Technical Report of the 1st Workshop on Latent Dynamics
(Jun 16, 2010, Tokyo, Japan)

潜在的グラフ構造からの異常検知*

井手剛†
Tsuyoshi Idé

Abstract: 潜在世界のダイナミックスを異常検知などの実応用につなげる時に問題になるのが、潜在構造の安定性という問題である。もし同定した潜在構造がわずかなデータの揺らぎにより、あるいは、反復アルゴリズムの初期値のような非本質的なパラメターによりがらりとその様相を変えるとしたら、見出された潜在構造が実世界の何らかの反映だと主張することは難しい。本稿では、共分散構造解析の限界を打破したとして一躍有名になった Meinshausen-Bühlmann 理論が、やはり多重共線性の下で困難を持つことを指摘し、それへの対処策について検討する。

1 Introduction

ネットワークやグラフからの知識発見は、データマイニングにおける最近の中心的な課題のひとつである。従来多くの研究は、グラフ構造もしくはそのデータベースを所与とし、それに対して何らかの機械学習的なタスクを行うことに注力きたが、ここ 2-3 年、グラフ構造の学習のための技法が急速に発展している。実用上の要請を考えても、グラフに対する詳細な知識が事前に得られることはむしろまれであり、グラフ構造それ自体をいわば潜在構造として扱い、潜在構造の学習もまた問題の一部であると捉える方が多くの場合自然である。

我々のグループではこれまで、変数間の依存関係が強い状況での、複数のセンサーデータからの異常検出・解析という問題に取り組んできた [9, 8, 12, 11, 10]。実用上、ノイジーなセンサーデータからの異常発見問題においては、以下のような条件が要請されることが多い。

1. データの非定常な変動に対応できること。特に、ノイズによる値のぶれに頑強であること。
2. システムのモジュール構造、もしくは変数のヘテロ性に対応できること。
3. 系全体が異常か否かのみならず、どの変数がどの程度異常かの情報が得られること。

我々は上記のような要請を満たす手法として、グラフィカル・ガウシアン・モデル（GGM）のスパース構造学習に基づく異常検知の手法を最近提案した [10]。我々の貢献のひとつは、上記のような実用上の要請の下で、望ましい構造学習の手法を同定したことである。とりわけ、伝統的な共分散構造解析の限界を打破したとして一躍有名になった Meinshausen-Bühlmann（MB）の算法 [14] が、実際上非常に不安定な結果を与えることを指摘し、Friedman らが提案した [7] のグラフィカル Lasso（以下 gLasso と呼ぶ）と呼ばれる算法が、実用上非常に優れていることを実験的に示した。

本稿では、GGM および MB の算法、gLasso などの理論を比較的詳細に解説し、多重共線性の下での構造不安定性について実験結果を紹介する。

図 1 に我々の問題設定をまとめておこう。ノイジーな多変量のセンサーデータを想定し、複数箇所で窓を取る。問題を簡単にするため、システムの正常稼動時 A と、異常が疑われる状況 B という 2 つの窓でデータを観測したとする。すなわち、2 つのデータセット

$$\mathcal{D}_{\rm A} \equiv \{\boldsymbol{x}_{\rm A}^{(n)} | \boldsymbol{x}_{\rm A}^{(n)} \in \mathbb{R}^M, n = 1, 2, ..., N_{\rm A}\}$$
$$\mathcal{D}_{\rm B} \equiv \{\boldsymbol{x}_{\rm B}^{(n)} | \boldsymbol{x}_{\rm B}^{(n)} \in \mathbb{R}^M, n = 1, 2, ..., N_{\rm B}\}.$$

が与えられたと考える。我々は主にセンサーデータに興味を持つので、インデックス n は典型的には時刻を表す離散値に対応する。$\mathcal{D}_{\rm A}$ および $\mathcal{D}_{\rm B}$ において、測定の回数 $N_{\rm A}$、$N_{\rm B}$ は一般には異なっていてもよい。データセット $\mathcal{D}_{\rm A}$ と $\mathcal{D}_{\rm B}$ が与えられた時、それぞれのデータにおいて変数間の依存関係を表すグラフの相違にどれだけ寄与したかを表す異常度を、各変数について計算せよ、というのが我々の問題である。この問題は統計学における 2 標本検定の問題と似ているが、知りたいのが個々の変数のスコアであるという点で異なる。

†IBM 東京基礎研究所, 242-8502 大和市下鶴間 1623-14（LAB-S7B）, e-mail: goodidea@jp.ibm.com,
IBM Research – Tokyo, 1623-14 Shimo-Tsuruma, Yamato-shi, Kanagawa 242-8502, Japan
Submitted Jun 15, 2010, revised Jun 17, 2010.

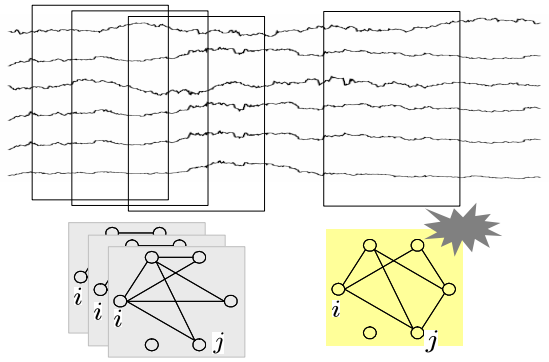

図 1: 問題設定。ノイジーなセンサーデータにおいて、正常稼動時と目される状況と、異常が疑われる状況のそれぞれに対して、(1) その共分散行列に基づいて疎なグラフを学習する。 (2) 次に、その 2 つの疎なグラフを比較して**それぞれの**変数の異常度を求める。

ここで、物理系にはよくあるように、一般に測定系はある冗長性があり、従って、多重共線性は前提と考える必要があることに注意されたい。また、データはノイジーで定常性を持たず、状態空間モデルを用いた時系列モデリング（これはある意味で Latent Dynamics を考える王道である）は簡単ではないことにも注意されたい。今のところ我々はシステムのダイナミックスを明示的に取り入れることはせず、滑走窓の形で系の非定常性を取り込む。

2 グラフィカル・ガウシアン・モデルの構造学習

本節では図 1 におけるステップ 1、すなわち、データからいかに疎なグラフを学習するかについて考える。このステップはデータ A と B に共通なので、以下しばらく両者を区別する添え字を落とし、どちらかを表すデータを、$\mathcal{D} = \{\boldsymbol{x}^{(n)} | n = 1, ..., N\}$ と書くことにする。データ $\mathcal{D}$ における M 個の変数はそれぞれ、平均ゼロ、標準偏差 1 に標準化されていると仮定する。この仮定の下、標本共分散行列 S は

$$\mathsf{S}_{i,j} \equiv \frac{1}{N} \sum_{n=1}^{N} x_i^{(n)} x_j^{(n)} \tag{1}$$

のように与えられる。これはデータの相関係数行列と同じものとなる。

2.1 精度行列と条件付き独立性

グラフィカル・ガウシアン・モデル（GGM）で考えるグラフは、M 個の変数のそれぞれを頂点とするグラフである。一般に、グラフィカル・モデルにおいて、頂点（もしくは変数）x_i と x_j をつなぐ辺が欠けている時、両者は、他のすべての変数を固定した時に条件付き独立である。逆も真である。頂点間の辺の有無を定義するために、GGM では次の M 次元正規分布

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \Lambda^{-1}) = \frac{\det(\Lambda)^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \Lambda \boldsymbol{x}\right) \tag{2}$$

を考える。ここで、det は行列式、$\Lambda \in \mathbb{R}^{M\times M}$ は精度行列を表す。$\mathcal{N}(\cdot|\boldsymbol{\mu}, \Sigma)$ は平均 $\boldsymbol{\mu}$、共分散行列 Σ の正規分布を表す記号である。先に述べたように、精度行列は共分散行列の逆行列である。

正規分布の仮定の下、x_i と x_j をつなぐ辺を欠く条件は下記のように書かれる。

$$\Lambda_{i,j} = 0 \quad \Rightarrow \quad x_i \perp\!\!\!\perp x_j \,|\, \text{other variables} \tag{3}$$

ここで $\perp\!\!\!\perp$ は統計的独立を示す。この条件 (3) は条件付き分布を明示的に書き下すことにより容易に理解することができる。これを以下示す。$(x_i, x_j)^\top$ をまとめて $\boldsymbol{x}_a$ と表し、これら以外の変数をやはりまとめて $\boldsymbol{x}_b$ と表しておく。中心化されたデータに対して、正規分布のよく知られた分割公式（例えば [2] の Sec. 2.3 参照）を用いて、求める条件付き分布は

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a| - \Lambda_{aa}^{-1}\Lambda_{ab}\boldsymbol{x}_b, \Lambda_{aa}^{-1}) \tag{4}$$

のようになる。ここで、$\boldsymbol{x}_a$ と $\boldsymbol{x}_b$ の分割に対応して、

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \tag{5}$$

と置いた。この場合、Λ_{aa} は 2×2 行列に過ぎないから、その逆行列は容易に求められ、(1,2) 成分は $\Lambda_{i,j}$ に比例する。したがって、もし $\Lambda_{i,j} = 0$ ならば、x_i と x_j は、他の変数を条件付けたときに統計的に独立である。

したがって、スパースなグラフを（GGM の範囲で）求めることは、スパースな精度行列を求めることと等価である。

2.2 共分散構造選択

共分散選択 [4] は疎構造学習のための標準的な手法である。簡単に言えばこれは、精度行列においてある小さい行列要素を 0 とおき、その条件を考慮した上で他の行列要素を推定し直す、という過程を繰り返す。しかしながら実用上は、まず標本共分散行列の逆行列を求めねばならないという問題に加え（実データは共分散行列はしばしばランク落ちする）、計算コストが高いこと、統計的検定の観点で必ずしも最適ではないことなどの欠点が知られていた。Drton と Perlman は統計的検定の最適

性の問題を詳しく検討し [5]、SIN と呼ばれる新しいアルゴリズムを提案した。ただしこれは、共分散行列が正則でなければならない要請を取り除いたわけではない。我々は測定系に冗長性があり、それゆえいくつかの変数が強い相関をもつという状況に興味があるので、SIN は我々の問題には有用とは言えない。

2.3 Meinshausen-Bühlmann の方法

MB の方法では [14]、ひとつの変数をターゲットにし、他の変数を入力とした L_1 正則化つきの回帰問題を解く。すなわち、ある変数 x_i に対し、

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\mathsf{Z}_i \boldsymbol{\beta} - \boldsymbol{y}_i||^2 + \mu ||\boldsymbol{\beta}||_1 \right\} \tag{6}$$

を解く。ただし、$\boldsymbol{y}_i \equiv (x_i^{(1)}, ..., x_i^{(N)})^\top$ と定義し、データ行列を

$$\boldsymbol{z}_i^{(n)} \equiv (x_1^{(n)}, .., x_{i-1}^{(n)}, x_{i+1}^{(n)}, ..., x_M^{(n)})^\top \in \mathbb{R}^{M-1} \tag{7}$$

に対して $\mathsf{Z}_i \equiv [\boldsymbol{z}_i^{(1)}, ..., \boldsymbol{z}_i^{(N)}]^\top$ と置いた。これに対応して、精度行列と分布のパラメターとしての共分散行列を、

$$\Lambda = \begin{pmatrix} L & \boldsymbol{l} \\ \boldsymbol{l}^\top & \lambda \end{pmatrix} \quad \Sigma \equiv \Lambda^{-1} = \begin{pmatrix} W & \boldsymbol{w} \\ \boldsymbol{w}^\top & \sigma \end{pmatrix} \tag{8}$$

と分割しておく。ここで行列の行と列は、x_i に関係する要素が最後の行と列と来るように適当に並び替えられているとする。これらの表現において、$W, L \in \mathbb{R}^{(M-1)\times(M-1)}$、$\lambda, \sigma \in \mathbb{R}$、$\boldsymbol{w}, \boldsymbol{l} \in \mathbb{R}^{M-1}$ である。

式 (6) は通常のいわゆる Lasso と同じであり、MB 理論の主張は、各変数に Lasso を解いてまとめれば、統計学的に一致性を持つ構造学習が行える、というものである。より詳しく書けば、まず、式 (6) を解いて、係数 $\boldsymbol{\beta}$ を求める。この係数は、「ターゲット変数」x_i を $\boldsymbol{\beta}^\top \boldsymbol{z}_i$ の形で予測するものであるから、ガウス分布の分割公式 (4) を眺めると、精度行列の対応する 1 列が、

$$\lambda = \frac{1}{\tilde{\sigma_i}^2}, \quad \boldsymbol{l} = -\frac{\boldsymbol{\beta}}{\tilde{\sigma_i}^2},$$

で与えられることが分かる。ただし、$\tilde{\sigma_i}^2$ は予測分散の推定値であり、最尤推定量を使う場合、

$$\tilde{\sigma_i}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_i^{(n)} - \boldsymbol{\beta}^\top \boldsymbol{z}_i^{(n)})^2.$$

のように与えられる。全ての変数について Lasso 回帰の問題を解くことにより精度行列の全要素を求めることができる。

3 グラフィカル Lasso

3.1 ラプラス事前分布による MAP 推定

GGM では、構造学習は多変量正規分布（式 (2)）の精度行列 Λ を求めることに帰着される。まず、疎な構造を得るための工夫は脇に置いて、データ $\mathcal{D}$ からどのように Λ を求めればよいか考えてみよう。最も自然な方法は、次の対数尤度を最大化することである。

$$\ln \prod_{t=1}^{N} \mathcal{N}(\boldsymbol{x}^{(t)} | \boldsymbol{0}, \Lambda^{-1}) = \text{const.} + \frac{N}{2} \{\ln \det(\Lambda) - \mathrm{tr}(\mathsf{S}\Lambda)\}$$

ここで tr は行列の対角和を表す。また、よく知られた恒等式 $\boldsymbol{x}^{(t)\top} \boldsymbol{x}^{(t)} = \mathrm{tr}(\boldsymbol{x}^{(t)} \boldsymbol{x}^{(t)\top})$ と式 (1) を使った。行列の微分に関するよく知られた公式

$$\frac{\partial}{\partial \Lambda} \ln \det(\Lambda) = \Lambda^{-1}, \quad \frac{\partial}{\partial \Lambda} \mathrm{tr}(\mathsf{S}\Lambda) = \mathsf{S} \tag{9}$$

を使えば、直ちに $\Lambda = \mathsf{S}^{-1}$ が最尤解であることが分かる。しかしながら、すでに何度か述べたように、標本共分散行列が正則であることは実用上はまれで、また、仮に正則であったとしても精度行列が疎になるということはほとんどありえない。このため、この解は実用的な価値に乏しい。

われわれも GGM に基づく構造学習を志向するが、上記の限界に基づき、解くべき問題を拡張する。すなわち、式 (2) を、精度行列 Λ が与えられた時の条件付き分布 $p_{\mathrm{G}}(\boldsymbol{x}|\Lambda)$ と見なし、Λ については、事前分布として、要素ごとに同一のラプラス分布を付す。すなわち、

$$p(\Lambda) = \prod_{i,j=1}^{M} \frac{\lambda}{2} \exp\left(-\lambda |\Lambda_{i,j}|\right) \tag{10}$$

である。この式から明らかに分かるように、この事前分布は、Λ の要素の値を 0 付近に束縛する効果を持つ。

そうして GGM の隣接行列 Λ^* を、事後確率最大（MAP: Maximum a posteriori）原理に従って求める。

$$\Lambda^* = \arg\max_{\Lambda} \left\{ \ln p(\Lambda) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}^{(n)} | \boldsymbol{0}, \Lambda) \right\} \tag{11}$$

それゆえ、ただの最尤推定を行うのではなく、次の L_1 制約項付きの最尤方程式を解くことにする。

$$\Lambda^* = \arg\max_{\Lambda} f(\Lambda; \mathsf{S}, \rho), \tag{12}$$

$$f(\Lambda; \mathsf{S}, \rho) \equiv \ln \det \Lambda - \mathrm{tr}(\mathsf{S}\Lambda) - \rho ||\Lambda||_1 \tag{13}$$

ここで $||\Lambda||_1$ は $\sum_{i,j=1}^{M} |\Lambda_{i,j}|$ により定義される。罰金項の重み ρ は入力パラメターとなるが、我々の文脈では、これは異常検知性能を最大化するように決定することができる。

3.2 ブロック勾配法

式 (12) は凸計画問題であり [1]、劣勾配法によって手軽に解くことができる。最近、Friedman、Hastie、および Tibshirani [7] は、グラフィカル Lasso （以下 gLasso と表す）と呼ばれる効率のよい劣勾配アルゴリズムを提案した。gLasso はまず、式 (12) の問題を、ブロック勾配法 [1, 6] という技術を用いて、L_1 制約付き回帰問題の集まりに帰着させる。「ブロック」というのは、上記行列方程式の特定の変数に着目して式変形を行うことに由来する。公式 (9) を用いると、式 (12) の勾配が

$$\frac{\partial f}{\partial \mathsf{\Lambda}} = \mathsf{\Lambda}^{-1} - \mathsf{S} - \rho\,\mathrm{sign}(\mathsf{\Lambda}) \tag{14}$$

と与えられることがわかる。ただし行列 $\mathrm{sign}(\mathsf{\Lambda})$ は、$\mathsf{\Lambda}_{i,j} \neq 0$ に対してはその (i,j) 要素が $\mathrm{sign}(\mathsf{\Lambda}_{i,j})$ で、また、$\mathsf{\Lambda}_{i,j} = 0$ に対しては $\in [-1,1]$ で与えられると定義する。

方程式 $\partial f/\partial \mathsf{\Lambda} = 0$ をブロック勾配法で解くために、ある特定の変数 x_i に着目し、$\mathsf{\Lambda}$ とその逆行列が (8) のように分割されているものとする。この x_i による分割に対応して、標本共分散行列 S も同様に分割するものとし、

$$\mathsf{S} = \begin{pmatrix} S^{\setminus i} & \boldsymbol{s} \\ \boldsymbol{s}^\top & s_{i,i} \end{pmatrix} \tag{15}$$

のように書いておく。

ここで方程式 $\partial f/\partial \mathsf{\Lambda} = 0$ の解を求めよう。$\mathsf{\Lambda}$ は正定値であるため、容易に証明できるように、その対角要素は正でなければならない。したがって、対角要素に関しては、勾配ゼロの条件は

$$\sigma = s_{i,i} + \rho \tag{16}$$

と書かれる。

$\boldsymbol{w}$ および $\boldsymbol{l}$ で表される非対角要素に関しては、他の変数をすべて固定したという条件の下での最適解は、

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||W^{\frac{1}{2}}\boldsymbol{\beta} - \boldsymbol{b}||^2 + \rho\,||\boldsymbol{\beta}||_1 \right\} = 0 \tag{17}$$

を解くことで求められる。ただし、$\boldsymbol{\beta} \equiv W^{-1}\boldsymbol{w}$、$\boldsymbol{b} \equiv W^{-1/2}\boldsymbol{s}$、$||\boldsymbol{\beta}||_1 \equiv \sum_l |\beta_l|$ である。

上式を示そう。分割公式 (8) に基づいて、方程式 $\partial f/\partial \mathsf{\Lambda} = 0$ の右上部分は直ちに

$$\boldsymbol{w} - \boldsymbol{s} - \rho\,\mathrm{sign}(\boldsymbol{l}) = 0 \tag{18}$$

と書かれる。$\mathsf{\Sigma\Lambda} = \mathsf{I}_M$ であるから、

$$\mathsf{\Sigma\Lambda} = \begin{pmatrix} WL + \boldsymbol{w}\boldsymbol{l}^\top & W\boldsymbol{l} + \lambda\boldsymbol{w} \\ \boldsymbol{l}^\top W + \lambda\boldsymbol{w}^\top & \boldsymbol{w}^\top\boldsymbol{l} + \sigma\lambda \end{pmatrix} = \begin{pmatrix} \mathsf{I}_{M-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 1 \end{pmatrix}. \tag{19}$$

を得る。この恒等式の右上部分を使うと、

$$\boldsymbol{l} = -\lambda W^{-1}\boldsymbol{w} = -\lambda\boldsymbol{\beta}, \tag{20}$$

であることが分かる。ただし、$\boldsymbol{\beta} \equiv W^{-1}\boldsymbol{w}$ である。$\mathsf{\Lambda}$ は正定であるから、λ は正でなければならない。従って、$\mathrm{sign}(\boldsymbol{l}) = -\mathrm{sign}(\boldsymbol{\beta})$ が成り立つ。これを用いると、式 (18) は次と等価であることが分かる。

$$\frac{\partial}{\partial\boldsymbol{\beta}} \left\{ \frac{1}{2}\boldsymbol{\beta}^\top W\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{s} + \rho\,||\boldsymbol{\beta}|| \right\} = 0 \tag{21}$$

$W^{-1/2}\boldsymbol{\beta}$ を $\boldsymbol{b}$ とおけば、この式が式 (17) と等価であることが分かる。この最適化問題をどう解くかについては Appendix を参照されたい。

さて、これを解いて $\boldsymbol{\beta}$ を得たとすれば $\mathsf{\Lambda}$ の対応する列を

$$\lambda = \frac{1}{\sigma - \boldsymbol{\beta}^\top W\boldsymbol{\beta}}, \quad \boldsymbol{l} = -\frac{\boldsymbol{\beta}}{\sigma - \boldsymbol{\beta}^\top W\boldsymbol{\beta}} \tag{22}$$

によって更新できる。ただしここで、式 (19) $\boldsymbol{w}^\top\boldsymbol{l} + \sigma\lambda = 1$ の右下部分と、式 (20) を用いた。また、式 (19) の右上部分を用いて、$\boldsymbol{w}$ を

$$\boldsymbol{w} = -W\boldsymbol{l}/\lambda.$$

のように更新することができる。ここで σ は式 (16) のために一定に保たれることに注意。したがって、グラフィカル Lasso アルゴリズムにおいては、$\mathsf{\Sigma} = \mathsf{\Lambda}^{-1}$ は $\mathsf{\Lambda}$ の副産物として得られ、明示的な逆行列の計算は不要である。

最終的な解 $\mathsf{\Lambda}^*$ を得るため、式 (17) を $x_1, x_2, ..., x_M, x_1, ...$ について解くことを収束するまで繰り返す。式 (16) のため、行列 W は必ず正則となることに注意。この点はこの算法の数値的安定性を示唆する。

3.3 MB の方法との関係

gLasso により導かれた座標ごとの最適化問題（式 (17)）には、Lasso に基づく構造学習法 (6) との明らかな類似がみられる。S の定義（式 (1)）を用いれば、もし条件

$$W = S^{\setminus i} \quad \text{and} \quad \rho = M\mu \tag{23}$$

が成り立てば、この問題が式 (17) と等価であることが分かる。W は $\mathsf{\Lambda}^{-1}$ の主対角行列であるので、ρ が小さい時には W と $S^{\setminus i}$ の間に何らかの密接な関係があることが推察されるが、$\rho > 0$ の時は両者は等しくなることはない。

さらに深刻なのは、多重共線性がある時の振る舞いである。Lasso においては、変数に多重共線性がある時、

強く相関した変数のグループのどれかが、例えば実装上の変数の順序付けの違いのようなほとんど偶然の要因で選択される。これは学習された構造が偶然の要因で大幅に変わるということを意味する。

結局、MB の算法は、gLasso と異なり、MAP 最適性のような明確な大局的最適性を持たず、実用上は非常に使いにくい、というのが結論になる。

4 相関異常度のスコアリング

本節では、論文 [10] で与えた相関異常度の定義を要約する。前節で論じた方法に基づいて、二つの疎な GGM $p_{\mathrm{A}}(\boldsymbol{x})$ および $p_{\mathrm{B}}(\boldsymbol{x})$ を得たとしよう。$\mathcal{D}_{\mathrm{A}}$ と $\mathcal{D}_{\mathrm{B}}$ の間の相違に対し、いかに個々の変数が寄与しているかを表すスコアを計算したい。確率モデル $p_{\mathrm{A}}(\boldsymbol{x})$ および $p_{\mathrm{B}}(\boldsymbol{x})$ が与えられている時、最も自然な相違度の尺度は、Kullback-Leibler（KL）距離である。しばらくの間、特定の変数 x_i に着目しよう。量

$$d_i^{\mathrm{AB}} \equiv \int \mathrm{d}\boldsymbol{z}_i \, p_{\mathrm{A}}(\boldsymbol{z}_i) \int \mathrm{d}x_i \, p_{\mathrm{A}}(x_i|\boldsymbol{z}_i) \ln \frac{p_{\mathrm{A}}(x_i|\boldsymbol{z}_i)}{p_{\mathrm{B}}(x_i|\boldsymbol{z}_i)} \quad (24)$$

は $p_{\mathrm{A}}(x_i|\boldsymbol{z}_i)$ と $p_{\mathrm{B}}(x_i|\boldsymbol{z}_i)$ の間の KL 距離の期待値を、分布 $p_{\mathrm{A}}(\boldsymbol{z}_i)$ によって計算したものである。$\boldsymbol{z}_i$ の定義は式 (7) を参照。式 (24) において A と B を入れ替えることで、d_i^{AB} の定義も得る。上式に現れる分布は正規分布のみであるから、この積分は解析的に実行できる。結果は

$$\begin{aligned} d_i^{\mathrm{AB}} &= \boldsymbol{w}_{\mathrm{A}}^{\top}(\boldsymbol{l}_{\mathrm{B}} - \boldsymbol{l}_{\mathrm{A}}) \\ &+ \frac{1}{2}\left\{ \frac{\boldsymbol{l}_{\mathrm{B}}^{\top}\mathsf{W}_{\mathrm{A}}\boldsymbol{l}_{\mathrm{B}}}{\lambda_{\mathrm{B}}} - \frac{\boldsymbol{l}_{\mathrm{A}}^{\top}\mathsf{W}_{\mathrm{A}}\boldsymbol{l}_{\mathrm{A}}}{\lambda_{\mathrm{A}}} \right\} \\ &+ \frac{1}{2}\left\{ \ln\frac{\lambda_{\mathrm{A}}}{\lambda_{\mathrm{B}}} + \sigma_{\mathrm{A}}(\lambda_{\mathrm{B}} - \lambda_{\mathrm{A}}) \right\} \end{aligned} \quad (25)$$

となる。ここで、Λ_{A} およびその逆行列 Σ_{A} をそれぞれ次のように分割した（式 (8) 参照）。

$$\Lambda_{\mathrm{A}} = \begin{pmatrix} L_{\mathrm{A}} & \boldsymbol{l}_{\mathrm{A}} \\ \boldsymbol{l}_{\mathrm{A}}^{\top} & \lambda_{\mathrm{A}} \end{pmatrix} \; \Sigma_{\mathrm{A}} \equiv \Lambda_{\mathrm{A}}^{-1} = \begin{pmatrix} W_{\mathrm{A}} & \boldsymbol{w}_{\mathrm{A}} \\ \boldsymbol{w}_{\mathrm{A}}^{\top} & \sigma_{\mathrm{A}} \end{pmatrix} \quad (26)$$

同様の分割は Λ_{B} および Σ_{B} にも適用される。定義 d_i^{BA} もまた、A と B を入れ替えることで得られる。式 (25) は、よく知られた分割公式 (4) を使えば容易に導出できる。

異常度の定義 (25) の各項は次のような明確な解釈を持つ。GGM の定義から、$\boldsymbol{l}_{\mathrm{A}}$ における非ゼロ要素の数は、頂点 x_i の次数と同じである。この意味で、第 1 項は主に近傍の生成および消滅に関する異常を検知する。第 2 項は、重み付きグラフとしての近傍グラフの「緊密さ」を表している。すなわち、仮に x_i が単一の辺を j に対して持つとすれば、この項は、対応する相関係数の間の差を、単一の変数に対する精度 λ_{A} および λ_{B} で割ったものに比例する。第 3 項は、変数間の関係の変化というよりは各変数ごとの精度もしくは分散の変化に結び付けられる。

最終的な異常度は、A と B の立場を入れ替えたスコアとあわせて、次のように定義するのが自然である。

$$a_i \equiv \max\{d_i^{\mathrm{AB}},\, d_i^{\mathrm{BA}}\} \quad (27)$$

5 実験

この節では、共線形性が強い場合の構造の安定性という切り口で、異なるいくつかの構造学習手法を比較する。

5.1 構造学習手法の比較

相関が強い変数を持つデータに対しては伝統的な共分散構造選択の手法の適用が難しいという事実を考えれば、最近新たに提案された L_1 制約付きの学習手法の安定性を調べてみるのは興味ある研究課題である。我々は gLasso （改めて `Glasso` と表記する）を他の 2 つの構造学習手法と比較した。

最初の比較対象は、Meinshausen と Bühlmann [14] により提案された手法（`Lasso` と表記する）である。彼らの手法は、各変数を目的変数としそれ以外を説明変数とする Lasso 回帰の問題を M 個独立に解くものである。彼らは、この手法がある種の統計的一致性を満たすことを示した。しかし実際上は、過剰に近傍を取り込む傾向があることが知られている [15, 3]。

そこで、もうひとつの比較対象として、適応 Lasso（adaptive lasso）[16] を使う構造学習手法を取り上げる。適応 Lasso（以下 `AdaLasso` と表記する）は 2 段階の線形回帰の手法であり、最初の回帰の結果を 2 度目の回帰の結果に使うことで「オラクル性」という性質を満たすようにする。理論的詳細は原論文 [16] を参照されたい。ここでは、文献 [3] において優れた結果を示した、2 段とも Lasso を用いる手法を使う。

我々はいくつかの変数が強い相関をもつ状況に興味があり、それゆえ S がランク欠損を起こしているというのが前提であるため、S の逆行列の存在を明示的に仮定する伝統的な共分散選択手法とその拡張 [13, 5] は比較の対象としない。

データと評価指標。求められたグラフ構造の安定性を調べる目的で、データにガウスノイズを印加する前後での構造の変化を調べた。用いたデータは第 **??**節で詳しく説明した *Actual spot rates* データである。時間軸を重複がないように 25 個に分け、連続した 100 日を含む小データを作った。そして罰金項の係数をいろいろと変えながら、それぞれの小データに対して何度も構造学習を

行った。その結果に対して、疎度（sparsity）を

$$(\text{疎度}) \equiv \frac{N_0}{M(M-1)}$$

で定義する。ここで、N_0 は Λ の非対角要素におけるゼロ要素の数である。

第 1 回目の構造学習の後、各小データに対して $x_i \leftarrow x_i + \epsilon_i$, のようにガウスノイズを加えた。ただし、$\epsilon_i$ は、平均ゼロの独立同一分布に従うガウスノイズを表す。ノイズの印加により、新たに生ずる辺と、消滅してしまう辺があるので、それらの数を数えて、「辺のフリップ確率」を形式的に

$$(\text{フリップ確率}) \equiv N_1/N_0,$$

で定義する。ここで N_1 は生成または消滅した辺の数である。

結果。図 2 に結果を示す。これは疎度の関数としてフリップ確率を示したものである。ガウスノイズの標準偏差は、平均 0、分散 1 に標準化した後のデータを対象に 0.1 とした。図から、`Lasso` および `AdaLasso` が、極めてノイズに脆弱であることが分かる。これらの方法だと、疎度が 0.5 の時に、フリップ確率は実に 50%にも及ぶ。これは要するに、推定されたグラフが、少なくともノイジーなデータに対しては、ほとんど**まったく信用できない**ことを示す。「真の」グラフ構造を求めることが必要な用途、例えばバイオインフォマティックスにおけるネットワーク推定問題などでは、この点に対して慎重な考察が必要であろう。一方、 `Glasso` はこれらよりははるかにノイズに対し安定である。

`Lasso` および `AdaLasso` の不安定性の大きな理由は、ある程度相関の強い変数がある時、その中のひとつの変数だけが選択されるという Lasso の傾向に帰すことができる。*Actual spot rates* データでは実際、BEF、FRF、DEM、NLG といった欧州通貨は互いに強く相関しあっている。この中のどれが説明変数として選ばれるかはほとんど偶然による。この種の変数節約の傾向は、汎化性能の観点から回帰問題では有用なものであるが、構造学習においては実用上深刻な欠陥と言える。

以上まとめると、変数ごとに独立した回帰問題を解くという `Lasso` および `AdaLasso` の構造学習手法は、データに強い相関を持つ変数群が含まれる場合は、安定した結果を与えない。対照的に `Glasso` は妥当に安定した結果を与える。

6　まとめ

相関異常の検知に対し疎な構造学習を用いるという手法を提案した。我々の問題は、2 つのデータセットの比較に基づいて、個々の変数の異常度を計算するというものであり、この意味で、データセット全体の相違度を求める 2 標本検定の枠組みのひとつの一般化になっている。

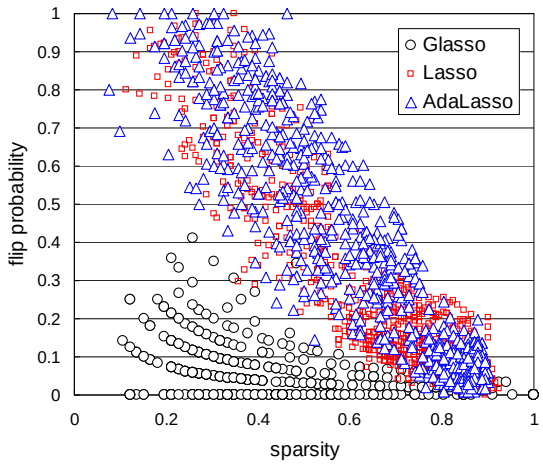

図 2: 疎度の関数としてプロットしたフリップ確率。`Lasso` と `AdaLasso` における著しい不安定性に注目。

我々は、最近提案された疎構造学習の手法のいくつかが共線形性の下で著しく不安定になり、したがって多くの場合、実センサーデータの解析には実用性が乏しいことを指摘した。しかしながら、gLasso アルゴリズムはこの深刻な問題に直面することなく、構造を学習できることを実験的に示した。

参考文献

[1] O. Banerjee, L. E. Ghaoui, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proc. Intl. Conf. Machine Learning*, pages 89–96. Press, 2006.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[3] P. Bühlmann. Variable selection for high-dimensional data: with applications in molecular biology. 2007.

[4] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[5] M. Drton and M. D. Perlman. A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

[6] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

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

[8] T. Idé and K. Inoue. Knowledge discovery from heterogeneous dynamic systems using change-point correlations. In *Proc. SIAM Intl. Conf. Data Mining*, pages 571–575, 2005.

[9] T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. In *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pages 440–449, 2004.

[10] T. Idé, A. C. Lozano, N. Abe, and Y. Liu. Proximity-based anomaly detection using sparse structure learning. In *Proceedings of 2009 SIAM International Conference on Data Mining*, 2009.

[11] T. Idé, S. Papadimitriou, and M. Vlachos. Computing correlation anomaly scores using stochastic nearest neighbors. In *Proc. IEEE Intl. Conf. Data Mining*, pages 523–528, 2007.

[12] T. Idé and K. Tsuda. Change-point detection using krylov subspace learning. In *Proc. 2007 SIAM Intl. Conf. Data Mining*, pages 515–520, 2007.

[13] S. L. Lauritzen. *Graphical Models*. Oxford, 1996.

[14] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[15] R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8(Suppl.2):S3, 2007.

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

A 劣勾配法による Lasso の解法

L_1 正則化項付きの 2 次計画問題 (17) は、変数ごとに劣勾配法を適用することで解くことができる。式 (17) の代わりに、等価な表現 (21) を考えよう。β_i について微分すると、

$$\sum_m W_{i,m}\beta_m - s_i + \rho\,\mathrm{sign}(\beta_i) = 0.$$

を得る。$\beta_i > 0$ に対して、この方程式に対する形式的な解は、

$$\beta_i = \frac{1}{W_{i,i}}(A_i - \rho),$$

で与えられる。ただし、

$$A_i \equiv s_i - \sum_{m\neq i} W_{i,m}\beta_m \tag{28}$$

と定義した。

$W_{i,i} > 0$ であるため、この解は $A_i > \rho$ を満たさなければならない。もしこの条件が満たされなければ、この目的関数の最小は $\beta_i = 0$ において得られる。なぜなら、この場合勾配が正であるからである。同様に、$\beta_i < 0$ の場合を考えると、各 i に対して次のような更新式を得る。

$$\beta_i \leftarrow \begin{cases} (A_i - \rho)/W_{i,i} & \text{for } A_i > \rho \\ 0 & \text{for } -\rho < A_i < \rho \\ (A_i + \rho)/W_{i,i} & \text{for } A_i < -\rho \end{cases}$$

この式を収束するまで繰り返すことで解を得る。