1F4-1

潜在グループ正則化学習におけるグループ構造の自動発見

Automatic Group Structure for Latent Group Regularized Learning

宮澤 桂　河原 吉伸　鷲尾 隆
Miyazawa Kei　Kawahara Yoshinobu　Washio Takashi

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Structurally-regularized learning has recently studied actively in high-dimensional settings because it results in sparse models and it could yield interpretable models than non-structurelly regularized learning methods, such as Lasso. In almost all existing structurally regularized learning, we need prior knowledge about structure in model parameters. In this paper, we develop a probabilistic model for estimating group structures for latent group Lasso, and a greedy optimization algorithm for this problem. We empirically show that group structures can be recovered when the data are generated from the model.

1. はじめに

近年，Group Lasso などに代表される構造正則化学習が注目を集め，よく研究されている．構造正則化学習とは，学習対象の変数間になんらかの離散構造を仮定し，それに沿った正則化項を用いる学習手法である．そのような正則化項を与えることによって，学習解の定性的解釈性を向上したり，モデルの予測精度をよくすることができる．さらに，離散構造に沿ったスパースな解を与えるため，高次元データの解析に有用である．

一方で，構造正則化学習を行うためにはデータ解析者が事前に離散構造を設定する必要があり，解析対象のデータに関する知識を持たない場合に適用することができない．そこで，解析者がデータに関する知識を持たない場合にも，データ中のなんらかのクラスタ構造を学習の中で発見しながら解を導く OSCAR [4] といった手法の研究が行われている．しかし，既存のそういった手法では Group Lasso などで用いられるような複雑な離散構造を与えることはできないため，複雑な離散構造を観測データから学習する手法の研究は重要な課題である．

そこで，本研究では Group Lasso に確率モデルを導入し，そのもとで，離散構造を最尤推定するアルゴリズムを提案する．また，そのアルゴリズムの妥当性をシミュレーション実験によって検証する．

本稿の構成は以下のようである．まず，2. では本研究で推定を行う変数間の離散構造について具体的な定義を与える．また，Group Lasso の概要とその拡張である潜在グループ正則化について説明する．3. では本研究で用いる確率モデルを示し，それに基づいて離散構造に関する尤度関数を定義する．4. では定義した尤度関数を最大化するための貪欲的最大化アルゴリズムを提案する．そして，5. で人工的に生成したデータを用いたシミュレーション実験を行い，6. で結論を述べる．

2. 先行研究

2.1 Group Lasso

Group Lasso は Lasso の拡張として提案された．Group Lasso では Lasso と同様にスパースな解を導くが，このとき，事前に与えられる離散構造に沿ったスパースな解となる．

連絡先: 宮澤桂, 大阪大学大学院工学系研究科, 大阪府箕面市小野原東 1-9-8,kei@ar.sanken.osaka-u.ac.jp

いま，モデルのパラメータ $\boldsymbol{w} = (w_1\ w_2\ \ldots\ w_P)^T$ の索引集合を $\mathcal{S} = \{1, 2, \ldots, P\}$ とし，事前に与えられる離散構造を $\mathcal{A} \subseteq 2^{\mathcal{S}}$ と表す．ここで，$2^{\mathcal{S}}$ は索引集合 $\mathcal{S}$ のべき集合である．このとき，Group Lasso における罰金項は，

$$\Omega(\boldsymbol{w}) = \sum_{A \in \mathcal{A}} ||\boldsymbol{w}_A||_2 \tag{1}$$

で与えられる．ここで，$\boldsymbol{w}_A \in \mathbb{R}^P$ はグループ $A \in \mathcal{A}$ に含まれていない数字を添え字にもつ要素が 0 であるベクトルで，

$$(\boldsymbol{w}_A)_i = \begin{cases} (\boldsymbol{w})_i & (i \in A) \\ 0 & (i \notin A) \end{cases}$$

となる．

以降の議論では，本節で与えた $\mathcal{A}$ のような変数が集合によってグループ分けされる離散構造をグループ構造と呼ぶことにする．

2.2 潜在グループ正則化

2.1 節で紹介した通常の Group Lasso は，与えられたグループ構造に沿ってグループ単位のスパースな解を与える．しかし，Group Lasso によって得られる解は特徴選択によって選ばれなかったグループの補集合となる．したがって，複数のグループに含まれる要素は，単一のグループにしか含まれない要素に比べて解として残りにくい．さらに，グループは解として選択されているにも関わらず，その中のいくつかの要素が解から欠けることは，解釈性の面からみても不都合な場合が多い．そこで，潜在変数 $\{\boldsymbol{v}_A\}_{A \in \mathcal{A}}$ をもちいた次の罰金項が提案されている [2] [3].

$$\Omega(\boldsymbol{w}) = \sum_{A \in \mathcal{A}} ||\boldsymbol{v}_A||_2 f(A)^{\frac{1}{2}} \tag{2}$$

ここで，$\boldsymbol{v}_A \in \mathbb{R}^P$ はグループ $A \in \mathcal{A}$ に含まれていない数字を添え字にもつ要素が 0 のベクトルである．また，$f(A)$ はグループ A の学習における重要度を表し，$f(A)$ の値が小さいほどグループ A は重要となる．そして，この正則化学習ではモデルのパラメータ $\boldsymbol{w} \in \mathbb{R}^P$ が

$$\boldsymbol{w} = \sum_{A \in \mathcal{A}} \boldsymbol{v}_A$$

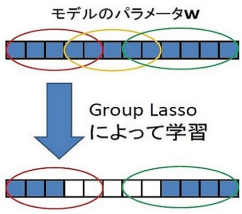

図 1: Group Lasso による解

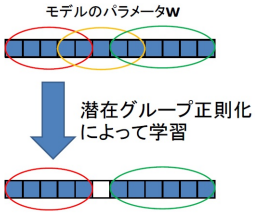

図 2: 潜在グループ正則化による解

で与えられる．この罰金項 (2) は潜在変数を用いてグループ構造を表しているので，潜在グループ正則化と呼ぶことにする．潜在グループ正則化を用いることによって，グループ単位でスパースな解が得られ，かつ，Group Lasso とは異なり，その解は選ばれたグループの結合となる．Group Lasso と潜在グループ正則化によって得られる解の違いを図 1 と図 2 に示す．図 1, 図 2 において，赤，黄，緑の円がそれぞれグループを表し，学習の結果, 赤と緑のグループが選択されている．Group Lasso による学習の結果では赤，緑のそれぞれのグループから黄のグループと共通の要素は消えている．一方，潜在グループ正則化では黄のグループにだけ含まれる要素のみが学習の結果から消える．

3. 確率モデルと問題の定式化

3.1 確率モデル

線形モデルのマルチタスク学習を考える．マルチタスク学習とは, 関連する複数のタスクを同時に学習させることで，これらのタスクに共通の要因を獲得させる手法である．本研究においては，フィッティングパラメータの値は異なるが，グループ構造は同一であるような複数の潜在グループ正則化学習を同時に学習することで，共通の要因であるグループ構造を獲得する．

いま，k 番目のタスクにおいて，出力 $\boldsymbol{y}^k \in \mathbb{R}^{N^k}$ が計画行列 $X^k \in \mathbb{R}^{N^k \times P}$ の線形モデルに分散 σ^2 のガウス雑音が加わったものとして，

$$\boldsymbol{y}^k \sim \mathcal{N}(X^k \boldsymbol{w}^k, \sigma^2 I) \tag{3}$$

で与えられるとする．ここで，$\boldsymbol{w}^k \in \mathbb{R}^P$ は k 番目のタスクにおけるモデルのパラメータ，I は $N^k \times N^k$ の単位行列である．

さらに，パラメータ $\boldsymbol{w}^k$ は潜在的にグループ構造 $\mathcal{A} \subseteq 2^{\mathcal{S}}$ をもち，潜在変数 $\{\boldsymbol{v}_A^k\}_{A\in\mathcal{A}}$ を用いて，

$$\boldsymbol{w}^k = \sum_{A\in\mathcal{A}} \boldsymbol{v}_A^k \tag{4}$$

で与えられるとする．ここで，$\boldsymbol{v}_A \in \mathbb{R}^P$ はグループ $A \in \mathcal{A}$ に含まれていない数字を添え字にもつ要素が 0 のベクトルである．

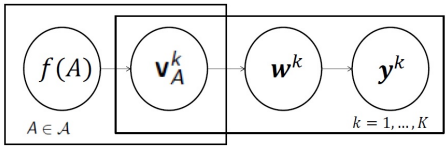

図 3: 確率モデルのグラフィカル表現

る．このとき，$\boldsymbol{y}^k$ が従う分布は

$$\boldsymbol{y}^k \sim \mathcal{N}\Big(X^k \sum_{A\in\mathcal{A}} \boldsymbol{v}_A^k, \sigma^2 I\Big) \tag{5}$$

となる．そして，潜在変数 $\{\boldsymbol{v}_A^k\}_{A\in\mathcal{A}}$ に事前分布

$$p(\boldsymbol{v}_A^k | f(A)) = q_A(||\boldsymbol{v}_A^k||_2 f(A)^{\frac{1}{2}}) f(A)^{\frac{|A|}{2}} \tag{6}$$

を与える. ここで， q_A は A についてその基数 $|A|$ のみに依存する裾の重い分布であり， $\{\boldsymbol{v}_A^k\}_{A\in\mathcal{A}}$ は各グループ A で互いに独立とする．また，$f(A)$ はそれぞれのグループ $A \in \mathcal{A}$ が固有にもつ分布のパラメータである．

以上の確率モデルを図 3 に示す．この確率モデルにおいて，$\{\boldsymbol{v}_A^k\}_{k=1,2,\ldots,K,A\in\mathcal{A}}$ の対数尤度は $\sum_{A\in\mathcal{A}} \log q_A(||\boldsymbol{v}_A^k||_2 f(A)^{\frac{1}{2}})$ で，これは前章 2.2 節で紹介した罰金項 (2) とよく似た式となる．したがって, 潜在変数 $\boldsymbol{v}_A^k$ に事前分布 (6) を用いて MAP 推定することは式 (2) を用いて学習を正則化することに対応する．ゆえに，周辺尤度，

$$\begin{aligned} &p(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^K | X^1, \ldots, X^K, \{f(A)\}_{A\in\mathcal{A}}, \sigma^2 I) \\ &= \prod_{k=1}^K \int p(\boldsymbol{y}^k | X^k, \{\boldsymbol{v}_A^k\}_{A\in\mathcal{A}}, \sigma^2 I) \prod_{A\in\mathcal{A}} p(\boldsymbol{v}_A^k | f(A)) d\boldsymbol{v}_A^k \end{aligned} \tag{7}$$

を $\{f(A)\}_{A\in\mathcal{A}}$ について最大化することによって，前章 2.2 節の潜在グループ正則化におけるグループの重要度 $f(A)$ を推定することができる [1].

以降の節において，この確率モデルにおけるグループ構造 $\mathcal{A}$ を推定することについて説明する．

3.2 問題の定式化

この節では，グループ構造 $\mathcal{A}$ を推定するために $\mathcal{A}$ に対する尤度を導出し，問題を定式化する．

フィッティングパラメータ $\boldsymbol{w}^k \in \mathbb{R}^P$ は潜在変数 $\boldsymbol{v}_A^k \in \mathbb{R}^P$ をもちいて式 (4) で与えられるから，$\boldsymbol{w}^k$ の確率密度関数を $\boldsymbol{v}_A^k$ の確率密度関数から求めることができるとし，$\boldsymbol{w}^k$ の確率密度関数を

$$p(\boldsymbol{w}^k | \mathcal{A}) = p(\boldsymbol{w}^k | \{p(\boldsymbol{v}_A^k | f(A))\}_{A\in\mathcal{A}}) \tag{8}$$

とおく．このとき，出力 $\boldsymbol{y}^k$ の分布がパラメータとしてフィッティングパラメータ $\boldsymbol{w}^k$ をもつと考えると，同時尤度は，

$$\begin{aligned} &p(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^K, \boldsymbol{w}^1, \ldots, \boldsymbol{w}^K | X^1, \ldots, X^K, \mathcal{A}, \sigma^2 I) \\ &= \prod_{k=1}^K p(\boldsymbol{y}^k | X^k, \boldsymbol{w}^k, \sigma^2 I) p(\boldsymbol{w}^k | \mathcal{A}) \\ &= \prod_{k=1}^K p(\boldsymbol{y}^k | X^k, \boldsymbol{w}^k, \sigma^2 I) p(\boldsymbol{w}^k | \{p(\boldsymbol{v}_A^k | f(A))\}_{A\in\mathcal{A}}) \end{aligned} \tag{9}$$

で与えられる．さらに，この式からフィッティングパラメータ $\boldsymbol{w}^k$ を積分消去して，周辺尤度，

$$p(\boldsymbol{y}^1,\ldots,\boldsymbol{y}^K|X^1,\ldots,X^K,\mathcal{A},\sigma^2 I) = \prod_{k=1}^{K}\int p(\boldsymbol{y}^k|X^k,\boldsymbol{w}^k,\sigma^2 I)p(\boldsymbol{w}^k|\mathcal{A})d\boldsymbol{w}^k \quad (10)$$

を得る．よって，本研究の目的であるグループ構造 $\mathcal{A}$ を推定する問題を，

$$\max_{\mathcal{A}\subseteq 2^{\mathcal{S}}}\prod_{k=1}^{K}\int p(\boldsymbol{y}^k|X^k,\boldsymbol{w}^k,\sigma^2 I)p(\boldsymbol{w}^k|\mathcal{A})d\boldsymbol{w}^k \quad (11)$$

と定式化する．ここで，$\mathcal{S}=\{1,\ldots,P\}$ はフィッティングパラメータの索引集合である．

また，本研究ではグループ構造に複数のグループに含まれる要素が存在しないことを仮定する．すなわち，

$$\forall A,B\in\mathcal{A},\ A\neq B\Rightarrow A\cap B=\emptyset \quad (12)$$

の命題が成り立つと仮定する．このとき, フィッティングパラメータ $\boldsymbol{w}^k$ の各要素は対応する単一の潜在変数 $\boldsymbol{v}_A^k$ の値と一致するため，確率密度関数は，

$$p(\boldsymbol{w}^k|\mathcal{A})=\prod_{A\in\mathcal{A}}p(\boldsymbol{v}_A^k|f(A)) \quad (13)$$

となる．

4. 最大化アルゴリズム

この節では，前節 (11) 式として定式化した問題を解く方法について説明する．

グループ構造 $\mathcal{A}$ に対する周辺尤度 $Lh(\mathcal{A})$ は，

$$Lh(\mathcal{A})=\max_{\{f(A)\}_{A\in\mathcal{A}}}\prod_{k=1}^{K}\int p(\boldsymbol{y}^k|X^k,\boldsymbol{w}^k,\sigma^2 I)\prod_{A\in\mathcal{A}}p(\boldsymbol{v}_A^k|f(A))d\boldsymbol{v}_A^k \quad (14)$$

となり，問題 (11) 式は，この値を最大にする $\mathcal{A}$ を求めることを表している．しかし，一般的に，組み合わせに関して式を最適化する問題は効率的にときにくい．もし，総当たり的にすべての組み合わせを試す場合，組み合わせの総数はおよそ 2^{2^P} 通りある．これは，$P=10$ 程度であったとしても現実的な時間で解くことが難しくなる．

そこで，本研究では以下の仮定をおく．まず，モデルの推定したい真のグループ構造を $\mathcal{A}$ とする．そして，グループ構造 $\mathcal{B}$ を $\forall B\in\mathcal{B}\exists A\in\mathcal{A},B\subseteq A$ を満たすグループ構造とする．また，グループ構造 $\mathcal{C}$ を $\forall C\in\mathcal{C}\exists B\in\mathcal{B},C\subseteq B$ を満たすグループ構造とする．このとき，$\mathcal{C}\neq\mathcal{B}$，すなわち，$|\mathcal{C}|>|\mathcal{B}|$ ならば，$Lh(\mathcal{C})<Lh(\mathcal{B})$ である．ただし， $\mathcal{B}$, $\mathcal{C}$ はともに，式 (12) の条件を満たしているとする．この仮定は，真のグループ構造における要素グループの部分集合のみで構成されたグループ構造同士の尤度を比較するとき，より真のグループ構造に近いグループ構造のほうが尤度が大きくなることを意味している．

この仮定をもとに作成した推定手法をアルゴリズム 1 に示す．以下ではこの動作と意図について説明する．

提案手法では，仮定をもとに推定すべき真のグループの部分集合から探索していき，徐々に真のグループに近づけていくことを目的としている．まず，すべての変数が別々のグループに属していると仮定し，探索を開始する．すなわち，

$$\mathcal{A}=\{\{1\},\ldots,\{P\}\}$$

とする．このとき，グループ構造 $\mathcal{A}$ は仮定におけるグループ構造 $\mathcal{B},\mathcal{C}$ と同じ条件をみたす．つぎに，$\mathcal{A}$ から 1 つグループを抜き出し，グループの核とする．この核に対して残りのグループを結合したとき，もし結合した要素が核と同じグループであるならば，尤度は増加する．したがって，アルゴリズム 1 では, 核となるグループを保持する集合として $\mathcal{A}'$ を定義し，$\mathcal{A}$ のグループを $\mathcal{A}'$ に含まれるグループに結合するか，核として別の新しいグループにするかを周辺尤度の比較によって決定している．この操作を $\mathcal{A}$ の要素すべてにたいして繰り返すことによって, 最終的にグループ構造が $\mathcal{A}'$ として推定できる．

Algorithm 1 提案手法の疑似コード

```
A ⇐ {{1},...,{P}}
A' ⇐ ∅
for i = 1 to P do
  NG ⇐ |A'|
  サイズ NG+1 の配列 l[] を用意する
  l[0] ⇐ グループ構造を A ∪ A' として周辺尤度を計算
  A ⇐ A\{i}
  j ⇐ 1
  for all A ∈ A' do
    A' ⇐ A'\A
    l[j] ⇐ グループ構造を A ∪ {A ∪ {i}} ∪ A' として周辺尤度を計算
    A' ⇐ A' ∪ {A}
    j ⇐ j + 1
  end for
  k ⇐ arg max_{j=0,...,NG} l[j]
  if k = 0 then
    A' ⇐ A' ∪ {{i}}
  else
    l[k] に対応する A' の k 番目の要素グループを A_k とする
    A' ⇐ A'\A_k
    A' ⇐ A' ∪ {A_k ∪ {i}}
  end if
end for
```

5. 評価実験

この章では，前章で述べた提案手法を人工データに適用して性能評価を行い，その結果について考察する．実験は 2 つ行う．実験 1 では提案手法によって実際にグループ構造が推定可能なことを確かめる．実験 2 ではタスク数 K の変化に対する推定精度を検証する．

5.1 実験条件の設定

本研究では，潜在変数 $\boldsymbol{v}_A^k$ の確率密度関数として多変量 t 分布をもちいた．すなわち，

$$p(\boldsymbol{v}_A^k|f(A),a)=f(A)^{\frac{|A|}{2}}\frac{\Gamma(a+\frac{|A|}{2})}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{|A|}{2}}\left(1+\frac{||\boldsymbol{v}_A^k||_2^2 f(A)}{2}\right)^{-a-\frac{|A|}{2}} \quad (15)$$

とした．入力値は $a=1.5$，ガウス分布の分散 $\sigma^2=1$ とし，タスク数 K とグループ構造 $\mathcal{A}$ は実験ごとに設定した．また，$\{X^k\}_{k=1,\ldots,K}$ は尤度の計算時間を短縮するために単位行列とした．

5.2 実験 1

実験 1 では，タスク数 $K=1000$，フィッティングパラメータの要素数 $P=10$ とし，グループ構造は，

表 1: 実験 1-1

	$\mathcal{A}$	$\{f(A)\}_{A\in\mathcal{A}}$
生成モデル	$\{\{1,2,3,4\},\{5,6\},\{7,8,9,10\}\}$	$(0.2, 0.5, 1.0)$
提案手法	$\{\{1,2,3,4\},\{5,6\},\{7,8,9,10\}\}$	$(0.17, 0.36, 0.64)$
Active Set Manner	$\{\{1,2,3,4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{1,3,4\},\{1,2,4\},\{2,3,4,5\}\}$	$(0.20, 0.38, 0.40, 0.62, 0.63, 0.65, 0.62, 5.27, 5.65, 7.99)$

表 2: 実験 1-2

	$\mathcal{A}$	$\{f(A)\}_{A\in\mathcal{A}}$
生成モデル	$\{\{1\},\{2\},\{3\},\{4\},\{5,6,7,8,9,10\}\}$	$(0.2, 0.2, 0.2, 0.2, 0.2)$
提案手法	$\{\{1\},\{2\},\{3\},\{4\},\{5,6,7,8,9,10\}\}$	$(0.17, 0.18, 0.15, 0.14, 0.18)$
Active Set Manner	$\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{9\},\{10\},\{7,8\},\{6,8,9\}\}$	$(0.18, 0.19, 0.16, 0.14, 0.19, 0.19, 0.18, 0.22, 0.21, 1.81)$

- $\mathcal{A} = \{\{1,2,3,4\},\{5,6\},\{7,8,9,10\}\}$, $f(A)$ は順に $0.2, 0.5, 1.0$ とする.
- $\mathcal{A} = \{\{1\},\{2\},\{3\},\{4\},\{5,6,7,8,9,10\}\}$, $f(A)$ はすべてのグループで 0.2 とする.

の 2 パターンを試した．また，提案手法の比較対象として，Active Set Manner による貪欲法 [1] も試した．ただし，Active Set Manner による探索では発見されるグループ数が多いので，$\frac{f(A)}{|A|}$ の値の小さいものから 10 グループだけを示す.

5.3 実験 2

実験 2 では，タスク数 K に対する推定精度の検証を行う．そこで，推定精度を測る尺度として，推定されたグループ構造における誤ったグループに含まれる要素の数を用いる．これは，推定されたグループ構造を真のグループ構造に一致するようにグループ間で要素の移動を行ったときに，要素の最小移動回数を数えたものである.

誤り要素数をもちいて以下の実験を行った．まず，グループ構造 $\mathcal{A}$ をグループ数 $|\mathcal{A}| = 3$ となるようにランダムに生成し，人工データをつくる．このとき，グループの重要度 $f(A)$ はすべてのグループで同じ値をもちいた．そして，この人工データに対してグループ構造の推定を行い，誤り要素数を求める．この試行を $K = 10, 20, \ldots, 100, 200, \ldots, 1000$ として，各タスク数 K に対して 20 回ずつ行い，それぞれの誤り要素数の平均を求める.

この実験は，$f(A) = 0.2$ と $f(A) = 20$ の 2 パターンについて行った.

5.4 実験結果

実験 1 の結果を表 1 から表 2 に示す．また，実験 2 の結果を図 4 に示す．図 4 は，行った 2 つのパターンについて，横軸を対数軸としてタスク数 K をとり，縦軸に平均誤り要素数をとっている．図 4 において，それぞれ青曲線は $f(A) = 0.2$, 赤曲線は $f(A) = 20$ のデータに対する曲線である.

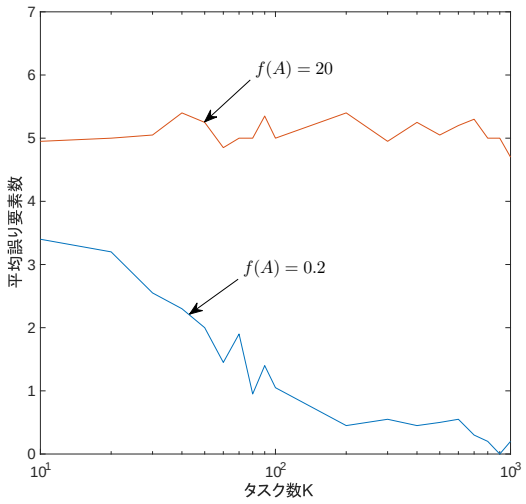

図 4: タスク数 K に対する誤り要素数の変化グラフ．青は $f(A) = 0.2$，赤は $f(A) = 20$.

6. まとめ

本稿では，潜在グループ正則化学習におけるグループ構造の自動発見を，潜在グループ構造の推定として定式化し，最尤推定による特定の条件下での解法を提案した．また，その提案手法によって実際にグループ構造の推定が可能かどうかをシミュレーションによって検証し，その推定精度を評価した．その結果，グループの重要度が高いグループについては提案手法によって推定することができ，タスク数の増加とともに推定精度もよくなることを確認した．ただし，提案手法は特定の条件下でしか使うことができないため，より一般の場合にグループ構造の推定を行うことが今後の課題となる．また，本研究では提案手法を実データに適用していない．今後は，この手法により実データから有益な情報を得る試みが必要である.

参考文献

[1] N.Shervashidze and F.Bach. (2015). Learning to Learn for Structured Sparsity. IEEE Trans. 4894-4902.

[2] L.Jacob, G.Obozinski and J.Vert. (2009). Group Lasso with Overlap and Graph Lasso. Proc. ICML 26, 433-440.

[3] G.Obozinski and F.Bach. (2012). Convex Relaxation for Combinatorial Penalties.

[4] D.Bondell and J.Reich. (2008). Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. The International Biometric Society. Biometrics, Vol. 64,Issue 1, 115-123.