by a third variable. Therefore, we need to remove such effects to estimate the essential dependency structure, which is obtained by searching conditional dependency among variables. If a d-dimensional random variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ is Gaussian, the conditional dependency between two variables is expressed by a precision matrix $\Lambda \in \mathbb{R}^{d\times d}$ (or inverse covariance). Under multivariate Gaussian distribution, the following property exists:

$$\Lambda_{jj'} = 0 \;\Leftrightarrow\; x_j \perp\!\!\!\perp x_{j'} \mid \text{other variables} \tag{1}$$

where $\perp\!\!\!\perp$ denotes statistical independence. With this property, GGM is defined as a graph where each node corresponds to a random variable x_j and the adjacency matrix is given by Λ. In a GGM, there is an edge between two nodes only if the corresponding two variables are conditionally dependent. In the case that only a few pairs of variables are dependent, most of the off-diagonal elements in Λ are zero and the corresponding graph expression is sparse, which allows us to visually inspect the underlying relations.

2.2 Sparse Estimation of GGM

The maximum likelihood estimator of a precision matrix is given as the inverse of the sample covariance matrix $\hat{\Sigma}$. This estimator is usually dense and the corresponding GGM is a complete graph, which states that every pair of variables is dependent. The difficulty arises here in that this occurs even when the true precision matrix is sparse and masks the underlying intrinsic relationships. To avoid this unfavorable property, Meinshausen and Bühlmann [8] proposed the use of Lasso for the sparse graph identification, which is later reformulated as a ℓ_1-regularized maximum likelihood problem [9,10]:

$$\begin{aligned} &\max_{\Lambda\in\mathbb{R}^{d\times d}} \; \ell(\Lambda; \hat{\Sigma}) - \rho\|\Lambda\|_1 \\ &\text{subject to } \; \Lambda \succ 0 \end{aligned} \tag{2}$$

where ρ is a regularization parameter and $\ell(\Lambda; \hat{\Sigma})$ is the log-likelihood of a Gaussian distribution defined as

$$\ell(\Lambda; \hat{\Sigma}) = \log\det\Lambda - \mathrm{tr}\left(\hat{\Sigma}\Lambda\right) \ . \tag{3}$$

The constraint is imposed since Λ must be positive definite as a valid precision matrix. The solution to (2) is sparse due to the effect of an additional ℓ_1-regularization term. An efficient algorithm, using a block coordinate descent [11], is available to solve this problem.

2.3 Learning Structural Changes

When comparing two GGMs representing similar models, some common edges may exist whose weights are close to one another. Zhang et al. [12] proposed the use of a Fused Lasso type regularization [16] to round these similar values to exactly the same value, thus allowing only the significant differences between two GGMs to be extracted. Their original idea is based on the work of Meinshausen and Bühlmann [8], which can naturally be transformed to an ℓ_1-regularized maximum likelihood type setting:

$$\begin{aligned} \max_{\Lambda_1,\Lambda_2} \ & \sum_{i=1}^{2} \left\{ \ell(\Lambda_i; \hat{\Sigma}_i) - \rho\|\Lambda_i\|_1 \right\} - \gamma \sum_{j\neq j'} |\Lambda_{1,jj'} - \Lambda_{2,jj'}| \\ & \text{subject to } \ \Lambda_1, \Lambda_2 \succ 0 \end{aligned} \tag{4}$$

where ρ and γ are regularization parameters. The last term forces the difference between certain elements of two matrices to be zero. They also provided an efficient technique for solving the subproblem of (4) which makes the entire procedure fast.

2.4 Multi-task Approach for Learning a Set of GGMs

The ordinary GGM estimation problem (2) aims to learn a single GGM from one dataset. Apart from the GGM estimation, it is known that jointly solving multiple similar tasks often improves the learning performance, which is referred to as multi-task learning [17]. Honorio et al. [13] assumed that all GGMs have the same topological structures, i.e., the same zero patterns in all precision matrices, and adopted the group-Lasso [18] approach, which is formulated as:

$$\begin{aligned} \max_{\{\Lambda_i\}_{i=1}^N} \ & \sum_{i=1}^{N} t_i \ell(\Lambda_i; \hat{\Sigma}_i) - \rho \sum_{j\neq j'} \max_i |\Lambda_{i,jj'}| \\ & \text{subject to } \ \Lambda_1, \Lambda_2, \ldots, \Lambda_N \succ 0 \end{aligned} \tag{5}$$

where ρ is a regularization parameter and $t_1, t_2, \ldots, t_N$ are non-negative constants. The regularization term ensures that the joint structure $\tilde{\Lambda}_{jj'} = \max_i |\Lambda_{i,jj'}|$ is sparse, where $\tilde{\Lambda}_{jj'} = 0$ denotes that the corresponding (j, j')-th entries are commonly zero in all N precision matrices.

3 Common Substructure Learning

The aim of the common substructure learning is to find a dependency structure between variables that is invariant to changes in the surrounding conditions. Formally, we have N covariance matrices $\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_N$ each of which is calculated from datasets sampled under different conditions. The task is to identify common elements shared by all precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$. To begin with, we assume that the number of variables in each condition is the same, i.e., all have d-dimensions. Also, the identities of each variable are the same, e.g., x_1 is always the value from the same sensor while surrounding conditions may change. Then, we define a common substructure of multiple GGMs as follows:

Definition 1 (Common Substructure of Multiple GGMs)
Let $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$ be the corresponding precision matrices of each GGM. Then, their common substructure is expressed by an adjacency matrix Θ defined as

$$\Theta_{jj'} = \begin{cases} \Lambda_{1,jj'} \ , & \textit{if } \Lambda_{1,jj'} = \Lambda_{2,jj'} = \ldots = \Lambda_{N,jj'} \\ 0 \ , & \textit{otherwise} \end{cases} . \tag{6}$$

The common substructure defined here has an edge between nodes only if the corresponding edge weights among all GGMs are equal. We expect to find such a substructure in the estimated precision matrices. To that end, we impose two regularizations and formulate the following problem:

$$\begin{aligned} &\max_{\{\Lambda_i\}_{i=1}^N} \sum_{i=1}^{N} t_i \, \ell(\Lambda_i; \hat{\Sigma}_i) - \sum_{j \neq j'} \left(\rho \max_i |\Lambda_{i,jj'}| + \gamma \max_{i,i'} |\Lambda_{i,jj'} - \Lambda_{i',jj'}| \right) \\ &\quad \text{subject to } \ \Lambda_1, \ \Lambda_2, \ \ldots, \ \Lambda_N \succ 0 \end{aligned} \tag{7}$$

where ρ, γ are regularization parameters and $t_1, t_2, \ldots, t_N$ are non-negative constants that satisfy $\sum_{i=1}^N t_i = 1$. Here, constants t_i are weighting parameters, usually chosen as $t_i = n_i / \sum_{i=1}^N n_i$ where n_i is the size of the i-th dataset. The second regularization term is a generalization of the one in (4) for $N \geq 3$, which ensures that some entries in the resulting precision matrices are common to all matrices. Since the second regularization does not impose any sparsity on the resulting precision matrices, we added the joint regularization term appearing in (5). The resulting common substructure Θ is obtained by applying Definition 1 to the estimated precision matrices $\hat{\Lambda}_1, \hat{\Lambda}_2, \ldots, \hat{\Lambda}_N$.

4 Algorithm

The problem (7) is a concave maximization with convex constraints. In this section, we introduce the solution algorithm based on the block coordinate descent method [19], where the approach is justified by the following theorem.

Theorem 1. *The solution sequence generated by the block coordinate descent for problem (7) is bounded and every cluster point*[2] *is a solution.*

[2] A point where the sequence converges.

4.1 Block Coordinate Descent

In the block coordinate descent, we fix elements in Λ_i corresponding to variables $x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_d$ and update entries related to a variable x_m. Since (7) is invariant for permutations of rows and columns in matrices, we can always arrange x_m-related entries located in the last row and column. Then, we partition each matrix into four parts, namely, one matrix, two vectors, and a scalar:

$$\Lambda_i = \begin{bmatrix} Z_i & \boldsymbol{z}_i \\ \boldsymbol{z}_i^\top & \omega_i \end{bmatrix} \ , \quad \hat{\Sigma}_i = \begin{bmatrix} P_i & \boldsymbol{p}_i \\ \boldsymbol{p}_i^\top & q_i \end{bmatrix} \ . \tag{8}$$

Now, we fix $Z_1, Z_2, \ldots, Z_N$ and derive the subproblem on $\{\boldsymbol{z}_i, \omega_i\}_{i=1}^N$:

$$\max_{\{\boldsymbol{z}_i, \omega_i\}_{i=1}^N} \sum_{i=1}^N t_i \left\{ \log\left(\omega_i - \boldsymbol{z}_i^\top Z_i^{-1} \boldsymbol{z}_i\right) - 2\boldsymbol{p}_i^\top \boldsymbol{z}_i - q_i \omega_i \right\} \\ -2 \sum_j \left(\rho \max_i |z_{ij}| + \gamma \max_{i,i'} |z_{ij} - z_{i'j}| \right) \tag{9}$$

where z_{ij} is the j-th entry of $\boldsymbol{z}_i$. By setting the derivative over ω_i to zero, we get:

$$\omega_i = \boldsymbol{z}_i^\top Z_i^{-1} \boldsymbol{z}_i + q_i^{-1} \ . \tag{10}$$

Here, $Z_i \succ 0$ and $\omega_i - \boldsymbol{z}_i^\top Z_i^{-1} \boldsymbol{z}_i = q_i^{-1} > 0$ guarantee the positive definiteness of Λ_i. Therefore, by choosing the initial Λ_i to be positive definite, that property is always preserved by the updating procedure of the block coordinate descent. Next, by substituting (10) into (9), we derive:

$$\min_{\{\boldsymbol{z}_i\}_{i=1}^N} \sum_{i=1}^N t_i \left(\frac{q_i}{2} \boldsymbol{z}_i^\top Z_i^{-1} \boldsymbol{z}_i + \boldsymbol{p}_i^\top \boldsymbol{z}_i \right) + \sum_j \left(\rho \max_i |z_{ij}| + \gamma \max_{i,i'} |z_{ij} - z_{i'j}| \right) . \tag{11}$$

Instead of solving this problem, we again adopt a coordinate descent approach and further decompose it into subproblems. We solve (11) only for elements related to the variable $x_{m'}$ $(m' \neq m)$ and fix the other entries. As before, we arrange the corresponding elements into the last of the vectors and matrices:

$$\boldsymbol{z}_i = \begin{bmatrix} \boldsymbol{v}_i \\ w_i \end{bmatrix} \ , \quad \boldsymbol{p}_i = \begin{bmatrix} \boldsymbol{r}_i \\ s_i \end{bmatrix} \ , \quad \tilde{Z}_i^{-1} = \begin{bmatrix} H_i & \boldsymbol{h}_i \\ \boldsymbol{h}_i^\top & g_i \end{bmatrix} \ . \tag{12}$$

Then, we derive the following subproblem of (11) over $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)^\top$:

$$\min_{\boldsymbol{w}} \ \frac{1}{2} \boldsymbol{w}^\top \mathrm{diag}(\boldsymbol{a}) \boldsymbol{w} - \boldsymbol{b}^\top \boldsymbol{w} + \rho \|\boldsymbol{w}\|_\infty + \gamma \max_{i,i'} |w_i - w_{i'}| \tag{13}$$

with coefficients $a_i = t_i q_i g_i$ and $b_i = -t_i \left(q_i \boldsymbol{h}_i^\top \boldsymbol{v}_i + s_i \right)$. The dual problem is

$$\min_{\boldsymbol{\xi}} \frac{1}{2} (\boldsymbol{b} - \boldsymbol{\xi})^\top \mathrm{diag}(\boldsymbol{a})^{-1} (\boldsymbol{b} - \boldsymbol{\xi}) \\ \text{subject to } \ |\mathbf{1}_N^\top \boldsymbol{\xi}| \leq \rho, \ \ \|\boldsymbol{\xi}\|_1 \leq \rho + 2\gamma \tag{14}$$

where $\boldsymbol{\xi} = \boldsymbol{b} - \mathrm{diag}(\boldsymbol{a}) \boldsymbol{w}$. This is the subproblem for the block coordinate descent of (7). In the next section, we show that this problem has three types of solutions each which can be derived efficiently.

4.2 Subproblem

First, we can see that subproblem (14) has a solution $\boldsymbol{\xi} = \boldsymbol{b}$ when $|\mathbf{1}_N^\top \boldsymbol{b}| \leq \rho$ and $\|\boldsymbol{b}\|_1 \leq \rho + 2\gamma$. In the case of $|\mathbf{1}_N^\top \boldsymbol{b}| > \rho$ or $\|\boldsymbol{b}\|_1 > \rho + 2\gamma$, the solution is on the boundary of the constraint set and can be classified into three types. Here, we give the solution procedure for each of these. The entire procedure is summarized in Algorithm 1.

1) The solution is on the boundary $\|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$: In this case, we ignore the first constraint in (14) and solve only for the second constraint. Moreover, this problem is shown to be equivalent to the following *continuous quadratic knapsack problem* [13]:

$$\min_{\boldsymbol{y}} \sum_{i=1}^{N} \frac{1}{2a_i} \left(|b_i| - y_i\right)^2 \quad \text{subject to } \ \boldsymbol{y} \geq 0 \,, \ \ \mathbf{1}_N^\top \boldsymbol{y} = \rho + 2\gamma \tag{15}$$

which relates to $\boldsymbol{\xi}$ by $\xi_i = \mathrm{sgn}(b_i) y_i$ where $\mathrm{sgn}(*)$ is a sign function. We give the solution procedure for this problem [13] in Section 4.3. Here, we note that the resulting $\boldsymbol{\xi}$ may violate the constraint $|\mathbf{1}_N^\top \boldsymbol{\xi}| \leq \rho$ since we have ignored it. In this case, we discard the solution and move on to the next case.

2) The solution is on the boundary $|\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho$: This time, we ignore the second constraint in (14) and solve

$$\min_{\boldsymbol{\xi}} \frac{1}{2} (\boldsymbol{b} - \boldsymbol{\xi})^\top \mathrm{diag}(\boldsymbol{a})^{-1} (\boldsymbol{b} - \boldsymbol{\xi}) \ \ \text{subject to } \ |\mathbf{1}_N^\top \boldsymbol{\xi}| \leq \rho \,. \tag{16}$$

This problem has the following single variable Lasso for its dual:

$$\min_{w_0} \ \frac{a}{2} w_0^2 - b w_0 + \rho |w_0| \tag{17}$$

with $a = \sum_{i=1}^N a_i$ and $b = \sum_{i=1}^N b_i$, and the solution is obtained as

$$w_0 = \mathrm{sgn}(b) \frac{(|b| - \rho)_+}{a} \tag{18}$$

where $(*)_+ = \max(*, 0)$ is a soft-thresholding operator. Again, the resulting value $\boldsymbol{\xi} = \boldsymbol{b} - w_0 \boldsymbol{a}$ may violate $\|\boldsymbol{\xi}\|_1 \leq \rho + 2\gamma$. In this case, the solution is on the edge of the intersection of two constraints, and is obtained by the next procedure.

3) The solution is on both boundaries $|\mathbf{1}_N^\top \boldsymbol{\xi}| = \rho$ and $\|\boldsymbol{\xi}\|_1 = \rho + 2\gamma$: Here, we solve (14) with two equality constraints. The procedure in this section is based on the following theorem.

Theorem 2. *Let the solution to (16) be $\tilde{\boldsymbol{\xi}}$. Then, the solution to (14) has the same signs as $\tilde{\boldsymbol{\xi}}$, i.e. $\tilde{\xi}_i \xi_i \geq 0$ for $1 \leq i \leq N$.*

From this result, we can factorize the objective function into the sum of two components $\sum_{\tilde{\xi}_i \geq 0} \frac{1}{2a_i}(b_i - \xi_i)^2$ and $\sum_{\tilde{\xi}_i < 0} \frac{1}{2a_i}(b_i - \xi_i)^2$. The constraint terms can also be expressed as $\sum_{\tilde{\xi}_i \geq 0} \xi_i + \sum_{\tilde{\xi}_i < 0} \xi_i = \rho$ (or $-\rho$) and $\sum_{\tilde{\xi}_i \geq 0} \xi_i - \sum_{\tilde{\xi}_i < 0} \xi_i = \rho + 2\gamma$. As a result, we derive two independent problems:

$$\min_{\boldsymbol{y}^+} \sum_{\tilde{\xi}_i \geq 0} \frac{1}{2a_i} \left(y_i^+ - b_i\right)^2 \quad \text{subject to } \boldsymbol{y}^+ \geq 0 \,, \quad \sum_{\tilde{\xi}_i \geq 0} y_i^+ = \alpha^+ \,, \tag{19}$$

$$\min_{\boldsymbol{y}^-} \sum_{\tilde{\xi}_i < 0} \frac{1}{2a_i} \left(y_i^- + b_i\right)^2 \quad \text{subject to } \boldsymbol{y}^- \geq 0 \,, \quad \sum_{\tilde{\xi}_i < 0} y_i^- = \alpha^- \,. \tag{20}$$

The solutions to these problems relate to $\boldsymbol{\xi}$ in that $\xi_i = y_i^+$ for $\tilde{\xi}_i \geq 0$ and $\xi_i = -y_i^-$ for $\tilde{\xi}_i < 0$. The parameters α^+ and α^- are $\rho+\gamma$ and γ, respectively if the solution is on $\mathbf{1}_N^\top \boldsymbol{\xi} = \rho$, and γ and $\rho+\gamma$, respectively, for $\mathbf{1}_N^\top \boldsymbol{\xi} = -\rho$. These problems are once again continuous quadratic knapsack problems and the solutions can be efficiently obtained by using the algorithm presented in [13]. We can derive the final solution by solving these problems for both cases $\mathbf{1}_N^\top \boldsymbol{\xi} = \rho$ and $\mathbf{1}_N^\top \boldsymbol{\xi} = -\rho$, and choosing the one with the smaller objective function value in (14).

4.3 Continuous Quadratic Knapsack Problem

In this section, we briefly summarize the algorithm for solving the following continuous quadratic knapsack problem presented in [13]:

$$\min_{\boldsymbol{y}} \sum_{i=1}^{N} \frac{1}{2c_i} \left(y_i - d_i\right)^2 \quad \text{subject to } \boldsymbol{y} \geq 0 \,, \quad \mathbf{1}_N^\top \boldsymbol{y} = \alpha \,. \tag{21}$$

Note that this formulation is common to (15), (19) and (20). From the KKT condition, the solution to this problem is given as $y_i(\nu) = \max(d_i - \nu c_i, 0)$ with some constant ν. Moreover, the optimal ν is what satisfies $\mathbf{1}_N^\top \boldsymbol{y}(\nu) = \alpha$. Since $\mathbf{1}_N^\top \boldsymbol{y}(\nu)$ is a decreasing piecewise linear function with breakpoints $\frac{d_i}{c_i}$, we can find a minimum breakpoint $\nu_0 = \frac{d_{i_0}}{c_{i_0}}$ that satisfies $\mathbf{1}_N^\top \boldsymbol{y}(\nu_0) \leq \alpha$ by sorting the N breakpoints. Then, the optimal ν is given as

$$\nu = \frac{\sum_{d_i - \nu_0 c_i \geq 0} d_i - \alpha}{\sum_{d_i - \nu_0 c_i \geq 0} c_i} \,. \tag{22}$$

4.4 Hyper-Parameters ρ and γ

The choice of hyper-parameters ρ and γ affects the resulting graphical models. There are several approaches for choosing these, such as cross validation [9,15] or the Bayesian information criterion [15]. Apart from selection techniques, the following result gives us some insight into ρ and γ, and is helpful for analyzing the data more intensively.

Algorithm 1. Pseudo Code for Common Substructure Learning

Input : sample covariances $\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_N$, regularization parameters ρ, γ
constants $t_1, t_2, \ldots, t_N > 0$, $\sum_{i=1}^N t_i = 1$
Output : precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$

initialize $\Lambda_i \leftarrow \hat{\Sigma}_i^{-1}$ for each $1 \leq i \leq N$;
repeat
  for $x_m : m = 1$ to d **do**
    for $x_{m'} : m' \neq m$ **do**
      if $|\mathbf{1}_N^\top \boldsymbol{b}| \leq \rho$ and $\|\boldsymbol{b}\|_1 \leq \rho + 2\gamma$ **then**
        $\boldsymbol{\xi} \leftarrow \boldsymbol{b}$;
      else
        solve continuous quadratic knapsack problem (15);
        if the solution does not satisfy $|\mathbf{1}_N^\top \boldsymbol{\xi}| \leq \rho$ **then**
          solve (16) with single variable Lasso;
          if the solution does not satisfy $\|\boldsymbol{\xi}\|_1 \leq \rho + 2\gamma$ **then**
            solve (19) and (20) for $[\alpha^+, \alpha^-] = [\rho + \gamma, \gamma]$;
            solve (19) and (20) for $[\alpha^+, \alpha^-] = [\gamma, \rho + \gamma]$;
            adopt one of the two solutions with the smaller value for (14);
          end if
        end if
      end if
      $\boldsymbol{w} \leftarrow \mathrm{diag}(\boldsymbol{a})^{-1}(\boldsymbol{b} - \boldsymbol{\xi})$;
      update (m, m')-th and (m', m)-th elements of Λ_i with w_i for $1 \leq i \leq N$;
    end for
    update (m, m)-th element of Λ_i by (10);
  end for
until $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$ converges

Proposition 1. *In the bivariate case, the off-diagonal elements of the precision matrices λ_i have the following property:*

$$|r_i| \leq \rho + 2\gamma \;\; \textit{for} \;\; 1 \leq i \leq N \;\; \textit{and} \;\; \left| \sum_{i=1}^N t_i r_i \right| \leq \rho \;\; \Rightarrow \;\; \lambda_i = 0 \tag{23}$$

where r_i is the covariance between two variables in the i-th dataset.

Although the result is specific to the bivariate case, we can interpret $\tilde{\gamma} = \rho + 2\gamma$ and ρ as thresholding parameters. If we wish to treat dependencies higher than some level as significant and expect them to be non-zero, $\tilde{\gamma}$ should not exceed that level. We can also see that ρ is the threshold for the average covariance and the parameter that controls the existence of common substructures.

Motivated by this result, we adopt a heuristic approach for the selection of γ. We interpret the parameter 2γ as the difference in characteristic scalings between r_i and $\tilde{r} = \sum_{i=1}^N r_i$. Here, we approximate the distributions of r_i and $\tilde{r} = \sum_{i=1}^N r_i$ with Gaussians and adopt their $1 - \alpha$ levels as their characteristic scalings. Then we set γ to be a half of their difference.

5 Simulation

In this section, we present numerical results of the proposed method both in a synthetic setting and using real world datasets.

5.1 Synthetic Experiment

The aim of this experiment is to evaluate the common substructure detection performance of the proposed method. For the sake of comparison, we adopted GLasso [11] as discussed in Section 2.2 and multi-task structure learning (MSL) [13] from Section 2.4 as baseline methods. Since neither method was designed for common substructure learning, we thresholded the variation in the estimated precision matrices $\hat{\Lambda}_1, \hat{\Lambda}_2, \ldots, \hat{\Lambda}_N$ and heuristically extracted the substructure $\hat{\Theta}$ by

$$\hat{\Theta}_{jj'} = \begin{cases} \hat{\theta}_{jj'} \ , & \text{if } \max_{i,i'} |\hat{\Lambda}_{i,jj'} - \hat{\Lambda}_{i',jj'}| < \epsilon \\ 0 \ , & \text{otherwise} \end{cases} \tag{24}$$

where ϵ is some given threshold for the maximum variation. Here, to avoid selecting zero edges as common substructures, we set $\theta_{jj'}$ to zero if $\hat{\Lambda}_{i,jj'} = 0$ for all i and one otherwise.

We generated sparse precision matrices in the following manner. First, we divided d variables $x_1, x_2, \ldots, x_d$ into non-overlapping subsets for each of the N conditions and generated small precision matrices[3] for each subset. In this step, we set some variable subsets and the corresponding matrices to be common to all N conditions so that the substructure could be shared by all GGMs. Finally, we combined these small matrices by adding some edges between them and derived N precision matrices $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$. In the experiment, we set the dimensionality of the data $d = 20$ and the number of conditions $N = 5$. We selected the size of the variable subsets to be 4 and therefore, the generated GGMs were composed of 5 cliques. The resulting GGM structure is shown in Figure 1.

For the simulation, we generated 100 samples according to the Gaussian distribution with Λ_i in each condition and scaled each variable to have a unit variance. We then compared the common substructure detection rates of the three methods. We repeated the simulations for 100 random realizations of the datasets and drew average ROC curves by varying the hyper-parameter ρ as shown in Figure 3. In this experiment, we chose parameter γ from the procedure presented in Section 4.4 with $\alpha = 0.05$. In Figure 3(a), we set the threshold $\epsilon = 10$ for GLasso and MSL, which means that almost all edges were actually treated as common substructures. The resulting curves clearly show that the proposed method outperforms the two baseline methods. If we set ϵ to a smaller value, e.g. $\epsilon = 1$ in Figure 3(b), the ROC curves for GLasso and MSL are no longer monotone increasing for ρ. Here, we note that $\epsilon = 1$ is already a very

[3] We set the diagonal elements in the matrix to one and the off-diagonals elements to a uniformly random value in $[-0.8, -0.1] \cup [0.1, 0.8]$, although this uniformity might be slightly skewed due to the positive definiteness constraint.

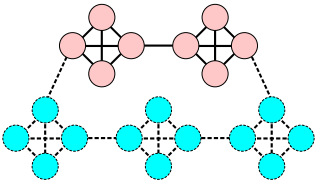


Fig. 1. A GGM structure: edges in the top two cliques (solid lines) are common dependencies, while others are not (dashed lines)

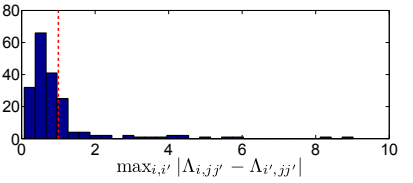
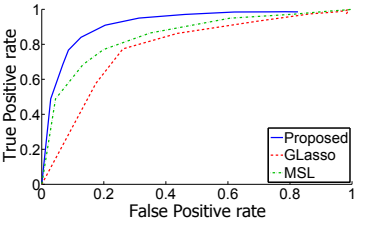
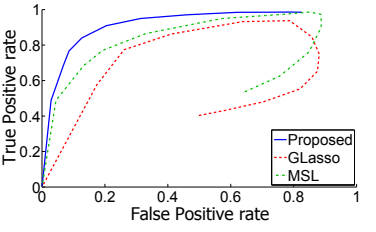


Fig. 2. Histogram of the variation in precision matrices estimated by GLasso with $\rho = 0.0032$. The vertical line denotes the threshold $\epsilon = 1$.



(a) $\epsilon = 10$



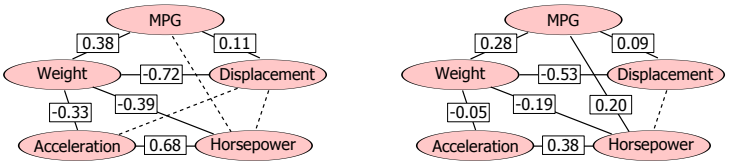
(b) $\epsilon = 1$

Fig. 3. ROC curves. The horizontal axis is the false positive detection rate of common substructures, while the vertical axis is the true positive rate.

optimistic choice. An example of the histogram showing the variation in precision matrices estimated by GLasso with $\rho = 0.0032$ is depicted in Figure 2. In this example, 74% of the estimated non-zero elements have variation less than $\epsilon = 1$ and are judged to be common dependencies. However, only 38% of the true common edges are actually included in the histogram below $\epsilon = 1$, while the other 62% are in the remaining 26% of the estimated non-zero elements. This means that the estimated edge weights using GLasso or MSL for true common substructures vary greatly across the matrices. This example clearly shows the limitation of the existing approaches in that common substructures can easily be masked by estimation variances.

5.2 Analysis of City-Cycle Fuel Consumption Data

We applied the proposed method to the *Auto MPG* dataset from the UCI Machine Learning Repository [20]. The dataset consists of 398 different car data entries containing MPG (Miles Per Gallon), number of cylinders, displacement, horsepower, weight, and acceleration data. Although the name of the car, its model year and the originating region are included in the data, we discarded these fields since they seem to be irrelevant to the other variables. We rearranged the data according to the number of cylinders, giving 199 entries for 4 cylinder cars, 83 for 6 cylinders, and 103 for 8 cylinders. We discarded the data for 3 and 5 cylinders since there were only few entries.



(a) $\rho = 0.1$, $\gamma = 0.062$ (b) $\rho = 0.2$, $\gamma = 0.062$

Fig. 4. Estimated dependency structures for MPG data. The solid lines denote common relations among cars with different numbers of cylinders while the dashed lines are varying dependencies. The numbers attached to solid lines denote the common edge weights.

We applied the proposed method to the 3 datasets containing data for cars with different numbers of cylinders. Each dataset was composed of 5 variables. Empirically, the number of cylinders is closely related to the displacement and the horsepower. The aim of the analysis was to find relations between variables that are irrelevant to the number of cylinders, which might be related to the underlying functional mechanism of cars. As pre-processing, we scaled each variable to have a unit variance.

Figure 4 shows the results for the two settings, $\rho = 0.1$ and 0.2. We chose γ based on the proposed heuristic. In the estimated graph, there are two major cliques composed of *weight, horsepower and acceleration* and *MPG, weight and displacement*, respectively. In the first clique, the relations between mass (weight), acceleration, and force (horsepower) are those expressed by Newton's motion equation. Since each variable has been scaled to unit variance, it is natural that the relation between them is steady. Data fields in the second clique, we believe, they are related to the quality of the car. Typically, expensive cars have many more features including a high specification engine which results in greater weight, higher displacement, and improved MPG. What the results suggest is that this tendency is common to cars with any number of cylinders. We conclude that the proposed method successfully found some reasonable common relations between variables without using any prior knowledge about the datasets.

5.3 Application to Anomaly Detection

In this section, the proposed method is applied to an anomaly detection problem. The task is to identify contributions of each variable to the difference between two datasets. Correlation anomalies [5], or errors on dependencies between variables, is known to be difficult to detect using existing approaches especially with noisy data. To overcome this problem, the use of sparse precision matrices was proposed by Idé et al. [5], since the sparse approach reasonably suppresses the pseudo correlation among variables caused by noise and improves the detection rate. Here, we propose to use the common substructure learning approach. There is a clear indication that the proposed method can further suppress the variation in the estimated matrices. In particular, we expect that dependency structures among healthy variables are estimated to be common, which reduces the risk that such variables are mis-detected and only anomalies are enhanced.

	best AUC	ρ
Proposed	**0.97**	0.05
GLasso	0.96	0.20
MSL	**0.97**	0.05

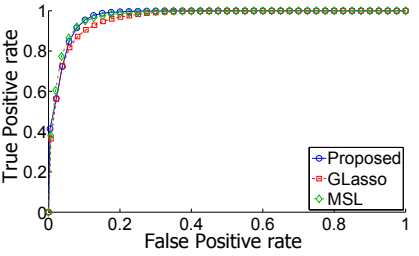
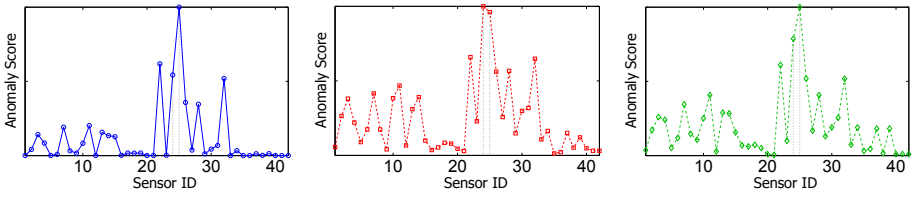


Fig. 5. Anomaly detection : best AUC values and corresponding ROC curves



(a) Proposed Method (b) GLasso (c) MSL

Fig. 6. Anomaly scores. All plots are normalized so that their maximum values are the same. Dotted lines denote true faulty sensors.

We evaluated the anomaly detection performances using the *sensor error* data [5]. The dataset comprised 42 sensor values collected from a real car in 79 normal states and 20 faulty states. The fault was caused by mis-wiring of the 24-th and 25-th sensors, resulting in correlation anomalies. We compared three methods, GLasso, MSL and our proposed method with the anomaly score proposed by Idé et al. [5] which is based on the KL-divergence between two datasets. Since sample covariances are rank deficient in some datasets, we added 10^{-3} on their diagonal to avoid the singularity. For simulation, we randomly sampled 20 datasets from the normal states and 5 datasets from the faulty states, and estimated sparse precision matrices with each method. We set the weight t_i in MSL and the proposed method as $t_i = \frac{1}{40}$ for normal datasets and $t_i = \frac{1}{10}$ for faulty datasets to balance the effects from the two states. Since the anomaly score was designed only for a pair of datasets, we calculated anomaly scores for each of 20×5 pairs and reported the average score and detection rate. We tested each method by varying the parameter ρ between 0.05 and 0.30.

We repeated the above procedure 100 times and drew ROC curves of the average anomaly detection rate with the best area under curve (AUC) results shown in Figure 5. First, we see that MSL and the proposed method surpass the detection rate of GLasso. This is because these two methods estimate precision matrices with joint regularizations. This reduces the estimation variance among matrices while GLasso conducts the estimation separately resulting in more varied estimators, which masks the correlation anomalies. Secondly, though the detection performances are competitive between MSL and the proposed method,

we can see further differences in the resulting anomaly scores in Figure 6. Clearly, the scores for the proposed method show lower significance for normal variables, especially for variables from 16 to 21 and 33 to 42, whereas anomaly variables are still enhanced. This is what we expected in the beginning; that is, the proposed method successfully reduces the nuisance effects and highlights only those variables with correlation anomalies. The remaining peaks at some normal variables are caused by the effect of the two faulty variables, since the correlation anomaly is calculated as faults of a pair of variables.

6 Conclusion

In this paper, we formulated the common substructure learning problem of multiple GGMs and presented an optimization algorithm based on the block coordinate descent. We further showed that the subproblem of the block coordinate descent has three types of solutions and can be solved efficiently with techniques for Lasso and the continuous quadratic knapsack problem. Numerical results on synthetic and real world datasets indicated the clear advantage of the proposed method over existing GGM structure learning methods.

Several future works have been identified: the analysis of the asymptotic property of (7), and the extension of the current formulation to the adaptive Lasso [21] type one to guarantee the *oracle property* [21] of the estimator. Applying the notion of commonness to more general dependency models is also an important work, e.g. non-linear relations or the commonness based on higher order moment statistics.

Acknowledgment. This work was partially supported by a JSPS Grant-in-Aid for Scientific Research(B) #22300054. The authors would like to thank Tsuyoshi Idé and his colleagues for providing the *sensor error* datasets for our simulation. We also acknowledge the many helpful comments from Shohei Shimizu.

References

1. Baillie, R.T., Bollerslev, T.: Common stochastic trends in a system of exchange rates. The Journal of Finance 44(1), 167–181 (1989)
2. Zhang, B., Li, H., Riggins, R.B., Zhan, M., Xuan, J., Zhang, Z., Hoffman, E.P., Clarke, R., Wang, Y.: Differential dependency network analysis to identify condition-specific topological changes in biological networks. Bioinformatics 25(4), 526–532 (2009)
3. Varoquaux, G., Gramfort, A., Poline, J.B., Thirion, B.: Brain covariance selection: better individual functional connectivity models using population prior. Arxiv preprint arXiv:1008.5071 (2010)
4. Ahmed, A., Xing, E.P.: Recovering time-varying networks of dependencies in social and biological studies. Proceedings of the National Academy of Sciences 106(29), 11878–11883 (2009)
5. Idé, T., Lozano, A.C., Abe, N., Liu, Y.: Proximity-based anomaly detection using sparse structure learning. In: Proceedings of the 2009 SIAM International Conference on Data Mining. SIAM, Philadelphia (2009)

6. Lauritzen, S.: Graphical models. Oxford University Press, USA (1996)
7. Dempster, A.P.: Covariance selection. Biometrics 28(1), 157–175 (1972)
8. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34(3), 1436–1462 (2006)
9. Yuan, M., Lin, Y.: Model selection and estimation in the gaussian graphical model. Biometrika 94, 19–35 (2007)
10. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. The Journal of Machine Learning Research 9, 485–516 (2008)
11. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3), 432–441 (2008)
12. Zhang, B., Wang, Y.: Learning structural changes of gaussian graphical models in controlled experiments. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (2010)
13. Honorio, J., Samaras, D.: Multi-task learning of gaussian graphical models. In: Proceedings of the 27th Conference on Machine Learning (2010)
14. Chiquet, J., Grandvalet, Y., Charbonnier, C.: Sparsity with sign-coherent groups of variables via the cooperative-lasso. Arxiv preprint arXiv:1103.2697 (2011)
15. Guo, J., Levina, E., Michailidis, G., Zhu, J.: Joint estimation of multiple graphical models. Biometrika 98(1), 1–15 (2011)
16. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B 67(1), 91–108 (2005)
17. Caruana, R.: Multitask learning. Machine Learning 28(1), 41–75 (1997)
18. Bach, F.R.: Consistency of the group lasso and multiple kernel learning. The Journal of Machine Learning Research 9, 1179–1225 (2008)
19. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications 109(3), 475–494 (2001)
20. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
21. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476), 1418–1429 (2006)

Appendix

Proof of Theorem 1: The non-differentiable term in (7), i.e., the regularization term, is continuous and convex, and is a sum of $\mathcal{O}(d^2)$ terms where each term is composed of variables $\Lambda_{1,jj'}, \Lambda_{2,jj'}, \ldots, \Lambda_{N,jj'}$. Moreover, (7) is continuous in a compact level set. Then, the claim follows from Theorem 4.1 in [19]. □

Proof of Theorem 2: We prove this for the case $\|\tilde{\boldsymbol{\xi}}\|_1 > \rho + 2\gamma$, otherwise $\tilde{\xi}$ is a solution to (14) and the claim holds. Let f be the objective function in (14) and $\boldsymbol{\xi}_0$ be one of the feasible solutions. Then, for $\boldsymbol{\xi}'_0 = \boldsymbol{\xi}_0 + \epsilon(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}_0)$ with $0 < \epsilon \leq 1$, $f(\boldsymbol{\xi}'_0) \leq f(\boldsymbol{\xi}_0)$ holds from the convexity of f. Therefore, $\boldsymbol{\xi}'_0$ is a better solution to problem (14) as long as $|\mathbf{1}_N^\top \boldsymbol{\xi}'_0| \leq \rho$ and $\|\boldsymbol{\xi}'_0\|_1 \leq \rho + 2\gamma$ are satisfied. The first condition always holds because $|\mathbf{1}_N^\top \boldsymbol{\xi}'_0| \leq (1-\epsilon)|\mathbf{1}_N^\top \boldsymbol{\xi}_0| + \epsilon|\mathbf{1}_N^\top \tilde{\boldsymbol{\xi}}| \leq \rho$. On the other hand, the latter condition $\|\boldsymbol{\xi}'_0\|_1 = \sum_{i=1}^N |\xi_{0,i} + \epsilon(\tilde{\xi}_i - \xi_{0,i})| \leq \rho + 2\gamma$ is no longer valid if $\|\boldsymbol{\xi}_0\|_1 = \rho + 2\gamma$ and $\mathrm{sgn}(\xi_{0,i}) = \mathrm{sgn}(\tilde{\xi}_i - \xi_{0,i})$, which results

in $\tilde{\xi}_i \xi_{0,i} \geq 0$. This is the necessary condition for the solution to (14). Otherwise, we can always improve the solution by the above procedure which contradicts its optimality. □

Proof of Proposition 1: Here, we use the alternative expression of (7):

$$\max_{\Theta, \{\Omega_i\}_{i=1}^N} \sum_{i=1}^N t_i \, \ell(\Theta + \Omega_i; \hat{\Sigma}_i) - \sum_{j \neq j'} \left(\rho |\Theta_{jj'}| + \tilde{\gamma} \max_i |\Omega_{i,jj'}| \right)$$
$$\text{subject to } \Theta + \Omega_1, \ \Theta + \Omega_2, \ \ldots, \ \Theta + \Omega_N \succ 0 \tag{25}$$

where $\Lambda_i = \Theta + \Omega_i$ and $\tilde{\gamma} = \rho + 2\gamma$. The equivalence can be proved by comparing their dual problems. In the bivariate case, let matrices Θ, Ω_i and $\hat{\Sigma}_i$ be

$$\Theta = \begin{bmatrix} 0 & \theta \\ \theta & 0 \end{bmatrix} , \quad \Omega_i = \begin{bmatrix} u_i & \omega_i \\ \omega_i & v_i \end{bmatrix} , \quad \hat{\Sigma}_i = \begin{bmatrix} a_i & r_i \\ r_i & b_i \end{bmatrix} .$$

Since $\sum_{i=1}^N t_i |\omega_i| \leq \max_i |\omega_i|$, the objective function (25) is upper-bounded by

$$\mathcal{L}(\theta, \{u_i, v_i, \omega_i\}_{i=1}^N; \{r_i\}_{i=1}^N) = \sum_{i=1}^N t_i \left\{ \log \left(u_i v_i - (\theta + \omega_i)^2 \right) \right.$$
$$\left. -(a_i u_i + b_i v_i) - 2(r_i \omega_i + \tilde{\gamma} |\omega_i|) \right\} - 2 \sum_{i=1}^N t_i r_i \theta - 2\rho |\theta| \, . \tag{26}$$

Moreover, this coincides with (25) if $\omega_i = 0$ for all i. Therefore, if $\omega_i = 0$ $(1 \leq i \leq N)$ is a maximizer of $\mathcal{L}$, it is also the solution to (25). From the derivative of $\mathcal{L}$ over ω_i, we get that $\omega_i = 0$ is a maximizer if

$$-(\tilde{\gamma} + r_i) \leq \frac{\theta}{u_i v_i - \theta^2} \leq (\tilde{\gamma} - r_i) \, . \tag{27}$$

This is a sufficient condition for (25) to have $\omega_i = 0$ $(1 \leq i \leq N)$ as its optimal value. If $|r_i| \leq \tilde{\gamma}$ holds for all i, the problem (25) coincides with the ℓ_1-regularized maximum likelihood (2) with the above constraints on θ:

$$\max_{\theta} \ \log \left(\tilde{u} \tilde{v} - \theta^2 \right) - \left(\tilde{a} \tilde{u} + \tilde{b} \tilde{v} \right) - 2 \left(\tilde{r} \theta + \rho |\theta| \right)$$
$$\text{subject to } \theta \text{ bounded by (27)} \, , \tag{28}$$

where $\tilde{r} = \sum_{i=1}^N t_i r_i$, and $\tilde{u}, \tilde{v}$ are diagonal components of the resulting common structure. Since the bound of θ involves 0, we see that $\theta = 0$ if $|\tilde{r}| \leq \rho$ from Proposition 1 in [5], and hence $\lambda_i = \theta + \omega_i = 0$. □