# 疎な相関グラフの学習による相関異常の検出

## Sparse structure learning for correlation anomaly detection

井手 剛*
Tsuyoshi Idé

IBM 東京基礎研究所
IBM Research, Tokyo Research Lab.

**Abstract:** ノイジーなセンサーデータからの異常検知の問題を考えた時、センサー同士の依存関係に現れる異常の検出は、実用上重要かつ困難な問題である。困難の由来はおおむね 2 つにまとめられる。ひとつは、センサー間の相関がノイズに対しきわめて脆弱なため、異常の兆候をノイズから切り分けるのが難しい点である。2 つ目は、複数の変数ペアにおいて何かの異常が観測されたとしても、その情報を個々のセンサーの異常度に帰着させるのが簡単でない点である。本論文では、前者への解決策として疎な構造学習の手法を用いることを、また後者に対しては、グラフィカル・ガウシアン・モデルから情報論的に自然に導かれる相関異常スコアを用いることを提案する。

## 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。現在のところ多くの研究は、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している。しかし一般には、グラフに対する詳細な知識が事前に得られることはまれであり、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自然な場合も多い [15]。また、共著関係のような、接続しているか否かの 2 値で表されるグラフより、何らかの実数値の重みを持った重み付きグラフを考える方が自然な場合も多い。

そのような代表例のひとつとして、本論文では実数値の多次元データ、特に時系列データを考える。ノイジーなセンサーデータの変数の間に存在しうる潜在構造としてのグラフを想像してみると、ノイズが存在する以上、完全に独立な変数というのは現実的にはありえず、必然的に、その依存関係を表す最も一般的なグラフは完全グラフとなってしまう。したがって、実数値データでは、ノイズによる見かけ上の依存関係を取り除いた、**疎な**グラフをデータから見出すというステップが必要である。

本論文では、次のような問題設定の下、グラフに基づく異常検知のタスクを考える [17]： ある多変量系の各変数の依存関係の強弱を表す完全グラフが 2 つ与えられた時、両者の相違に最も寄与する頂点を特定せよ。また、その寄与度を計算せよ。多変量系全体としての相違のみならず、どのように相違しているかを個々のノードの異常度を使って表そうとしている点に注意されたい。

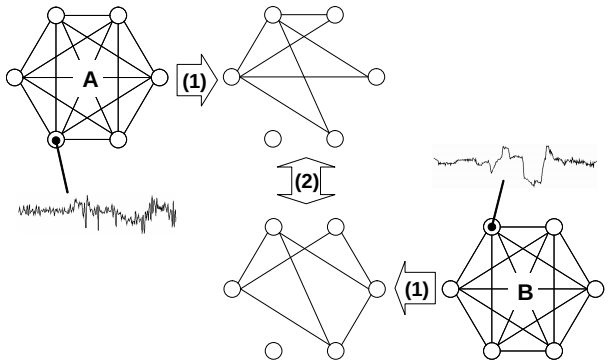


図 1: 問題設定。ノイジーなセンサーデータ A と B が与えられた時、(1) その共分散行列に基づいて疎なグラフを学習する。 (2) 次に、その 2 つの疎なグラフを比較して**それぞれの**変数の異常度を求める。

図 1 に我々の問題設定を要約する。我々の目標は、典型的にはノイジーなセンサーデータから推定された依存グラフの依存関係に生じうる異常（これを**相関異常**と呼ぶ）を、各変数の異常度として定量化することである。主として相関異常に着目する理由は、第 1 に、各変数個別に出る異常の検出にはそれなりに既存手法もあるが、相関異常の方は使える既存手法がほとんどないためである。しかも人手での検出に深い勘と経験を必要とするため、これを半自動化する実際上のメリットが大きい。第 2 に、相関異常は、コンピュータシス

*連絡先： IBM 東京基礎研究所
242-8502 神奈川県神奈川下鶴間 1623-14 （LAB-S7B）
E-mail: goodidea@jp.ibm.com

テム [15] や自動車などの複雑な機械系 [17] に広く見出されるためである。ノイズの影響を大きく受けているかもしれない密な依存グラフから出発して、我々はまず、ノイズの影響を除去した疎な構造を見出すことを試みる。次いで、見出された構造を比較して、各変数の異常度を計算する。

統計学の分野では、実数値多変量データからの構造学習という問題は、Dempster [6] を嚆矢とする共分散構造選択理論の枠組みで扱われてきた。その名の示すようにこの理論は、多変量正規分布を仮定して、疎な構造を実現するためのある制約の下に、データに対し精度行列（共分散行列の逆行列）を反復的に当てはめる。最近、伝統的な共分散構造選択法の欠点を解決する目的で、新たな疎構造学習の手法がいくつか提案されている [20, 21, 5, 3, 7, 9]。中でも、Meinshausen と Bühlmann [21] は、疎構造学習というタスクを各変数の近傍選択の問題として捉え、それを、ひとつの変数を目的変数とし残りを説明変数とする Lasso（$L_1$ 制約付き線形回帰）の問題として定式化した。$L_1$ 正規化により、多くの線形結合の係数は厳密にゼロとなり、非ゼロの係数を持つ説明変数は目的変数の近傍と見なされる。実用的な見地からは、彼らの方法の最大のメリットのひとつは、変数の数（$M$）が、標本数（$N$）よりも大きいような状況でも形式的には疎構造学習が可能であるという点である。この点は伝統的な共分散選択理論と決定的に違う。これまでの共分散選択手法では、共分散行列がランク欠損を起こしているような状況では意味のある結果を出すのが難しかった。$M > N$ であったり、あるいはいくつかの変数が強く相関しているような時、共分散行列はランク落ちするから、この欠点は実用的には深刻なものであった。

この論文では、相関異常をスコアリングするというタスクに対し、$L_1$ 制約付きの疎構造学習手法を用いることを初めて提案する。一般のセンサーデータ解析でよくあるように、データには相関の強い変数対がいくつかあると仮定する。我々は Meinshausen-Bühlmann の方法 [21] がそのような場合に不安定化すること、そしてグラフィカル Lasso と呼ばれる方法 [9] がノイズに頑強で、したがって我々の用途に適することを実験的に示す。我々はまた、グラフィカル・ガウシアン・モデルに基づいて、相関異常度の定義を情報理論的に自然な方法で導く。

この論文の構成は以下のとおりである。第 2 節は、具体例を交えつつ我々の問題設定を述べ、グラフィカル・ガウシアン・モデルの本質的な部分を解説する。第 3 節は、統計学における 2 標本検定との比較を視野に入れつつ、関連研究を概観する。第 4 節は、疎構造学習の方法について議論する。第 5 節は我々の相関異常度の定義について述べる。続く第 6 節において実データを用いた実験結果を説明する。最後に、第 7 節において本論文の要約を行う。

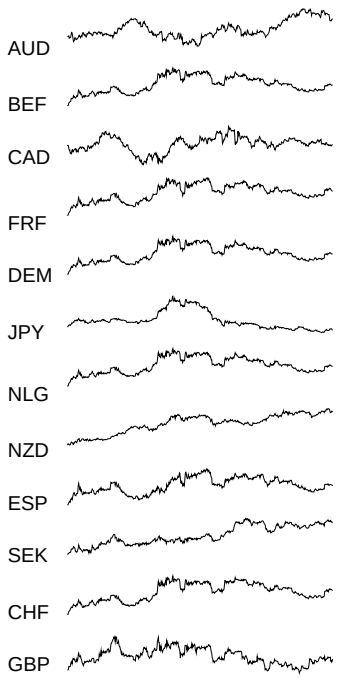


図 2: *Actual spot rates* データ。1996 年 8 月 9 日までの 567 日間を示したもの。

表 1: *Actual spot rates* データにおける略号一覧。

| | | | |
|---|---|---|---|
| AUD | Australian Dollar | NLG | Dutch Guilder |
| BEF | Belgian Franc | NZD | New Zealand Dollar |
| CAD | Canadian Dollar | ESP | Spanish Peseta |
| FRF | French Franc | SEK | Swedish Krone |
| DEM | German Mark | CHF | Swiss Franc |
| JPY | Japanese Yen | GBP | UK Pound |

# 2 準備

この節では、グラフィカル･ガウシアン･モデル（GGM）の要点をまとめ、我々の問題設定をやや形式的に述べる。

## 2.1 問題設定

物理的なセンサーを介して採取されるデータや経済時系列データにおいて、複数の変数が強い相関を持つことは珍しいことではない。図 2 はそのような例であり、これは、対ドルの通貨レートの日ごとのスポット値を示している[1]。図中の略号は表 1 に示してある。当然予想されるように、BEF、FRF、DEM、NLG などの

[1]UCR アーカイブ [18] の *Actual spot rates* データの最後の 567 日分を取ったものである。元データは 1986 年 10 月 9 日から 1996 年 9 月 8 日までの 10 年にわたる。

欧州通貨は強い相関を持つことが分かる。実際、最初の 4 つの通貨について変数対ごとの散布図を示した図 3 では、BEF と FRF がほとんど完全相関の関係にあることが分かる。一方、他の通貨との間の散布図には複雑なトラジェクトリが描かれている。

この図は相関異常解析の重要な出発点を示唆している。複雑なダイナミクスを持つ系では、もし相関が完全相関から遠ければ、ペアごとのトラジェクトリは複雑で不安定なものになる。実際、図 3 に対応して計算された相関係数を表 2 に示すが、(BEF, FRF) を除いて括弧の中と外の値に非常に大きな食い違いがあることが分かる。このような状況では、異なるデータセットを比較し異常を判断する足場になりえるのは、変数同士の密接な類似性のみである。密接な類似性を「近傍」という言葉で言い表すとすれば、異常検知のための我々の出発点は次の仮説である。

**仮説 1 (近傍の保存)** 系が正常稼動していれば、実験条件のばらつきに対して、各変数の近傍グラフはほとんど不変である。

近傍グラフの形式的な定義は以下順次行う。この仮説に基づき、この論文では、相関異常の検知という問題を、各変数の近傍グラフの変化を定量的に捉える問題と等値する。これはもちろん仮説に過ぎないが、我々の観察によれば、複雑なダイナミックスをもつ系では多くの実用的な状況で妥当であると考えている。

ここで本論文で取り扱う問題の定義を行う。2 つのデータセット

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

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。$\mathcal{D}_{\mathrm{A}}$ および $\mathcal{D}_{\mathrm{B}}$ において、測定の回数 $N_{\mathrm{A}}$、$N_{\mathrm{B}}$ は一般には異なるが、変数の数 $M$ は同一であるとする。また、対応する各変数は物理量として同一のものとする。例えば、データセット $\mathcal{D}_{\mathrm{A}}$ における $\boldsymbol{x}_A^{(n)}$ の第 1 次元が大気圧を表すものであるとすれば、$\mathcal{D}_{\mathrm{B}}$ における第 1 次元もまた同じ物理量を（ある異なる状況の中で）測定したものである。

我々の解くべき問題は次のとおりである。

**定義 1 (相関異常スコアリング問題)** データセット $\mathcal{D}_{\mathrm{A}}$ と $\mathcal{D}_{\mathrm{B}}$ 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。

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。次節で両者の関係について若干議論する。

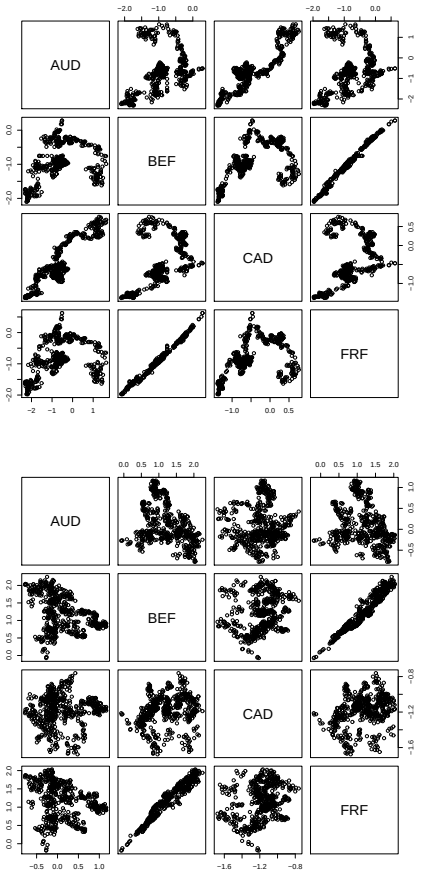


図 3: *Actual spot rates* データにある 4 つの通貨に関する散布図プロット。上： 最初の 500 日間、下： 最後の 567 日間。

## 2.2 グラフィカル・ガウシアン・モデル

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

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

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 $\Sigma$ 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。

$$\Lambda_{i,j} = 0 \quad \Rightarrow \quad x_i \perp\!\!\!\perp x_j \,|\, \text{other variables} \quad (2)$$

表 2: 図 3 に示す散布図プロットに対応した相関係数。カッコ前の数字は上の図（カッコの中は下の図）に対応する。

| | BEF | CAD | FRF |
|---|---|---|---|
| AUD | 0.31 (-0.37) | 0.91 (0.04) | 0.26 (-0.23) |
| BEF | | 0.46 (0.19) | 0.99 (0.97) |
| CAD | | | 0.41(0.30) |

ここで ⫫ は統計的独立を示す。

条件 (2) 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。$(x_i, x_j)^\top$ 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 [4] の Sec. 2.3 参照）を用いて、求める条件付き分布は

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a| - \Lambda_{aa}^{-1}\Lambda_{ab}\boldsymbol{x}_b, \Lambda_{aa}^{-1}) \tag{3}$$

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

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \tag{4}$$

と置いた。この場合、$\Lambda_{aa}$ は $2 \times 2$ 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。

これらの結果を用いれば、近傍、近傍グラフ、近傍選択を次のように定義できる。

**定義 2 (近傍)** $\Lambda_{i,j} \neq 0$ ならば、頂点 $x_i$（$x_j$）は $x_j$（$x_i$）の**近傍**である。$x_i$ の**近傍グラフ**とは、$x_i$ の近傍をすべて含み、したがって、$x_i$ と近傍との間に辺が張られたグラフである。**近傍選択**とは、各頂点の近傍をすべて列挙することである。

我々の最初の目標は、疎な $\Lambda$ を見出すことである。この $\Lambda$ の要素は、本質的な依存関係にある変数対に対しては非ゼロの値をとり、おそらくはノイズにより弱く関係しあっているだけの変数対にはゼロとなることを期待する。そのような疎な $\Lambda$ は、ノイズに由来しないような本質的な依存関係を表現するものと考えられ、したがって、相関異常の検知に有用なはずである。しかし実世界のノイジーなデータでは、標本共分散行列 $\mathsf{S}$ の要素が厳密にゼロになることはありえず、それゆえ $\Lambda$ もまた一般には疎にならない。その上、もし強く相関している変数が存在すれば $\mathsf{S}$ はランク欠損を起こし、その逆行列は存在すらしない。たとえ $\mathsf{S}$ が理論上は正則であったとしても、$M$ が数 10 を越えると、逆行列が数値的に不安定になることはよくあることである。伝統的な共分散構造選択 [6] では、まず $\Lambda$ の存在を仮定し、その要素を最尤法のある制約を満たすようにひとつずつゼロにしてゆく。したがって、$\mathsf{S}$ の逆が求まらないようなデータを扱うのが難しかった。この点が、一般の系の依存性解析において、共分散構造選択理論の利用が妨げられてきた主たる理由であった。我々は、いくつかの変数が強い相関を持つことを明示的に前提にし、以下述べるような $\mathrm{L}_1$ 制約を課した最尤推定に基づく構造選択手法により、上記実用上の問題を取り除く。

本節の最後に、GGM によりグラフを学習することと、対象とするセンサーデータの生成モデルが多変量正規分布であると仮定することが同じでは**ない**ことを強調しておきたい。図 3 からも明らかなように、データの分布自体は決して多変量正規分布で表されるようなものではない。そのようなデータを GGM で学習するということは、線形相関という色眼鏡でデータを見直す（逆に言えば複雑な非線形相関は捨象してしまう）ということを意味する。これは一般には無謀な簡単化であるが、我々の目標からすれば正しいアプローチであると言える。なぜなら、我々は近傍保存仮説を破るような状況だけに興味があるからである。

# 3 関連研究

## 3.1 グラフからの異常発見

グラフ系列からの異常検知もしくは変化検出は実用上非常に重要な問題である [15]。Sun ら [24] は単一のグラフ上で頂点同士の近接度を計算することで、異常な頂点を同定する手法を提案した。彼らのタスクは、頂点の同定を含むという点で我々の問題と似ているが、一般の 2 つのグラフではなく、単一の 2 部グラフのみを対象にしているという点で問題設定が異なる。Sun らはまた、グラフの頂点のクラスタリングを行うことで、グラフ系列からの変化検出の問題を考えた [23]。グラフにおける頂点のクラスタリングは近傍選択問題と似ているが、彼らの手法、あるいはこれまで提案されたほとんどの手法は、密なグラフに対して適用するのが一般に難しい。Tong ら [25] もまたグラフ系列からの変化検出問題を扱ったが、彼らの目標は個々の頂点に対して異常度を計算することではなかった。Xuan と Murphy [26] は多次元時系列を分割するというタスクに取り組んだ。このタスク自体は我々のものと異なるが、彼らの使った $\mathrm{L}_1$ 制約付きの最尤推定に基づく構造学習手法は本質的には我々のそれと共通である。

### 3.2 構造学習

共分散選択 [6] は疎構造学習のための標準的な手法である。しかしながら実用上は、先に述べた逆行列の存在に関わる問題に加え、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 [7]、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。

そのような場合、$L_1$ 制約付き回帰による方法 [21, 5, 3, 9] や、ベイズ的な疎構造学習 [20] の方法が選択肢となりうる。しかし現時点では、相関が強い変数が混じるような状況において、どれが最善の疎構造学習なのかについて一般的な合意は形成されていない。因果グラフの学習というタスクにおいて、Arnold ら [2] は、SIN [7] および Lasso [21] を含む複数の疎構造学習の手法を比較した。ただ、彼らの目標は因果グラフそれ自体を学習することが目標であり、また、相関の強い変数が混じる状況に考察を加えているわけではない。

### 3.3 2 標本検定

2 標本検定とは、2 つのデータセットの相違を検知し評価することを目的にした統計的検定のことである。形式的に言えば、2 標本検定は $p_{\rm A} = p_{\rm B}$ か $p_{\rm A} \neq p_{\rm B}$ かを判定するためのものである。ここで、$p_{\rm A}$ および $p_{\rm B}$ は、それぞれ $\mathcal{D}_{\rm A}$ と $\mathcal{D}_{\rm B}$ を表す確率分布である。2 標本検定は統計学において非常に長い歴史を持ち、Kolmogorov-Smirnov 検定 [10] や最近傍検定 [14] など、それぞれ特色のある各種の方法が知られている。2 標本検定は我々の問題に近いが、異なるのは、我々は単に $p_{\rm A}$ と $p_{\rm B}$ が全体としてどの程度異なるかを述べるだけでなく、個々の変数の寄与度を計算することを目標にしているという点である。

2 標本検定に関連して、カーネル法を用いた統計的独立性の検定が最近注目を集めている。Gretton らは 2 標本検定と独立性検定に対して、カーネル関数を用いた新たな統計量を提案した [12, 13]。Fukumizu らは条件的独立性を検定するために、再生核ヒルベルト空間で定義された共分散作用素を用いることを提案した [11]。これらの研究はある点において GGM の拡張とみなすことができ、図 3 に示したような複雑な相関を持つ系の構造学習に使える可能性はある。しかしその場合、独立性の定義の一般化と共に、暗黙に線形相関を前提とする仮説 1 もまた一般化する必要がある。実用的な要請を満たしつつ、これをどのように実行するかについてはまだ知られていることはない。この点の議論の詳細は本論文では扱わないが、興味深い今後の課題といえよう。

## 4 疎構造学習

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 $\mathsf{S}$ は

$$\mathsf{S}_{i,j} \equiv \frac{1}{N} \sum_{n=1}^{N} x_i^{(n)} x_j^{(n)} \tag{5}$$

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。

### 4.1 正規化項付き最尤推定

GGM では、構造学習は多変量正規分布（式 (1)）の精度行列 $\Lambda$ 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 $\Lambda$ 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 (5) を使った。行列の微分に関するよく知られた公式

$$\frac{\partial}{\partial \Lambda} \ln\det(\Lambda) = \Lambda^{-1}, \quad \frac{\partial}{\partial \Lambda} \mathrm{tr}(\mathsf{S}\Lambda) = \mathsf{S} \tag{6}$$

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。

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。

$$\Lambda^* = \arg\max_{\Lambda} f(\Lambda; \mathsf{S}, \rho), \tag{7}$$

$$f(\Lambda; \mathsf{S}, \rho) \equiv \ln\det\Lambda - \mathrm{tr}(\mathsf{S}\Lambda) - \rho ||\Lambda||_1 \tag{8}$$

ここで $||\Lambda||_1$ は $\sum_{i,j=1}^{M} |\Lambda_{i,j}|$ により定義される。$L_1$ 制約の一般的性質から、$\Lambda$ の多くの要素が厳密にゼロになることが期待される。罰金項の重み $\rho$ は入力パラメターとなるが、幸いにしてこれは、どの程度の相関係数の値までノイズ由来のものとみなすかについての閾値という明確な意味を持っている。これについては後で述べる。

## 4.2 グラフィカル Lasso アルゴリズム

式 (7) は凸計画問題であり [3]、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 [9] 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 (7) の問題を、ブロック勾配法 [3, 8] 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。公式 (6) を用いると、式 (7) の勾配が

$$\frac{\partial f}{\partial \Lambda} = \Lambda^{-1} - \mathsf{S} - \rho \,\mathrm{sign}(\Lambda) \tag{9}$$

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

方程式 $\partial f/\partial \Lambda = 0$ 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、$\Lambda$ とその逆行列を次のように分割しよう。

$$\Lambda = \begin{pmatrix} L & \boldsymbol{l} \\ \boldsymbol{l}^\top & \lambda \end{pmatrix} \quad \Sigma \equiv \Lambda^{-1} = \begin{pmatrix} W & \boldsymbol{w} \\ \boldsymbol{w}^\top & \sigma \end{pmatrix} \tag{10}$$

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。この $x_i$ による分割に対応して、標本共分散行列 $\mathsf{S}$ も同様に分割するものとし、

$$\mathsf{S} = \begin{pmatrix} S^{\setminus i} & \boldsymbol{s} \\ \boldsymbol{s}^\top & s_{i,i} \end{pmatrix} \tag{11}$$

のように書いておく。

ここで方程式 $\partial f/\partial \Lambda = \mathsf{0}$ の解を求めよう。$\Lambda$ 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

$$\sigma = s_{i,i} + \rho \tag{12}$$

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、

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||W^{\frac{1}{2}}\boldsymbol{\beta} - \boldsymbol{b}||^2 + \rho \, ||\boldsymbol{\beta}||_1 \right\} = 0 \tag{13}$$

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。導出については我々の別論文 [16] を参照されたい。これは $L_1$ 制約付きの 2 次計画問題であり、これもやはり座標ごとの劣勾配法 [9] で効率的に解くことができる。

最終的な解 $\Lambda^*$ を得るため、式 (13) 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 (12) 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。実際、後に示すように、いくつかの変数に強い相関があっても、この算法は安定して妥当な解を与える。

## 4.3 Lasso との関係

gLasso により導かれた座標ごとの最適化問題（式 (13)）には、Lasso に基づく構造学習法との明らかな類似がみられる。Meinshausen-Bühlmann の方法 [21] は、それぞれの $x_i$ ごとに分離した Lasso 回帰の問題

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} ||\mathsf{Z}_i\boldsymbol{\beta} - \boldsymbol{y}_i||^2 + \mu ||\boldsymbol{\beta}||_1 \right\} \tag{14}$$

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

$$\boldsymbol{z}_i^{(n)} \equiv (x_1^{(n)}, .., x_{i-1}^{(n)}, x_{i+1}^{(n)}, ..., x_M^{(n)})^\top \in \mathbb{R}^{M-1} \tag{15}$$

に対して $\mathsf{Z}_i \equiv [\boldsymbol{z}_i^{(1)}, ..., \boldsymbol{z}_i^{(N)}]^\top$ と置いた。

$\mathsf{S}$ の定義（式 (5)）を用いれば、もし条件

$$W = S^{\setminus i} \quad \text{and} \quad \rho = \mu \tag{16}$$

が成り立てば、この問題が式 (13) と等価であることが分かる。$W$ は $\Lambda^{-1}$ の主対角行列であるので、$\rho$ 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。この点で、gLasso は Meinshausen-Bühlmann [21] における最適化問題とよく似ているものの、微妙に異なる問題を解いていると言える。これは一見小さな違いであるが、構造の安定性という性質に劇的な相違をもたらすことを次節で見る。

## 4.4 $\rho$ の値の選択

これまで、罰金項の係数 $\rho$ をある所与の定数と見なしてきた。正規化理論に基づく多くの機械学習の問題

において、罰金項の係数をどのように選ぶかはなかなか難しい問題である。ただ、我々の問題においてはむしろ、それを入力パラメターとして扱ったほうが便利である。なぜなら、我々のゴールは何か「真の」構造をひとつ決めるというものではないからである。

次の結果は、$\rho$ の値をどう決めるかについての有用な手がかりになる。

**命題 1** *$x_i$ と $x_j$ $(i \neq j)$ からなる $2 \times 2$ の問題を考えた時、式 (7) の意味で最適な $\Lambda$ の非対角要素は次のように与えられる。*

$$\Lambda_{i,j} = \begin{cases} -\frac{\mathrm{sign}(r)(|r|-\rho)}{(1+\rho)^2-(|r|-\rho)^2} & \text{for } |r| > \rho \\ 0 & \text{for } |r| \leq \rho \end{cases}$$

*ただし、$r$ は 2 つの変数の間の相関係数である。*

証明の概略は我々の別論文 [16] を参照されたい。

これはわずか $M = 2$ の時の解ではあるが、$\rho$ の値を決める際の便利な足場になりうる。例えば、相関係数の絶対値にして大体 0.5 以下をノイズによるものとして信用しないこととしよう。すると、$\rho$ として入れるのはその値以下でなければならない（おそらく、$\rho = 0.3$ あたりが有用であろう）。$\rho$ の値が 1 に近いと、得られる近傍グラフは非常に小さいものになるはずで、逆に 0 近いと、得られるグラフはすべての変数同士が結合した完全グラフに近いものになろう。

近傍保存の仮説を考えると、この性質は実用上非常に重要なものである。第 2 節において、高度に動的な多くの系において、相関係数の値は、その大きさがほとんど 1 でない限りは非常に不安定となることを見た。しかし、単に小さい行列要素の値をある閾値を用いてゼロと置いてしまっては、GGM としての数学的一貫性が失われる（例えば精度行列が正定でなくなる）。何とか結果まで出したとしても、それは閾値の選択にかなり敏感なものになろう。本節で説明した疎構造学習の方法は、理論的に首尾一貫した仕方でノイズによる望ましくない影響を除去してくれるという利点がある。

また、例えば固定した近傍数 $k$ による $k$ 近傍グラフを作るようなやり方と対照的に、この疎構造学習の方法は変数ごとに最適な近傍数を自動的に選択してくれる。もしある変数が他に超然と孤立していれば、選択された近傍数は自動的にゼロになるはずである。また、変数同士の関係において、強く結合したクラスターがもしあれば、クラスター内部の変数はお互いにお互いを近傍として認識するだろう。適応的な近傍選択を行うというのは Lasso ベースの従来手法 [21] でも同じことなのだが、後で見るように、強く相関している変数群があるときの振る舞いは一筋縄ではいかないので注意を要する。

# 5 相関異常度のスコアリング

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。この節では、これらからいかに個々の変数の異常度を計算するかについて論ずる。

## 5.1 期待 Kullback-Leibler 距離

我々に残る仕事は、$\mathcal{D}_{\mathrm{A}}$ 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定量的に計算することである。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

$$d_i^{\mathrm{AB}} \equiv \int \mathrm{d}\boldsymbol{z}_i \, p_{\mathrm{A}}(\boldsymbol{z}_i) \int \mathrm{d}x_i \, p_{\mathrm{A}}(x_i|\boldsymbol{z}_i) \ln \frac{p_{\mathrm{A}}(x_i|\boldsymbol{z}_i)}{p_{\mathrm{B}}(x_i|\boldsymbol{z}_i)} \tag{17}$$

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 (15) を参照。式 (17) において A と B を入れ替えることで、$d_i^{\mathrm{AB}}$ 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

$$\begin{aligned} d_i^{\mathrm{AB}} &= \boldsymbol{w}_{\mathrm{A}}^{\top}(\boldsymbol{l}_{\mathrm{B}} - \boldsymbol{l}_{\mathrm{A}}) \\ &+ \frac{1}{2}\left\{ \frac{\boldsymbol{l}_{\mathrm{B}}^{\top}\mathsf{W}_{\mathrm{A}}\boldsymbol{l}_{\mathrm{B}}}{\lambda_{\mathrm{B}}} - \frac{\boldsymbol{l}_{\mathrm{A}}^{\top}\mathsf{W}_{\mathrm{A}}\boldsymbol{l}_{\mathrm{A}}}{\lambda_{\mathrm{A}}} \right\} \\ &+ \frac{1}{2}\left\{ \ln\frac{\lambda_{\mathrm{A}}}{\lambda_{\mathrm{B}}} + \sigma_{\mathrm{A}}(\lambda_{\mathrm{B}} - \lambda_{\mathrm{A}}) \right\} \end{aligned} \tag{18}$$

となる。ここで、$\Lambda_{\mathrm{A}}$ およびその逆行列 $\Sigma_{\mathrm{A}}$ をそれぞれ次のように分割した（式 (10) 参照）。

$$\Lambda_{\mathrm{A}} = \begin{pmatrix} L_{\mathrm{A}} & \boldsymbol{l}_{\mathrm{A}} \\ \boldsymbol{l}_{\mathrm{A}}^{\top} & \lambda_{\mathrm{A}} \end{pmatrix} \quad \Sigma_{\mathrm{A}} \equiv \Lambda_{\mathrm{A}}^{-1} = \begin{pmatrix} W_{\mathrm{A}} & \boldsymbol{w}_{\mathrm{A}} \\ \boldsymbol{w}_{\mathrm{A}}^{\top} & \sigma_{\mathrm{A}} \end{pmatrix} \tag{19}$$

同様の分割は $\Lambda_{\mathrm{B}}$ および $\Sigma_{\mathrm{B}}$ にも適用される。定義 $d_i^{\mathrm{BA}}$ もまた、A と B を入れ替えることで得られる。式 (18) は、よく知られた分割公式 (3) を使えば容易に導出できる。

異常度の定義 (18) 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 $\lambda_{\mathrm{A}}$ および $\lambda_{\mathrm{B}}$ 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。

### 5.2 異常度の定義

上に定義した $d_i^{\mathrm{AB}}$ および $d_i^{\mathrm{BA}}$ は、第 $i$ 番目の頂点の周りの近傍グラフの変化を定量的に測る指標である。この量が大きければ大きいほど、$x_i$ が絡む変化は大きくなる。したがって、近傍保存の仮説を前提にすれば、第 $i$ 変数の異常度を次のように定義するのが自然である。

$$a_i \equiv \max\{d_i^{\mathrm{AB}},\ d_i^{\mathrm{BA}}\} \tag{20}$$

この定義は、我々の以前の提案 [17] の自然な拡張になっている。以前の方法の問題点のひとつは、近傍の数をある定数 $k$ に固定したことである。また、相違度を定義するにあたっては、直感的に定義した指標を採用したため、例えば、 $x_i \to -x_i$ のような変化を検知できないという点も実用的な問題であった。ここでは、異常度について情報論的に自然な尺度を使うことを提案した。それにより、KL 距離の意味で、分布に影響を及ぼしえる任意の変化を反映できるようになったわけである。

### 5.3 手法の要約

相関異常度を計算する我々の方法は 2 段階からなる。第 1 段は疎な構造を見出し、第 2 段目は各変数についての異常度を計算する。

1. 入力:
   - 基準データ $\mathcal{D}_{\mathrm{A}}$ および検査対象データ $\mathcal{D}_{\mathrm{A}}$.
   - 罰金項の係数 $\rho$.
2. 出力: 個々の変数の異常度 $a_1, ..., a_M$.
3. 算法:
   (a) 相関係数行列 $\mathsf{S}_{\mathrm{A}}$ および $\mathsf{S}_{\mathrm{B}}$ を式 (5) により計算する。
   (b) gLasso を用いて疎な精度行列 $\Lambda_{\mathrm{A}}$ および $\Lambda_{\mathrm{B}}$ を計算し、また、副産物としてそれらの逆行列 $\Sigma_{\mathrm{A}}$ および $\Sigma_{\mathrm{B}}$ を求める。
   (c) 相違度 $d_i^{\mathrm{AB}}$ および $d_i^{\mathrm{BA}}$ を式 (18) から求め、異常度 $a_i$ を $i = 1, ..., M$ に対して計算する。

最後に、上記算法の計算量について簡単に見ておく。式 (5) からわかるように、相関係数行列を計算するための計算量は $O(M^2N)$ である。精度行列を計算するためには、gLasso は最悪で $O(M^3)$ という計算量を必要とする。gLasso アルゴリズムの振る舞いは完全に分かっているわけではないが、元になる行列が疎な場合は大幅に小さい計算時間で済むことが知られている [9]。構造学習の計算量の最適化についてはまだ研究の余地があり、今後の課題といえよう。

## 6 実験

この節ではまず、共線形性が強い場合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。次いで、実際のセンサーデータを用いて我々の異常検知指標の性能を評価する。

### 6.1 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 $\mathrm{L}_1$ 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 [21] 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 [22, 5]。そこで、もうひとつの比較対象として、適応 Lasso（adaptive lasso）[27] 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 [27] を参照されたい。ここでは、文献 [5] 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 [19, 7] 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 2 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 $\Lambda$ 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 4 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。

この小節をまとめると、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。

## 6.2 異常度の比較

ここでは 4 つの考えうる異常度の指標について比較を行う。最初のものが我々の提案する指標で（`KL` と表す）、GGM から導かれる条件付き KL 距離の平均を計算するものである。第 2 および第 3 の指標は、以前の論文 [17] において使われた指標であり、次のように書かれる。

$$d_i^{\mathrm{AB}} = \left| \frac{\tilde{\boldsymbol{l}}_{\mathrm{A}}^{\top}(\boldsymbol{s}_{\mathrm{A}} - \boldsymbol{s}_{\mathrm{B}})}{(1 + \tilde{\boldsymbol{l}}_{\mathrm{A}}^{\top}\boldsymbol{s}_{\mathrm{A}})(1 + \boldsymbol{s}_{\mathrm{B}}^{\top}\tilde{\boldsymbol{l}}_{\mathrm{A}})} \right| \tag{21}$$

ここで $\tilde{\boldsymbol{l}}_{\mathrm{A}}$ は、$x_j$ is 1 if $x_j$ が $x_i$ の近傍である時に 1、そうでないときに 0 となるような 2 値の「接続ベクトル」である。また、共分散行列 $\mathcal{D}_{\mathrm{A}}$ に対しては式 (11) と同様な分割がなされているものとする。接続ベクトルを得るために、第 2 の指標では `Glasso` を用い（これを stochastic neighborhood + `Glasso` の意味で `SNG` と呼ぶ）、第 3 の指標では、文献 [17] と同様の $k$ 最近傍法を用いる（これを stochastic neighborhood + $k$-NN の意味で `SNN` と呼ぶ）。

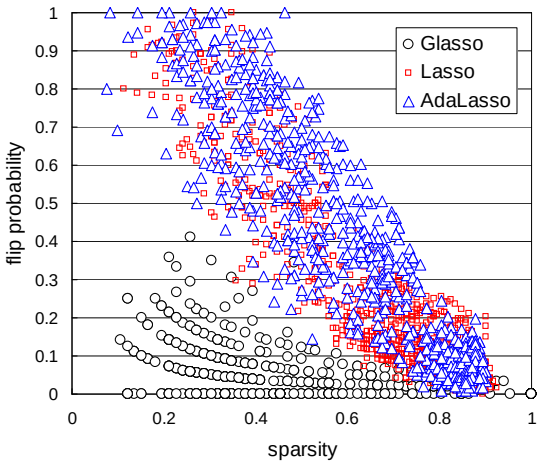


図 4: 疎度の関数としてプロットしたフリップ確率。`Lasso` と `AdaLasso` における著しい不安定性に注目。

表 3: 比較された異常度指標の一覧と、最善の AUC 値。

| symbol | neighborhood | metric | best AUC |
|---|---|---|---|
| `KL` | `Glasso` | 式 (18) | **0.96** ($\rho = 0.3$) |
| `SNG` | `Glasso` | 式 (21) | 0.93 ($\rho = 0.7$) |
| `SNN` | $k$-NN | 式 (21) | 0.87 ($k = 2$) |
| `LR` | `Glasso` | 式 (22) | 0.81 ($\rho = 0.5$) |

最後に、第 4 の指標は、統計学における最も基本な変化検出の指標である尤度比に基づき

$$d_i^{\mathrm{AB}} = 1 - \prod_{n=1}^{N_{\mathrm{A}}} \frac{p_{\mathrm{A}}(x_{Ai}^{(n)}|\boldsymbol{z}_{Ai}^{(n)})}{p_{\mathrm{B}}(x_{Ai}^{(n)}|\boldsymbol{z}_{Ai}^{(n)})} \tag{22}$$

で定義される。もしデータ $\mathcal{D}_{\mathrm{A}}$ が完全に $p_{\mathrm{A}}$ および $p_{\mathrm{B}}$ で説明されれば、$d_i^{\mathrm{AB}}$ は 0 になるはずである。さもなくば、1 よりも小さい正のある値をとるはずである。上記の定義においては、$d_i^{\mathrm{BA}}$ は A と B を入れ替えることで容易に得られる。最終的なスコアは、式 (17) 同様、$\max\{d_i^{\mathrm{AB}}, d_i^{\mathrm{BA}}\}$ で求められる。

**データ**。ここでは *sensor_error* というデータを用いた。これはあるプロトタイプの自動車の多数回の走行データに基づいて生成したシミュレーションデータである[2]。もともとこのデータは、急ブレーキ時の挙動を

[2]生データは、定常的なトレンドがなくなるように前処理され、す

調べる目的で生成されたため、時系列は高度に非定常であり、データの主要部を時系列的に予測できるモデルを学習することは困難である。データは、正常時における 79 回の試験走行と、異常を含む状態での 20 回の走行を含む。それぞれの走行はおよそ $N = 150$ 点を含み、変数の数は $M = 44$ である。$N$ は $M$ と同じ程度であり、$N$ が大きい時の性質を利用した漸近解析は難しい。データに含まれる異常は、センサーの配線ミスによるものであり、$x_{24}$ および $x_{25}$ に異常センサーが現れるように変数の並びを調整した。個別のセンサーは正常であるがゆえ、配線ミスによる異常は検出が最も難しい異常のひとつである。

図 5 はある特定の走行において、$M = 44$ の中の 4 つのセンサーについて変数対ごとの散布図を示したものである。注意深く図を見ると、$x_{24}$ と $x_{32}$ の散布図が、正常時は強い逆相関を呈しているのに対し、異常時にはそれが失われていることが分かる。これは実際に $x_{24}$ の配線ミスに起因するものであるが、散布図における軌跡の複雑さや不安定性を見れば、これを検知することの難しさが了解される。実際、Wishart 分布理論に基づく相関係数の仮説検定の手続き [1] を用いた場合、意味のある $p$ 値に対し、相関係数行列のほとんどすべての行列要素に対して、参照データとの同一性が棄却されてしまい、異常変数の同定どころの話ではない。

**評価尺度**。我々の問題設定では、参照データと検査対象データの間に、$20 \times 79 = 1580$ 通りの可能なテストの方法がある。これらの結果を要約するため ROC 曲線（Receiver Operating Characteristic curve）を用いることにする。ROC 曲線は、ここでは、検知率（本当に異常な変数がどれだけ異常と指し示されたか）とデータ割合（異常度の高い順にいくつのデータを見たか）の関係として定義する。この場合、ROC 曲線の横軸は、上位いくつの変数を見たかに対応して、$0, \frac{1}{M}, \frac{2}{M}, ..., 1$ という値をとる。ROC カーブを更に要約した指標として、通常行われるように、ROC 曲線の下の面積を表す AUC（Area Under Curve）を採用する。

**結果**。図 6-8 に、$\rho = 0.3, 0.5, 0.7$ に対する ROC 曲線を示す。参考のため、点線はランダムに異常変数を選択した場合を表す。SNN に関しては、最善の AUC を与える $k = 2$ に対応する同一の曲線を各 $\rho$ の図に描いている。4 つの指標を比べると、まず LR が他よりも非常に悪い結果となっていることが分かる。この理由のひとつは、LR がモデルの生成とスコアの計算双方に（$N_{\mathrm{A}}$ もしくは $N_{\mathrm{B}}$ 個の）全データ点を使っていることであろう。このデータは非常にノイジーなので、そのような方針はノイズの望ましくない効果をより強調する方向に働くことが考えられる。

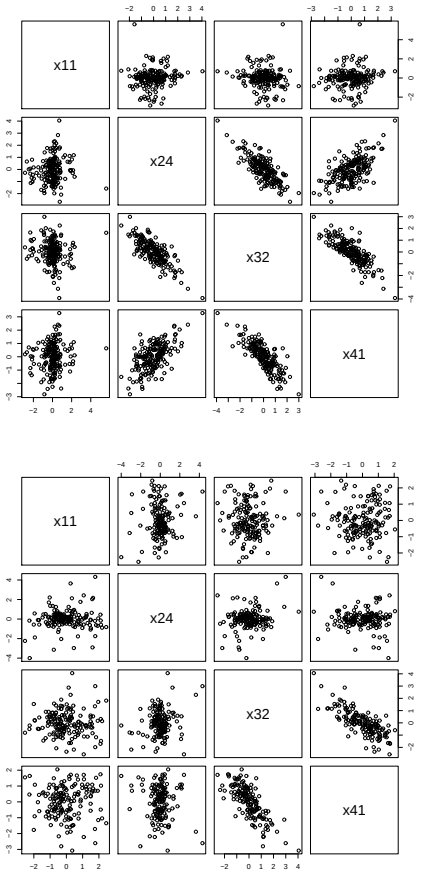


図 5: *sensor_error* の、変数対ごとの散布図の一部。上：第 10 番目の参照用の走行データ。下： 第 3 番目の異常時データ。

表 3 は各手法における最善の AUC 値をまとめたものである。KL と SNG が SNN より顕著に優れており、これは適応的近傍選択機能が非常に有用であることを示している。計算結果を詳しく見ると、$\rho = 0.3$ という値に対しては、このデータだと、大雑把に言って、相関係数の大きさにしておよそ 0.6 以下の結合が枝狩りされる（そしてこのとき、参照用データの疎度は約 0.9 程度である）。近傍保存の仮説と図 5 に示すようなデータの大きな揺らぎを考えれば、相関係数の枝狩りとしてこれは妥当な閾値だと言える。$\rho$ が 0.5 より大きくなると、SNG が KL よりも良い AUC 値を与えることは興味深い。このような大きな $\rho$ では、得られる構造は非常に疎となり（$\rho = 0.7$ における疎度は 0.98 であった）、データセット間の相違に対する個々の変数の寄与は、$\sigma_{\mathrm{A}}$ や $\lambda_{\mathrm{B}}$ で表される個々の変数の分散の効果が相対的に大きな割合を占める。SNG は、個々の分散の項を含まない単純な定義を採用しているので、個々の分散の値のランダムな揺らぎに対してはむしろ頑強なの

べての変数のサンプリング間隔が一致するように適当に補間した。生成した 79+20 個の相関係数行列のデータは Web にて公開予定である。

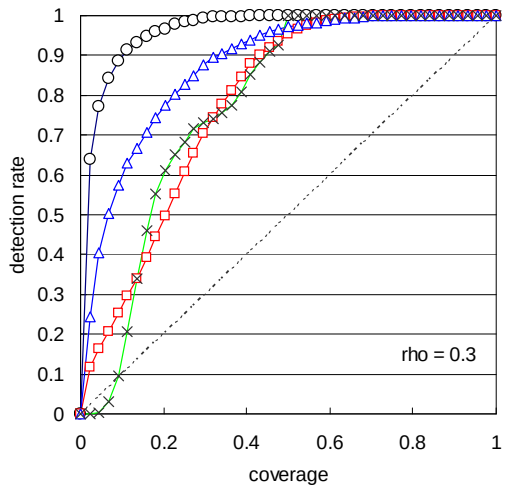


図 6: $\rho = 0.3$ に対する ROC 曲線。KL (○)、SNG (□)、SNN (△)、LR (×) の比較。

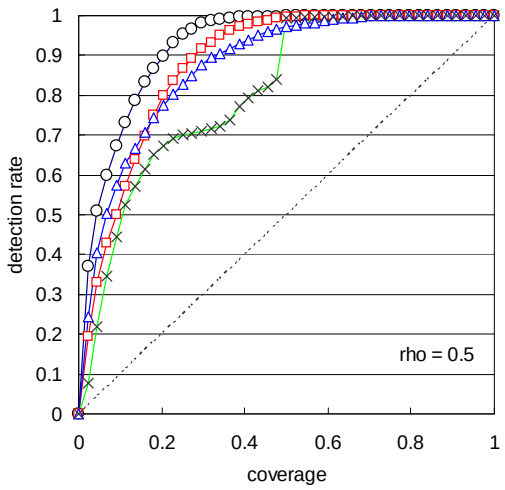


図 7: $\rho = 0.5$ に対する ROC 曲線。KL (○)、SNG (□)、SNN (△)、LR (×) の比較。

であろう。

## 7 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。我々の知る限り、本論文は、疎構造学習を用いてこの種の問題を扱った初めての仕事である。

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。

我々はまた、いくつかの相関異常度のスコアリング手法を、実際の機械系のセンサーデータに対して適用し、新たに提案した期待 KL 距離に基づく尺度が優れた異常検知性能を持つことを示した。

## 参考文献

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 3rd. edition, 2003.

[2] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical Granger methods. In *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pages 66–75, 2007.

[3] 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.

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

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

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

[7] 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.

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

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

[10] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7:697–717, 1979.

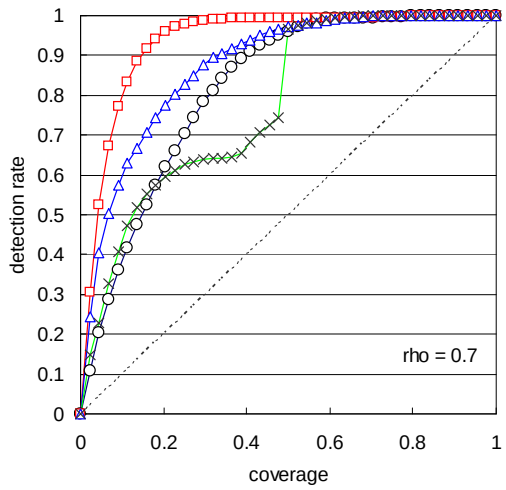


図 8: $\rho = 0.7$ に対する ROC 曲線。KL (○)、SNG (□)、SNN (△)、LR (×) の比較。


[11] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*.

[12] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

[13] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, 2008.

[14] Z. Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Annals of Statistics*, 16:772–783, 1988.

[15] 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.

[16] 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.

[17] 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.

[18] E. Keogh and T. Folias. The UCR time series data mining archive [`http://www.cs.ucr.edu/~eamonn/TSDMA/index.html`]. 2002.

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

[20] F. Li and Y. Yang. Using modified lasso regression to learn large undirected graphs in a probabilistic framework. In *Proc. National Conf. Artificial Intelligence*, pages 801–806, 2005.

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

[22] 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.

[23] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. GraphScope: parameter-free mining of large time-evolving graphs. In *Proc. the 13th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pages 687–696, 2007.

[24] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Proc. IEEE Intl. Conf. Data Mining*, pages 418–425, 2005.

[25] H. Tong, S. Papadimitriou, P. S. Yu, and C. Faloutsos. Proximity tracking on time-evolving bipartite graphs. In *Proc. 2008 SIAM Intl. Conf. Data Mining*, pages 704–715, 2008.

[26] X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *Proc. the 24th Intl. Conf. Machine Learning*, pages 1055–1062, 2007.

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