OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Flexible Structured Graphical LASSO with Latent Variables

Kazuki Koyama KAZUKI.KOYAMA@NTT.COM
Keisuke Kiritoshi K.KIRITOSHI@NTT.COM
Tomomi Okawachi T.OKAWACHI@NTT.COM
Tomonori Izumitani TOMONORI.IZUMITANI@NTT.COM
NTT Communications Corp., Tokyo, Japan

Abstract

The Graphical LASSO method represents sparse inter-variable relationships in the form of a precision matrix. In this study, we propose a new method "Latent Structured Graphical LASSO" that reflect prespecified group structure of variables by introducing a regularization scheme similar to the Latent Group LASSO framework. To represent importance of each group and variable relationship, a group weight and latent variables that are decomposed from each element of a precision matrix are used. We adopt a hierarchical Bayesian model with a prior of multivariate Student's t-distribution for sparsity enhancement and an optimization method based on a variational EM algorithm. We applied the method to two real-world datasets, namely, actual spot rates and credit card fraud, and verified the effectiveness using sparseness, AUC, and correlation anomaly scores. The results indicate that the method can extract sparse relationships between variables considering underlying group structure.

1. Introduction

Knowledge discovery from high-dimensional networks and graphs is an important issue in data mining for many social and science phenomena. In particular, if a system is composed of multiple factors, it is natural for us to assume that there are interdependencies between the variables and to understand the system by extracting these relationships from the data [10]. To extract a sparse Gaussian graphical model, Banerjee et al. [3] and Friedman et al. [7] have proposed the Graphical LASSO, which assumes that the precision matrix of a Gaussian likelihood follows a Laplace distribution introducing sparsity. However, because the Graphical LASSO assumes that each relationship is drawn from the same distribution, this assumption is inappropriate for problems where the relationships correspond to several different classes. In other words, when structures are hidden between relationships, the Graphical LASSO based on a simple $L1$ regularized sparse model often cannot capture behavior well, because an equivalent penalty is added to all relationships [20, 22, 23].

In this paper, we propose a new method for extracting structured-sparsity inherent in a Gaussian graphical model by applying the Latent Group LASSO framework [1, 2, 13, 18]. It is natural to consider that some kind of structured-sparsity exists in a Gaussian graphical model, so it is important for interpretation to express such features with models. Here, structured-sparsity means that we regard one or more relationships as a group allowing duplication, compare the relevance of each group, and infer a sparse precision matrix along the group structure. Although there are various studies that introduce structure to the Graphical LASSO [5, 8, 15, 19, 21], the proposed method differs greatly in that we set up a stochastic model for individual groups and infer the Gaussian graphical model, especially using latent variables. In particular, Tao et al. proposed a method based

on overlapping group norm [19], but the proposed method is more flexible in that it optimizes based on a stochastic model and also tunes the relevance of individual groups.

2. Related Work

2.1. Group LASSO with Latent Variables

The Group LASSO [22] provides sparse solutions for each group along a given discrete structure. It achieves differently structured sparsity with appropriate sparsity-including norms that often correspond to convex relaxations of combinatorial penalties on the support (i.e., non-zero pattern) of the parameter vectors. While most of these norms induce intersection-closed sets of non-zero patterns, Jacob et al. [13] and Bach et al. [1, 2] introduced a different latent formulation of sparsity-inducing norms that yields union-closed sets of non-zero patterns, using the latent variables. In this paper, we denote this method as the Latent Group LASSO.

Let the index set $\mathcal{I} = \{1, \ldots, M\}$ of the model parameters be $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_M]^\top$, and let $\mathcal{G} \subseteq 2^{\mathcal{I}}$ be a discrete structure given in advance. Here, $2^{\mathcal{I}}$ is the power set of $\mathcal{I}$. In the Latent Group LASSO, the parameter vector $\boldsymbol{\omega}$ is represented as a sum of latent vectors $\boldsymbol{\nu}_G$, which are identically zero at indices not in $G \in \mathcal{G}$. Let the weights $\mathcal{W}(G)$ set our prior belief in subset G being relevant where $\mathcal{W} : 2^{\mathcal{I}} \to \mathbb{R}^+$. Here, $\mathbb{R}^+$ is the set of positive real numbers. In particular, a smaller $\mathcal{W}(G)$ means that subset G is more relevant; if G is irrelevant, then $\mathcal{W}(G) = \infty$. The corresponding regularization term of the Latent Group LASSO is then

$$\Omega(\boldsymbol{\omega}) = \sum_{G \in \mathcal{G}} \|\boldsymbol{\nu}_G\|_2 \mathcal{W}(G)^{\frac{1}{2}}. \tag{1}$$

Note that $\boldsymbol{\nu}_G \in \mathbb{R}^M$ is a vector such that all its components with indices in $\mathcal{I} \setminus G$ are zero, and $\boldsymbol{\omega}$ is given by $\boldsymbol{\omega} = \sum_{G \in \mathcal{G}} \boldsymbol{\nu}_G$. Figure 3 in Appendix A shows the difference between the solutions obtained with the Group LASSO and the Latent Group LASSO. The Latent Group LASSO infers not only latent variables $\{\boldsymbol{\nu}_G\}_G$ but also the relevance $\{\mathcal{W}(G)\}_G$ from the data. Since it is not similar to the Group LASSO, we need to introduce a probabilistic model for optimization. The specific optimization method adapted for the proposed method is described in Appendix C.

2.2. Formulation of Graphical LASSO

Let $\mathcal{D} = \{\boldsymbol{x}^{(n)} | \boldsymbol{x}^{(n)} \in \mathbb{R}^M, n = 1, \ldots, N\}$ be the observation data normalized to mean 0 and standard deviation 1. Furthermore, if the data matrix is $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}]^\top$, the sample covariance matrix is given as $\boldsymbol{\Upsilon} \equiv \boldsymbol{X}\boldsymbol{X}^\top / N$. The purpose of the Graphical LASSO [3] is to find a sparse precision matrix $\boldsymbol{\Lambda}$ such that $\Lambda_{ij} \neq 0$ if x_i and x_j have an essential dependency, while $\Lambda_{ij} = 0$ if they are only weakly related due to non-essential factors. Actually, the Graphical LASSO is a convex programming problem, and Friedman et al. [7] proposed an efficient subgradient algorithm to solve this problem. Here, focusing on a specific variable x_i, we appropriately rearrange $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma} \equiv \boldsymbol{\Lambda}^{-1}$, and $\boldsymbol{\Upsilon}$ so that the elements related to x_i are the last row and column and decomposed as follows.

$$\boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{L}_i & \boldsymbol{l}_i \\ \boldsymbol{l}_i^\top & \lambda_i \end{bmatrix}, \begin{bmatrix} \boldsymbol{S}_i & \boldsymbol{s}_i \\ \boldsymbol{s}_i^\top & \sigma_i \end{bmatrix}, \begin{bmatrix} \boldsymbol{U}_i & \boldsymbol{u}_i \\ \boldsymbol{u}_i^\top & \upsilon_i \end{bmatrix} \tag{2}$$

Since $\boldsymbol{\Lambda}$ is a positive definite matrix, its diagonal elements must be positive. Therefore, the optimal solution of the Graphical LASSO satisfies the following two equations.

$$\sigma_i^* = v_i + \rho \tag{3}$$

$$\boldsymbol{\omega}_i^* = \arg\min_{\boldsymbol{\omega}_i} \left\{ \frac{1}{2} \left\| \boldsymbol{S}_i^{-\frac{1}{2}} \boldsymbol{u}_i - \boldsymbol{S}_i^{\frac{1}{2}} \boldsymbol{\omega}_i \right\|^2 + \rho \left\| \boldsymbol{\omega}_i \right\|_1 \right\} \tag{4}$$

where $\boldsymbol{\omega}_i \equiv \boldsymbol{S}_i^{-1} \boldsymbol{s}_i$ and ρ is the regularization coefficient. To obtain the optimal sparse precision matrix $\boldsymbol{\Lambda}^*$, (3) and (4) are repeated for $x_1, \ldots, x_M, x_1, \ldots$ until convergence.

3. Proposed Method

In this section, we formally describe the proposed method, called the "Latent Structured Graphical LASSO," which applies latent structured regularization learning to the Graphical LASSO framework. We introduce a group structure $\mathcal{G}$ into the off-diagonal components of a precision matrix $\boldsymbol{\Lambda}$, obtain the sparse optimal precision matrix $\boldsymbol{\Lambda}^*$ along the group structure, and adopt the Latent Group LASSO framework in the regularization term to achieve it.

Let $\boldsymbol{y}_i \equiv \boldsymbol{S}_i^{-\frac{1}{2}} \boldsymbol{u}_i$ (corresponding to the response vector) and $\boldsymbol{Z}_i \equiv \boldsymbol{S}_i^{\frac{1}{2}}$ (corresponding to the design matrix) in (4) for simplicity of notation. Note that we regard Λ_{ij} and Λ_{ji} as the same parameter since the precision matrix is generally a symmetric matrix. That is, since $\boldsymbol{\Lambda} \in \mathbb{R}^{M \times M}$ for $\mathcal{D} = \{\boldsymbol{x}^{(n)} | \boldsymbol{x}^{(n)} \in \mathbb{R}^M, n = 1, \ldots, N\}$, the number of essential parameters for off-diagonal components in this case is $M(M-1)/2$. We flexibly set the group structure $\mathcal{G}$ in these parameters according to prior knowledge and criteria. Since unnecessary groups are reduced as a result of the optimization, we can arbitrarily set a possible group G in the group structure $\mathcal{G}$ (see Appendix B).

In the proposed method, we set the group structure $\mathcal{G}$ for all off-diagonal components, so we optimize them collectively. For this reason, we combine variables as $\boldsymbol{\omega} = [\boldsymbol{\omega}_1^\top, \ldots, \boldsymbol{\omega}_M^\top]^\top$ and $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_M^\top]^\top$. Moreover, let $\boldsymbol{Z}$ be the block diagonal matrix of $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M$, where they are arranged diagonally in $\boldsymbol{Z}$. This combinations include duplicates due to the symmetry of the precision matrix, but we adopt them for simplicity of notation. Of course, we can obtain an equivalent optimal solution even if we formulate without duplicates. From the preparation so far, we solve the following optimization problem, adding a regularization term based on the Latent Group LASSO.

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega} = \sum_{G \in \mathcal{G}} \boldsymbol{\nu}_G} \left\{ \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\omega} \right\|^2 + \sum_{G \in \mathcal{G}} \left\| \boldsymbol{\nu}_G \right\|_2 \mathcal{W}(G)^{\frac{1}{2}} \right\} \tag{5}$$

We can obtain the sparse optimal off-diagonal components $\boldsymbol{l}^* = [\boldsymbol{l}_1^{*\top}, \ldots, \boldsymbol{l}_M^{*\top}]^\top$ of the precision matrix $\boldsymbol{\Lambda}^*$ according to $\boldsymbol{l}_i^* = -\lambda_i \boldsymbol{\omega}_i^*$ from $\boldsymbol{\omega}^*$ obtained in this way.

For the diagonal components, since we use a different regularization term in (5) from that in (4), we cannot use simple updating rules like (3). Therefore, fixing the off-diagonal components to $\boldsymbol{l}^*$, we obtain them by maximizing the likelihood function for $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_M]^\top$, i.e.,

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} \left\{ \log \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x}^{(n)} \middle| \boldsymbol{0}, \boldsymbol{\Lambda}^{-1}(\boldsymbol{\lambda}, \boldsymbol{l}^*) \right) \right\} \tag{6}$$

where $\boldsymbol{\Lambda}(\boldsymbol{\lambda}, \boldsymbol{l}^*)$ means the precision matrix in which the diagonal components are $\boldsymbol{\lambda}$ and the off-diagonal components are $\boldsymbol{l}^*$. Thus, even if we update the diagonal components, since (6) is the

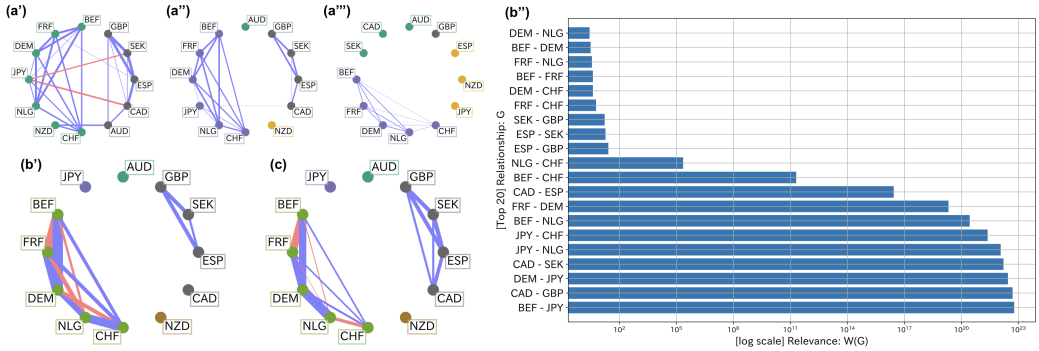

Figure 1: Graph structure and currency clustering calculated from actual spot rates dataset.

maximum likelihood estimation based on observation data $\mathcal{D}$, we have confirmed that the properties that the precision matrix must satisfy such as regularity and positive definiteness are satisfied. In numerical calculations, we extract the final optimal solution by repeating the update with (5) and (6) until $\mathbf{\Lambda}^*$ converges.

We optimize the off-diagonal components with (5) by applying the method proposed by Shervashidze & Bach [18]. They introduce $\beta > 0$ as a hyperparameter that enhances more sparsity. In this paper, we present some results tuned for β and summarize the more detailed optimization flow of the proposed method in Appendix C and D.

4. Experiments

4.1. Actual Spot Rates

This experiment used real data on daily spot prices (foreign currency in dollars), and we considered the effect of changes in input graph structure $\mathcal{G}$ on estimation. Here, the currencies used in this dataset are AUD (Australia), BEF (Belgium), CAD (Canada), CHF (Switzerland), DEM (Germany), ESP (Spain), FRF (France), GBP (United Kingdom), JPY (Japan), NLG (Netherlands), NZD (New Zealand), and SEK (Sweden) [11].

After normalization, we show the results of estimation using the Graphical LASSO, where (a') $\rho = 0.3$, (a'') $\rho = 0.6$, and (a''') $\rho = 0.9$, and the proposed method, where (b') $\beta = 1.5$ and $\mathcal{G}$ is assigned one relationship to one group. Then (c) $\beta = 2.1$ and $\mathcal{G}$ has a group structure based on the relationships between variables as described later, in Figure 1. In each of the figures, the absolute values of the precision matrix represent the thickness of the edge on the same scale, in addition, red means positive values and blue means negative values. Moreover, the colors of each node mean clustering based on the affinity propagation [6]. Throughout, we may infer results such that the continental nations of Western Europe, which include Germany (DEM) and countries that could be called part of the "Franc Economic Zone," have a deeper relationship. However, in the Graphical Lasso, the existence of a connection tends to be ambiguous because all relations are given the same weight. In contrast, as shown in (b') and (b'') showing the top 20 weights in ascending order of $\mathcal{W}(G)$, the proposed method can clearly extract relationships even with actual data, and at the same time, can estimate the degree of irrelevance $\mathcal{W}$ satisfying desirable properties of $\mathcal{W}(G) \to \infty$ corresponding to $\boldsymbol{\nu}_G \to 0$. Due to this, we consider $\mathcal{W}(G)$ to be a useful index for quantitatively evaluating a group structure. By the way, looking at (b'), the color of the node suggests the existence

Table 1: Sparsity and AUC scores for test data.

	MLE	OAS	BS	GL	$\beta = 0.0$	$\beta = 0.1$	$\beta = 0.2$
SPARSITY	0.9149	0.9152	0.9326	0.9175	0.9184	0.9190	**0.9377**
AUC	0.9494	0.9494	0.9502	0.9495	0.9494	0.9494	**0.9505**

of this cluster in the affinity propagation method. Therefore, we regard this as prior knowledge and add a group structure that makes all the connections in {BEF, CHF, DEM, FRF, NLG} and {CAD, ESP, GBP, SEK} to $\mathcal{G}$ used in (b') and induce {CAD, ESP, GBP, SEK} to be easily created. The result of using such an input group structure is (c). Consequently, we can obtain a new graph that considers the effects of clusters {CAD, ESP, GBP, SEK} on the basis of our prior knowledge.

4.2. Credit Card Fraud Detection

In this experiment, we compare the performance of the precision matrices extracted through the anomaly detection task with a credit card fraud detection dataset. Due to confidentiality issues, This dataset contains features V1, V2, ..., V28 that are the result of a principal component analysis (PCA) transformation and features "Time" and "Amount" which have not been transformed [17]. Here, we used 29 features, V1, V2, ..., V28, and "Amount," and then, the test data included 1.0% fraudulent transactions. After normalization, we inferred the precision matrices from the Graphical LASSO (GL), which tuned the regularization coefficients by likelihood cross-validation, and the proposed method for several β. In addition, we define the anomalies d_i for x_i used in this task as $d_i \propto -\log p(x'_i|\boldsymbol{x}'_{-i}, \mathcal{D})$, where $\mathcal{D}$ is the training dataset, and $\boldsymbol{x}'_{-i}$ means a vector obtained by removing the i-th element from test sample $\boldsymbol{x}'$. Moreover, we defined the outlier for each test sample as a 2-norm of the outlier vector, i.e., $\|\boldsymbol{d}\|_2$.

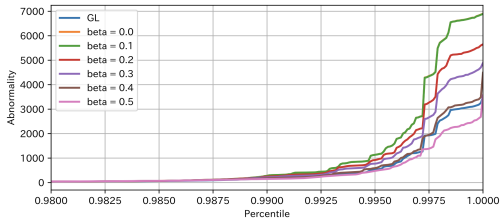

Figure 2: Percentile value of anomalies $\|\boldsymbol{d}\|_2$.

Figure 2 shows the outliers $\|\boldsymbol{d}\|_2$ of the test data for each method plotted against each percentile. From this figure, up to the 99th percentile, the outliers were almost the same for all methods, but then, the outliers of the proposed method increased dramatically compared with the Graphical LASSO (GL). This trend is very reasonable because the test data contained 1% fraudulent transactions.

Furthermore, as shown by Table 1, we calculated the sparsity of a precision matrix and the area under curve of the receiver operating characteristic (AUC) scores. Here, the maximum likelihood estimation (MLE), oracle approximating shrinkage (OAS) and basic shrinkage (BS) were added as comparison methods [4, 14]. Moreover, in this experiment, we used the Gini index proposed by Hurley & Rickard [9] as the most robust measurement for evaluating sparsity. The range of the Gini index is $[0, 1]$, and the higher its value, the sparser a precision matrix. As a result, we believe that the proposed method is highly effective even in tasks that apply precision matrices such as abnormality detection and change point detection while maintaining sufficient sparsity.

5. Conclusion and Future Work

In this paper, we proposed a new regularized method that introduces latent variables and a group structure for all relationships to the Graphical LASSO framework. Using the proposed method, we can incorporate our own interests and prior knowledge into the proposed method and extract a precision matrix along the way from the data. As future work, we are trying to create a more accurate model by optimizing the pre-input group structure $\mathcal{G}$ itself and to extend the method to a mixture normal distribution, as studied for the Graphical LASSO [12].

References

[1] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[2] Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[3] Onureena Banerjee, Laurent El Ghaoui, Alexandre d'Aspremont, and Georges Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In William W. Cohen and Andrew W. Moore, editors, *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 89–96. ACM, 2006.

[4] Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero III. Shrinkage algorithms for mmse covariance estimation. *IEEE Trans. Signal Processing*, 58(10):5016–5029, 2010.

[5] Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B*, 76(2):373–397, 2014.

[6] Brendan J J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 2007.

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[8] Alex Gibberd and J.D.B. Nelson. Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 2015.

[9] Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[10] Tsuyoshi Idé and Hisashi Kashima. Eigenspace-based anomaly detection in computer systems. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 440–449, New York, NY, USA, 2004. Association for Computing Machinery.

[11] Tsuyoshi Idé, Aurelie C. Lozano, Naoki Abe, and Yan Liu. Proximity-based anomaly detection using sparse structure learning. In *SDM*, pages 97–108. SIAM, 2009.

[12] Tsuyoshi Idé, Ankush Khandelwal, and Jayant Kalagnanam. Sparse gaussian markov random field mixtures for anomaly detection. In Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu, editors, *ICDM*, pages 955–960. IEEE Computer Society, 2016.

[13] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *ICML*, volume 382 of *ACM International Conference Proceeding Series*, pages 433–440. ACM, 2009.

[14] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411, 2004. ISSN 0047-259X.

[15] Benjamin M. Marlin and Kevin P. Murphy. Sparse gaussian graphical models with unknown block structure. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *ICML*, volume 382 of *ACM International Conference Proceeding Series*, pages 705–712. ACM, 2009.

[16] Jason A. Palmer, David P. Wipf, Kenneth Kreutz-delgado, and Bhaskar D. Rao. Variational em algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems 18*, pages 1059–1066. MIT Press, 2006.

[17] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *SSCI*, pages 159–166. IEEE, 2015.

[18] Nino Shervashidze and Francis R. Bach. Learning the structure for structured sparsity. *IEEE Trans. Signal Processing*, 63(18):4894–4902, 2015.

[19] Shaozhe Tao, Yifan Sun, and Daniel Boley. Inverse covariance estimation with structured groups. In *IJCAI*, International Joint Conference on Artificial Intelligence, pages 2836–2842, 2017.

[20] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[21] Veronica Tozzo, Federico Tomasi, Margherita Squillario, and Annalisa Barla. Group induced graphical lasso allows for discovery of molecular pathways-pathways interactions. *CoRR*, abs/1811.09673, 2018.

[22] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[23] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

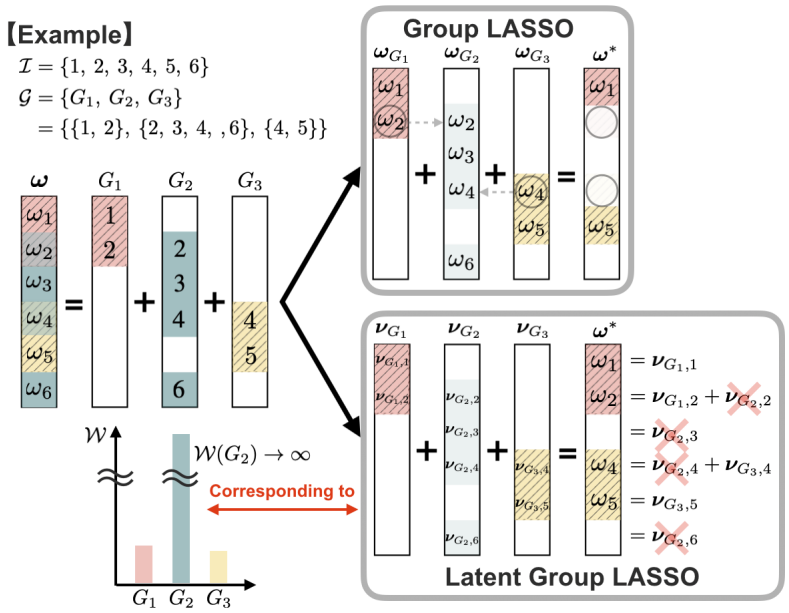

Figure 3: Difference between Group LASSO and Latent Group LASSO.

Appendix A. Group LASSO vs. Latent Group LASSO

Fig. 3 shows the difference between the solutions obtained with Group LASSO and Latent Group LASSO. We consider the group structure of $\mathcal{G} = \{G_1, G_2, G_3\}$ for $\boldsymbol{\omega} = [\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6]^\top$. In Fig. 3, red, green, and yellow respectively represent G_1, G_2, and G_3. As a result of learning, red and yellow groups are selected. As a result of learning with Group LASSO, variables belonging to G_2 among those belonging to G_1 and G_3 are reduced to 0. With Latent Group LASSO, however, variables contained only in G_2 are reduced to 0.

Appendix B. Input Structure

To discover the structure underlying the precision matrix $\boldsymbol{\Lambda}$, we consider what group structure $\mathcal{G}$ should be given in advance. Basically, as a result of optimization, the weight of unnecessary groups is $\mathcal{W}(G) \to \infty$, and the corresponding latent variable is $\boldsymbol{\nu}_G \to 0$, so it is no problem to include all possible groups in $\mathcal{G}$, allowing duplication. If we consider combinatorial explosions, we can set $\mathcal{G}$ on the basis of some criteria. For example, as shown in section 4.1, we may determine $\mathcal{G}$ on the basis of characteristics specific to observation data, such as country or region. Alternatively, if we do not have a priori knowledge, we can use the results obtained by other methods such as the Graphical LASSO in advance. In any case, by setting $\mathcal{G}$ according to our interests, we can compare the relevance of G through $\mathcal{W}(G)$.

Actually, the proposed method also includes a regularization term equivalent to the Graphical LASSO, that is, if $\mathcal{G} = \{\{1\}, \ldots, \{M(M-1)/2\}\}$ and $^\forall\mathcal{W}(G)$ $(G \in \mathcal{G})$ are always fixed to the same value, (5) is equivalent to regularization term (4) of the Graphical LASSO corresponding to $\rho = \mathcal{W}(G)^{\frac{1}{2}}$. Note that there are $M(M-1)/2$ off-diagonal parameters in the M-dimensional data due to the symmetry of the precision matrix. Although there are some errors due to differences

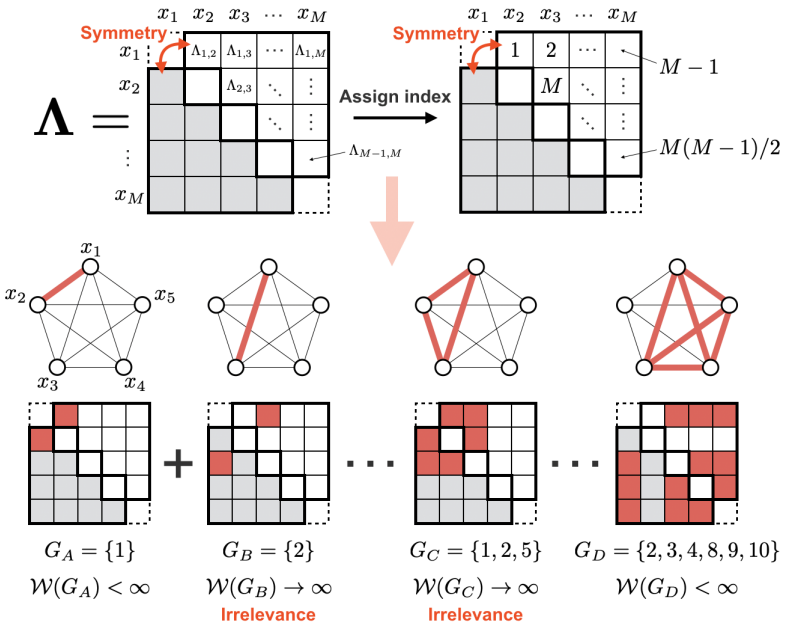

Figure 4: Group structure introduced to off-diagonal components. Note that index assignment we have shown is example, and lower part of this figure shows case where $M = 5$. For such assignments, we can group single connection, such as G_A and G_B, or group multiple connections based on domain knowledge of $\boldsymbol{x}$, such as G_C or G_D. After optimization of proposed method, irrelevant groups become $\mathcal{W}(G) \to \infty$ and therefore corresponding $\boldsymbol{\nu}_G \to 0$. If $\mathcal{W}(G) < \infty$, we can evaluate relevance with $\mathcal{W}(G)$ magnitude.

in optimization methods, we have confirmed in this situation that the proposed method extracts a precision matrix similar to the Graphical LASSO. Therefore, we can implicitly compare the results obtained with the Graphical LASSO by including $\{1\}, \ldots, \{M(M-1)/2\}$ in $\mathcal{G}$ in the proposed method. Figure 4 schematically illustrates the contents of this section.

Appendix C. Optimization Flow

The Latent Group LASSO uses K linear regression problems with design matrices $\boldsymbol{Z}^k$ and response vectors $\boldsymbol{y}^k$ for $k \in \{1, 2, , \ldots, K\}$ obtained by dividing observation data $\mathcal{D}$ into K pieces. For each $\boldsymbol{Z}^k$ and $\boldsymbol{y}^k$, the classical Gaussian linear model with i.i.d. noise of variance σ^2, i.e.,

$$\boldsymbol{y}^k \sim \mathcal{N}(\boldsymbol{Z}^k \boldsymbol{\omega}^k,\ \sigma^2 \boldsymbol{I}), \tag{7}$$

is assumed with the Latent Group LASSO. Here, $\boldsymbol{I}$ is an identity matrix. Moreover, for each k and group structure $\mathcal{G}$, $\boldsymbol{\omega}^k$ is represented as a sum of latent vectors $\{\boldsymbol{\nu}_G^k\}_{G \in \mathcal{G}}$ such as

$$\boldsymbol{\omega}^k = \sum_{G \in \mathcal{G}} \boldsymbol{\nu}_G^k. \tag{8}$$

Therefore, the distribution followed by $\boldsymbol{y}^k$ is equal to

$$\boldsymbol{y}^k \sim \mathcal{N}\left(\boldsymbol{Z}^k \sum_{G\in\mathcal{G}} \boldsymbol{\nu}_G^k,\ \sigma^2 \boldsymbol{I}\right). \tag{9}$$

For the prior distribution of $\{\boldsymbol{\nu}_G^k\}_{G\in\mathcal{G}}$, it is assumed that the following properties are satisfied. First, $\{\boldsymbol{\nu}_G^k\}_{G\in\mathcal{G}}$ are jointly independent. Second, for $^\forall G \in \mathcal{G}$, $\boldsymbol{\nu}_G^k$ has an isotropic density with inverse scale parameter $\mathcal{W}(G)$, i.e.,

$$p(\boldsymbol{\nu}_G^k|\mathcal{W}(G)) = q_G(\|\boldsymbol{\nu}_G^k\|_2 \mathcal{W}(G)^{\frac{1}{2}}) f(G)^{\frac{|G|}{2}}, \tag{10}$$

where q_G is a heavy-tailed distribution that induces a sparse solution and only depends on G through its cardinality $|G|$. Finally, as $\boldsymbol{\nu}_G^k$ are assumed independent,

$$p(\boldsymbol{\omega}^k|\mathcal{W}) = \prod_{G\in\mathcal{G}} p(\boldsymbol{\nu}_G^k|\mathcal{W}(G)). \tag{11}$$

In this statistical model, the log likelihood of parameter vectors is $\sum_{G\in\mathcal{G}} \log q_G(\|\boldsymbol{\nu}_G^k\|_2 \mathcal{W}(G)^{\frac{1}{2}})$, which very closely resembles norm (1). Consequently, maximum a posteriori (MAP) estimation using prior distribution (10) for the latent variable $\boldsymbol{\nu}_G^k$ corresponds to regularization learning using (1). As a result, by maximizing the marginal likelihood for $\{\mathcal{W}(G)\}_{G\in\mathcal{G}}$, i.e.,

$$p(\{\boldsymbol{y}^k\}_k|\mathcal{W}) = \prod_{k=1}^{K} \int p\left(\boldsymbol{y}^k \middle| \boldsymbol{Z}^k\boldsymbol{\omega}^k, \sigma^2\boldsymbol{I}\right) p\left(\boldsymbol{\omega}^k \middle| \mathcal{W}\right) \mathrm{d}\boldsymbol{\omega}^k, \tag{12}$$

we can obtain the optimal solution of $\mathcal{W}(G)$ of a group G in the Latent Group LASSO.

Empirically, when the variance of the prior distribution $p(\boldsymbol{\nu}_G^k|\mathcal{W}(G))$ is smaller than the variance σ^2 of the likelihood, $\mathcal{W}(G)$ may be underestimated. To solve this problem, Shervashidze & Bach [18] introduced $\beta > 0$ as a control parameter of the estimation result and proposed a method of obtaining an appropriate estimation result of $\mathcal{W}(G)$ by tuning β. Specifically, since we usually do not have prior knowledge about $p(\mathcal{W}(G))$, a uniform distribution in $p(\mathcal{W}(G))$ is implicitly assumed with (11). The method using control parameter β makes $\mathcal{W}(G)$ be overestimated as $p(\mathcal{W}(G)) \propto \mathcal{W}(G)^\beta$. In this case, (11) becomes

$$p(\boldsymbol{\omega}^k|\mathcal{W}) = \prod_{G\in\mathcal{G}} p(\boldsymbol{\nu}_G^k|\mathcal{W}(G)) p(\mathcal{W}(G)). \tag{13}$$

After the above optimization procedure, if we want to determine $\boldsymbol{\omega}^*$ uniquely, we can use statistics for $\{\boldsymbol{\omega}^{k*}\}_k$, such as the mean, or the solution obtained by optimizing (5) with the Group LASSO using $\mathcal{W}^*$.

In this paper, we used a multivariate Student's t-distribution as the probability density function of latent variable $\boldsymbol{\nu}_G^k$, i.e.,

$$p(\boldsymbol{\nu}_G^k|\mathcal{W}(G),\ \theta) = \left(\frac{\mathcal{W}(G)}{2\pi}\right)^{\frac{|G|}{2}} \frac{\Gamma\left(\theta + \frac{|G|}{2}\right)}{\Gamma(\theta)} \left(1 + \frac{\|\boldsymbol{\nu}_G^k\|_2^2 \mathcal{W}(G)}{2}\right)^{-\theta-\frac{|G|}{2}},$$

where θ is a parameter governing the shape of the distribution. The smaller θ is, the heavier-tailed the distribution (for $\theta \leq 1$, there is no finite variance). We carried out all experiments with $\theta = 1.5$ and $K = 100$.

Algorithm 1 Latent Structured Graphical LASSO

Input:
Set K-divided normalized data $\{\mathcal{D}^k\}_k$, group structure $\mathcal{G}$, and hyper-parameters of q_G
Calculate $\{\boldsymbol{\Upsilon}^k\}_k$ from $\{\mathcal{D}^k\}_k$
Initialize $\{\boldsymbol{\Lambda}^k\}_k, \{\boldsymbol{\Sigma}^k\}_k \leftarrow \{\boldsymbol{\Upsilon}^{k^{-1}}\}_k, \{\boldsymbol{\Upsilon}^k\}_k$
repeat
  Calculate $\{\boldsymbol{y}^k\}_k$ and $\{\boldsymbol{Z}^k\}_k$
  Optimize hierarchical Bayesian model $p\left(\{\boldsymbol{y}^k\}_k | \mathcal{W}\right)$ and update $\{\boldsymbol{\nu}_G^k\}_{k,G}$, $\{\mathcal{W}(G)\}_G$, and σ
  $\{\boldsymbol{\omega}^k\}_k \leftarrow \{\sum_G \boldsymbol{\nu}_G^k\}_k$
  for $i = 1$ **to** M **do**
    $\{\boldsymbol{l}_i^k\}_k \leftarrow \{-\lambda_i^k \boldsymbol{\omega}_i^k\}_k$
  end for
  $\{\boldsymbol{\lambda}^k\}_k \leftarrow \left\{ \arg\max_{\boldsymbol{\lambda}^k} \mathcal{N}\left(\mathcal{D}^k \middle| \mathbf{0}, \boldsymbol{\Lambda}^{k^{-1}}\left(\boldsymbol{\lambda}^k, \boldsymbol{l}^k\right)\right) \right\}_k$
  Calculate $\{\boldsymbol{\Lambda}^k\}_k$ and $\{\boldsymbol{\Sigma}^k\}_k$ from $\{\boldsymbol{l}^k\}_k$ and $\{\boldsymbol{\lambda}^k\}_k$
  if $\{\boldsymbol{\Lambda}^k\}_k$ converges **then**
    Break this loop
  end if
until

Appendix D. Latent Structured Graphical LASSO Algorithm

We summarize the optimization flow of the proposed method in Algorithm 1. For numerical calculations, a variational EM algorithm using the Palmer et al. method allows us to obtain closed-form updates that optimize the marginal likelihoods (12) [16]. Based on this update, the computational complexity required for one iteration is $\mathcal{O}(P^3)$, which corresponds to the inverse of the $P \times P$ matrix with the number of latent variables as P.