依存関係にスパース性を入れる
— グラフィカル lasso の話 *

井手剛 (Tsuyoshi Idé)
IBM Research, T. J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10592, USA.
tide@us.ibm.com

2016 年 11 月 26 日

1　実世界で現れるいくつかのスパース性

スパース化の話を初めて聞いたのはもう 10 年以上前、機械学習の勉強を始めて間もない頃だった。たとえば回帰問題で、特徴量がちょっとくらい多くても、「えるわん」をかませれば自動的に無駄な特徴が消えてなくなる、みたいな話は、初学者の私にはとてつもなく神秘的に聞こえた。重要でない回帰係数の絶対値がある程度小さくなるのなら想像もできるが、きっかりゼロ。「えるわん」、すげー。現代の機械学習のブレイクスルーをいくつか挙げろと言われたら、深層学習、カーネル法に加えて、スパース化の理論を間違いなく挙げたい。

さて、データがスカラーの出力 y と M 次元ベクトル $\boldsymbol{x}$ の組で $\mathcal{D} = \{(y^{(n)}, \boldsymbol{x}^{(n)}) \mid n = 1, \ldots, N\}$ のように与えられているとき、たとえば線形回帰においてスパースな解を実現するためのもっとも基本的な方法は

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{N} \|\mathsf{X}^\top \boldsymbol{\beta} - \boldsymbol{y}\|^2 + \rho \|\boldsymbol{\beta}\|_1 \right\} \tag{1}$$

を解くことである。これを lasso 回帰と呼ぶ。ただし、$\mathsf{X} \equiv [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}]$ はデータ行列で、あらかじめ y も x も標本平均がゼロに標準化されているとする。また、$\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{M} |\beta_i|$ である（β_i は係数ベクトル $\boldsymbol{\beta}$ の第 i 成分、$|\beta_i|$ はその絶対値）。この項がくだんの ℓ_1 正則化項で、ρ はこの正則化項をどれだけ重視するかを表す正の定数である。これを解くことにより、回帰係数がスパースに、すなわち、β_1 から β_M のうち多くの係数が厳密に 0 になる。これは、変数がたくさんあるときに、余分な変数を減らすという意味でのスパース化である（図 1 の (1)）。

---

* Tsuyoshi Idé, “Making relationships simple – A story of graphical lasso”, Iwanami Data Science, Vo.5 (2016) pp.48-63. 岩波データサイエンス刊行委員会編、岩波データサイエンス Vol.5、pp.48-63、2016。

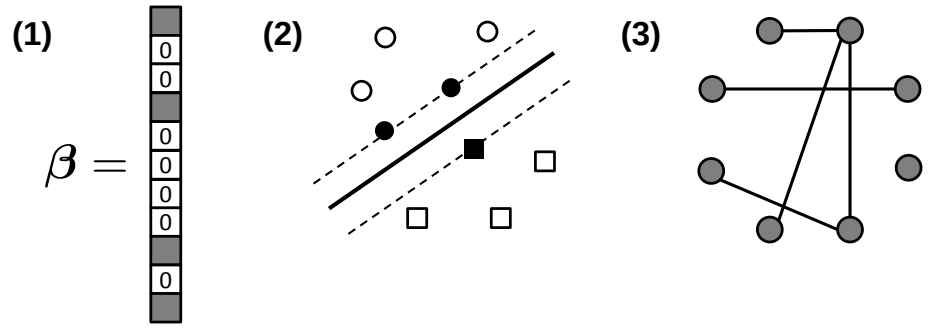

図 1　機械学習で出てくる主な 3 つの種類のスパース性。(1) 少ない変数だけでモデルを表す。(2) 少ない標本だけでモデルを表す。(3) 少ない関係だけに割り切る。

一方、変数ではなく、余分な標本を消すという意味でのスパース化も使われている。たとえば、$\mathcal{D} = \{\boldsymbol{x}^{(n)} \mid n = 1, \ldots, N\}$ という M 次元の標本 N 個から、パラメター $\boldsymbol{\theta}$ を持つ確率モデル $p(\boldsymbol{x}|\boldsymbol{\theta})$ を使って密度推定を行うことを考える。基本的な戦略は、対数尤度

$$\min_{\boldsymbol{\theta}} \left\{ \sum_{n=1}^{N} \ln p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta}) \right\} \tag{2}$$

を最大化するように $\boldsymbol{\theta}$ を決めることである（最尤推定)。実用上よくあるシナリオは、観測データがノイズで汚染されており、必ずしもすべての標本が信用できないという場合である。このようなときは、上記の各項に重み w_n をつけ、さらに、N 次元ベクトルとしての $\boldsymbol{w} = (w_1, \ldots, w_N)^\top$ に対する適切な正則化項を付加して、$\boldsymbol{\theta}$ と $\boldsymbol{w}$ に双方について同時に最適化する問題を考えることができる[*1]。

あるいは、lasso 回帰と並んで有名なサポートベクトルマシン（支持ベクトル学習器）という分類器では、各標本に対して係数 $\alpha_1, \ldots, \alpha_N$ を割り当て、それに矩形制約という制約を付けて 2 次計画問題を解くことで、標本にわたるスパース性を実現している。すなわち、2 値分類の分類面に遠く、分類面の決定に直接寄与しない標本の重みはゼロになる（図 1 の (2))。

本章の主題は、上記二つのスパース化とはまた別の、第 3 のスパース性についてである。たとえば、ある機械を決まった数のセンサーで常時監視しているような場合を考える。それぞれのセンサーは物理的な意味を持っており、自明に無駄なセンサーというものはない（もしそういうものあればそもそも測らない)。このような場合に興味があるのは、変数同士の**関係**である。たとえば、

[*1] $p(\boldsymbol{x}^{(n)}|\boldsymbol{\theta})$ および正則化項の具体的な例としてはこの論文 [8] を参照。

ある部分に不具合が生じているのであれば、その部分に関係するセンサーの値を見ることで何か診断ができるかもしれない。

しかし一般に、ノイジーなデータの関係について断定的に何かを言うのは難しい。「風が吹けば桶屋が儲かる」のたとえの通り、何かは何かにいつでも微妙に関係しており、関係がないことの証明はいわゆる悪魔の証明である。ゆえ、もしスパース化の手法により、何かと何かが関係がないと言い切ることができれば、実に有用である。この場合のスパース化は、変数個々ではなく、変数同士の関係に入る（図 1 の (3)）。技術的には、問題は 2 つある。ひとつはどのように「関係」を定義するのかという問題で、もうひとつは、いかに正しく（実データのノイズに負けずに）関係を求めるかという点である。まずは前者につき次節で見ていこう。

2 「関係」を定義する

変数同士の「関係」を見積もる仕方はいろいろ考えられる。誰でも知っているのが相関係数である。先ほど同様、データ $\mathcal{D} = \{\boldsymbol{x}^{(n)} \mid n = 1, \ldots, N\}$ が与えられるとする。表記の簡単化のため、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する。この仮定の下、標本共分散行列 S は

$$\mathsf{S} \equiv \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}^{(n)} \boldsymbol{x}^{(n)\top} \tag{3}$$

のように与えられる。標準化の仮定の下では、これはデータの相関係数行列と同じものとなる。

しかし相関係数は、実世界の関係性を表す上で大いに問題がある。実例としてある経済時系列のデータ[*2]を見てみよう。これは 1986 年 10 月 9 日から 1996 年 9 月 8 日までの 10 年間にわたり、対ドル通貨レートの日次スポット値を記録したデータで、表 1 に示す 12 の通貨のデータから成る。例として 4 つの通貨だけを選び、最初の 500 日間と最後の 567 日間を切り出して変数対ごとの散布図にしたのが図 2 である。これに対応して、相関係数を計算したものを表 2 に示す。ベルギーとフランスの間には通貨協定があり、相関係数はほとんど 1 で安定しているが、その他の通貨間の関係の相関係数はまるで安定していないことがわかる。たとえば、CAD と AUD は前半では相関係数が 0.91 と非常に高いが、後半ではほとんど無相関である。つまり、相関係数は、ノイジーなデータにおいては、完全相関か完全逆相関に近い場合を別にして、信頼できる指標にはならないのである。

問題はそれだけにとどまらない。相関係数の危険性を如実に示すエピソードとして「教会と殺人のパラドックス」という話がある [10]。ある国の国勢調査において、都市ごとに教会の数と殺人や強盗などの重大犯罪事件の件数には非常にきれいな比例関係が見出された。ではこれから、「教会が多ければ多いほど強盗の数は多い」と結論できるだろうか、という話である。答えはもちろんノーで、これは人口という第 3 の変数を介して間接的に現れた相関と考えるべきである。しかし素朴に相関係数を計算するだけではそのような効果を除外することができない。

[*2] `Actual spot rates` データ。`http://ide-research.net/datasets.html` でダウンロード可能。

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

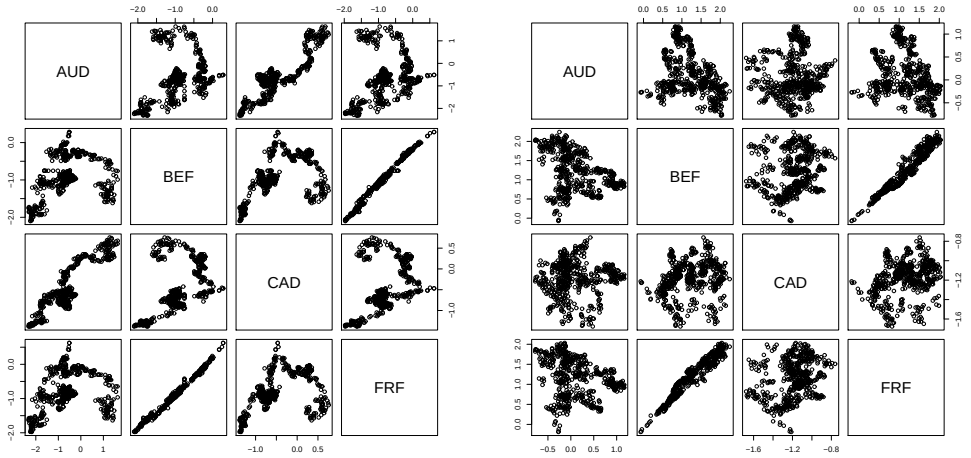

図 2　`Actual spot rates` データにある 4 つの通貨に関する散布図プロット。上： 最初の 500 日間、下： 最後の 567 日間。

表 2　図 2 に示す散布図プロットに対応した相関係数。括弧の前の数字は左の図、括弧の中の数字は右の図に対応する。

	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)

そこでグラフィカルモデルである。相関係数の問題は、値が 1 または −1 に近い場合を除いて値が容易にブレてしまうことであった。無相関というのは理論上はありえても、ノイジーな実データではまずありえない。2 つの変数の間の関係について「関係がない」ということをもう少し正確に定義する必要がある。「関係がない」ということの統計学的な自然な言い換えは「統計的に独立である」ということになろう。他の変数もあるので、「2 つの変数が、他の変数を与えた時に条件つき独立となっていれば、その 2 つは無関係である」と定義してみる。「独立」の代わりに「条件つき独立」とすることで、ほかの変数を介して間接的に現れた相関を区別し、「教会と殺人のパラドックス」のようなことを排除するわけである。この考え方でグラフと確率分布を結びつけるモデルを

一般に**対（つい）マルコフグラフ**（pairwise Markov graph）と呼ぶ。

確率分布と変数の関係を表すグラフが対応したということは、関係を求める問題（これを**構造学習**（structure learning）と呼ぶ）が、データ $\mathcal{D}$ から確率分布を学習する問題に帰着できることを意味する。どのような確率モデルを最初に仮定するかに対応し、さまざまな構造学習の手法が考えられる。特に多変量正規分布（ガウス分布）を用いる構造学習のモデルを、**ガウス型グラフィカルモデル**（Gaussian graphical model）と呼ぶ。ノイジーな実数データに関する限り、2 次を超える高次の統計量を扱うのはノイズへの脆弱性の観点から実用上困難なので、ガウス型グラフィカルモデル以外の選択肢は非常に乏しいのが現実である。

M 次元多変量正規分布の確率密度関数は次の形で与えられる。

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Lambda^{-1}) \equiv \frac{(\det \Lambda)^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Lambda (\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{4}$$

ここで Λ は精度行列と呼ばれ、概念上は共分散行列の逆行列に対応するものである。det は行列の行列式を表す。$\boldsymbol{\mu}$ はもちろん平均である。 対マルコフグラフの定義に従い、たとえば、x_1 と x_2 についての、他の変数 $x_3, \ldots, x_M$ を与えたときの条件つき分布 $p(x_1, x_2|x_3, \ldots, x_M)$ を求めてみよう。条件つき確率の定義から、

$$p(x_1, x_2 \mid x_3, \ldots, x_M) = \frac{p(x_1, \ldots, x_M)}{p(x_3, \ldots, x_M)} \tag{5}$$

だから、x_1, x_2 の関数として $p(x_1, x_2|x_3, \ldots, x_M)$ は $p(x_1, \ldots, x_M)$ に比例している。仮定 $\boldsymbol{\mu} = \mathbf{0}$ を使った上で、式 (4) から x_1, x_2 に関係する項をすべて拾うと

$$p(x_1, x_2 \mid x_3, \ldots, x_M) \propto \exp\left\{-\frac{1}{2}(\Lambda_{1,1}x_1^2 + 2\Lambda_{1,2}x_1x_2 + \Lambda_{2,2}x_2^2)\right\}$$

となることがただちにわかる。$x_3, \ldots, x_M$ を与えた時に x_1 と x_2 が統計的に独立であるためには、$p(x_1, x_2 \mid x_3, \ldots, x_M)$ が x_1 に関する部分と x_2 に関する部分の積として書ける必要があるが、この条件は上式から明らかに、$\Lambda_{1,2} = 0$ である。逆が成り立つことも明らかなので、一般的に書くと、ガウス型グラフィカルモデルの意味においては、「**もし $\Lambda_{i,j} = 0$ なら x_i と x_j は無関係**」ということである。

以上まとめると次のようになる。

- 相関係数は、変数間の関係を定義するためには役不足。グラフィカルモデルが必要。
- ガウス型グラフィカルモデルは、多変量正規分布を使って実現した対マルコフグラフ。
- ガウス型グラフィカルモデルでは、構造学習は、精度行列を推定する問題に帰着できる。

したがって、関係性にスパース性を入れるというのは、スパースな精度行列を求めることと等価である。それを具体的にどうするかを次節で見てゆこう。

3　グラフィカル lasso の衝撃

精度行列というのは定義上、共分散行列の逆行列である。だとすれば、標準化されたデータから式 (3) により標本共分散行列を求め、単にその逆行列を計算すればよいように思える。しかし疎な行列を実現するのは簡単ではない。相関係数の絶対値がある閾値以下のものをゼロにする、というような手作業的な方法だと「**正定値性**（positive definiteness）」という条件が満たされず、グラフィカルモデルとしての解釈ができないためである。Λ が正定値であるとは、Λ のすべての固有値が正であるということである。これが成り立たないと、式 (4) における指数関数の肩が正の無限大に発散しうるので、確率分布の規格化条件を満たすことが不可能になる。

また、そもそも、変数の数が多くなると逆行列の計算自体が難しい。計算時間の点もさることながら、困るのは「ランク落ち」というやつである。式 (3) において、データ行列を $\mathsf{X} \equiv [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}]$ とすると、$\mathsf{S} = \frac{1}{N}\mathsf{X}\mathsf{X}^\top$ となる。このように 2 つの行列の「2 乗」の形で書かれている行列の場合、数式上では行列の固有値が負になることはない[*3]。したがってその逆行列も負の固有値は持たないのだが、問題は、数値的にはほぼゼロの固有値がよく現れることである。そうなると逆行列が求まらず、やはり確率分布とのつながりを失ってしまう。

この正定値性という条件は、行列を求める最適化問題特有のものである。lasso 回帰より数段問題が難しいと言える。そのため、1972 年の Dempster [3] による貪欲法的な手法の提案の後、30 年以上の長きにわたり、疎構造学習（ガウス型グラフィカルモデルに対しては**共分散選択**（covariance selection）とも呼ばれる）という手法は、たかだか数個程度の変数の間の関係を求めるための思考補助ツールという位置づけを出ることはなかった [*4]。

状況が一変したのは、2006 年に Banerjee らが、ガウス型グラフィカルモデルの疎構造学習を ℓ_1 正則化つき凸最適化問題として定式化して以降である [1]。まず、スパース性を忘れて、精度行列 Λ についての最尤推定の問題を考える。各変数を平均ゼロ、標準偏差 1 に標準化したデータを考えている前提だと、対数尤度は

$$\ln \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}^{(n)} \mid \boldsymbol{0}, \Lambda^{-1}) = \sum_{n=1}^{N} \left\{ -\frac{M}{2}\ln(2\pi) + \frac{1}{2}\ln\det\Lambda - \frac{1}{2}\boldsymbol{x}^{(n)\top}\Lambda\boldsymbol{x}^{(n)} \right\} \tag{6}$$

$$= \text{c.} - \frac{N}{2}\left\{ \mathrm{Tr}\left(\Lambda \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}^{(n)}\boldsymbol{x}^{(n)\top} \right) - \ln\det\Lambda \right\} \tag{7}$$

$$= \text{c.} - \frac{N}{2}\left\{ \mathrm{Tr}(\Lambda\mathsf{S}) - \ln\det\Lambda \right\} \tag{8}$$

のように書ける。最後の式変形では式 (3) を使った。Tr は行列の跡（対角和）、c. は定数項を表す。尤度を最大にするためには、最後の式の $\{\cdot\}$ の中身を最小化すればよい。Banerjee らは、それに

[*3] X の特異値分解を $\mathsf{X} = \mathsf{U}\mathsf{Z}\mathsf{V}^\top$ とすると $\mathsf{S} = \frac{1}{N}\mathsf{U}\mathsf{Z}^2\mathsf{U}^\top$ となる。対角行列 Z^2 の要素はゼロか正。

[*4] とはいえ、マーケティングなどの分野では、共分散選択は一貫して重要な技術であり続けた。スパース化の技術が登場する前の古典技術の様子は、朝野ら [11] にわかりやすくまとめられているので一読を勧める。

lasso 風の正則化項を加えた問題が凸最適化問題となっていることを示した。解くべきは次のような最適化問題である。

$$\min_{\Lambda} f(\Lambda) \quad \text{subject to} \quad f(\Lambda) = \mathrm{Tr}(\Lambda S) - \ln\det\Lambda + \rho\|\Lambda\|_1 \tag{9}$$

括弧内の第 3 項が精度行列に対する ℓ_1 正則化項で、$\|\Lambda\|_1 = \sum_{i,j=1}^{M} |\Lambda_{i,j}|$ と定義される。ただし $\Lambda_{i,j}$ は Λ の (i,j) 成分、$|\Lambda_{i,j}|$ はその絶対値である。lasso 回帰の経験によれば、この項を付け加えることで疎な解が得られそうである。しかも、$\ln\det\Lambda$ のおかげで、Λ が負の固有値を持つことは許されず、また固有値がゼロに近づくと目的関数が正の無限大に発散するので、括弧内の最小化が達成できれば、自動的に正定な行列が解となる。手作業的に正定値性を満たす必要がない。

問題はこの最適化問題をどう具体的に解くかであるが、やや驚くべきことに、「微分してゼロとおく」という高校数学以来おなじみの方法でさわやかに解けるのである。もちろん、$\|\Lambda\|_1$ というカドのある関数のおかげで、普通の意味での微分は定義できないので、「**劣勾配法**（subgradient method）」というあやしい名前が付いているが、基本は一緒である。行列の微分に関するよく知られた公式 [2]

$$\frac{\partial}{\partial\Lambda}\ln\det\Lambda = \Lambda^{-1}, \quad \frac{\partial}{\partial\Lambda}\mathrm{Tr}(S\Lambda) = S \tag{10}$$

を使うと、式 (9) の勾配が

$$\frac{\partial f}{\partial\Lambda} = \Lambda^{-1} - S - \rho\,\mathrm{sign}(\Lambda) \tag{11}$$

と与えられることがわかる。ただし $\mathrm{sign}(\Lambda)$ は、(i,j) 要素が $\mathrm{sign}(\Lambda_{i,j})$ で与えられる行列である。符号関数 $\mathrm{sign}(\cdot)$ は、引数の符号に応じて ±1 という値をとる。引数が 0 なら符号は決まらないので、±1 ではない何かの値をとると約束する。

方程式 $\partial f/\partial\Lambda = 0$ は行列の方程式となるが、行列のひとつの行と列に着目して順繰りに解くことができる。ある特定の変数 x_i に着目し、それが最後の行と列に来るように変数の名前を付け替えたと考え、Λ とその逆行列、および標本共分散行列 S が

$$\Lambda = \begin{pmatrix} L & \boldsymbol{l} \\ \boldsymbol{l}^\top & \lambda \end{pmatrix}, \quad \Sigma \equiv \Lambda^{-1} = \begin{pmatrix} W & \boldsymbol{w} \\ \boldsymbol{w}^\top & \sigma \end{pmatrix}, \quad S = \begin{pmatrix} S^{-i} & \boldsymbol{s} \\ \boldsymbol{s}^\top & s_{i,i} \end{pmatrix} \tag{12}$$

のように分割されているものとする。L, W, S^{-i} は、それぞれ Λ, Σ, S から x_i に対応する行と列を抜いた $(M-1)\times(M-1)$ 行列である。一方、$\boldsymbol{l}, \boldsymbol{w}, \boldsymbol{s}$ は M 次元の列ベクトルになる。

Λ は正定値であるため、その対角要素は正でなければならない（証明は [10] の 10.4.2 節参照）。したがって、式 (11) の右下の対角要素に関する部分について、勾配ゼロの条件は

$$\sigma = s_{i,i} + \rho \tag{13}$$

と書かれる。これでまず σ は求まった。

次に非対角要素について考える。分割 (12) に基づいて、方程式 $\partial f/\partial\Lambda = 0$ の右上部分は直ちに

$$\boldsymbol{w} - \boldsymbol{s} - \rho\,\mathrm{sign}(\boldsymbol{l}) = 0 \tag{14}$$

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

$$\Sigma\Lambda = \begin{pmatrix} WL + \boldsymbol{w}\boldsymbol{l}^\top & W\boldsymbol{l} + \lambda\boldsymbol{w} \\ \boldsymbol{l}^\top W + \lambda\boldsymbol{w}^\top & \boldsymbol{w}^\top\boldsymbol{l} + \sigma\lambda \end{pmatrix} = \begin{pmatrix} \mathsf{I}_{M-1} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{pmatrix}. \tag{15}$$

が成り立つ。この恒等式の右上部分を使うと、

$$\boldsymbol{l} = -\lambda W^{-1}\boldsymbol{w} = -\lambda\boldsymbol{\beta}, \tag{16}$$

であることが分かる。ただし、$\boldsymbol{\beta} \equiv W^{-1}\boldsymbol{w}$ である。Λ 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 (14) は次の方程式と等価であることが分かる。

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left\{\frac{1}{2}\boldsymbol{\beta}^\top W\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{s} + \rho\|\boldsymbol{\beta}\|_1\right\} = \mathbf{0} \tag{17}$$

さらに、$W^{-1/2}\boldsymbol{\beta}$ を $\boldsymbol{b}$ とおけば、この式が

$$\min_{\boldsymbol{\beta}}\left\{\frac{1}{2}\|W^{\frac{1}{2}}\boldsymbol{\beta} - \boldsymbol{b}\|^2 + \rho\|\boldsymbol{\beta}\|_1\right\} \tag{18}$$

と等価であることが分かる。

式 (18) は、冒頭に示した lasso 回帰と同じ形をしている。Banerjee ら [1] によりすでに指摘されていたこの事実を、Friedman ら [5] はさらに深く掘り下げ、この lasso 回帰の問題が、別の劣勾配法により極めて効率よく解けることを示した。上式 (17) において β_i での微分を実行すると

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

を得る。この方程式に対する形式的な解は

$$\beta_i = \frac{1}{W_{i,i}}[A_i - \rho\,\mathrm{sign}(\beta_i)] \tag{19}$$

で与えられる。ただし、

$$A_i \equiv s_i - \sum_{m\neq i} W_{i,m}\beta_m \tag{20}$$

と定義した。

共分散行列の正定値性より $W_{i,i} > 0$ が成り立つことに注意すると、式 (19) においては、A_i と $\pm\rho$ の大小関係により β_i の符号が左右されることがわかる。たとえば $A_i > \rho$ なら、右辺はどうやっても負にはならないので、$\beta_i > 0$ しかありえない。したがって、カッコの中身は $A_i - \rho$ しかありえない。$A_i < \rho$ なら右辺はどうやっても正にはなれないので $\beta_i < 0$ であり、したがって、カッコの中身は $A_i + \rho$ しかありえない。

興味深いのは $|A_i| < \rho$ の場合である。この場合は、右辺は正にも負にもなりえる。右辺が正になるのはカッコの中身が $A_i + \rho$ となる場合であるが、これは $\mathrm{sign}(\beta_i) = -1$ のときに生ずる。しかしこれは右辺も左辺も正という仮定に反している。この場合、唯一の可能性は $\beta_i = 0$ となること（したがって $\mathrm{sign}(\beta_i)$ が何か ±1 以外の値をとること）しかない。右辺が負になるときも同様の

矛盾を導けるので、$\beta_i = 0$ しかありえないことがわかる。結局、各 i に対して次のような更新式を得る。

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

この式を収束するまで繰り返すことで解を得る。これを解いて $\boldsymbol{\beta}$ を得たとすれば Λ の対応する列を

$$\lambda = \frac{1}{\sigma - \boldsymbol{\beta}^\top W \boldsymbol{\beta}}, \quad \boldsymbol{l} = -\frac{\boldsymbol{\beta}}{\sigma - \boldsymbol{\beta}^\top W \boldsymbol{\beta}} \tag{21}$$

によって更新できる。ただしここで、式 (15) $\boldsymbol{w}^\top \boldsymbol{l} + \sigma\lambda = 1$ の右下部分と、式 (16) を用いた。また、式 (15) 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 (13) のために一定に保たれることに注意。したがって、この算法においては、$\Sigma = \Lambda^{-1}$ は Λ 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。

以上の計算は、他の変数を固定してひとつつの変数について解く、ということを行っているため、最終的な解 Λ^* を得るためには、式 (18) を $x_1, x_2, ..., x_M, x_1, x_2, ...$ と収束するまで何周でも繰り返す。Friedman らはこの手法を**グラフィカル lasso**（graphical lasso）と名づけた [5]。グラフィカル lasso は、それまで知られていた共分散選択手法に比べて数値的安定性、高速性の観点で圧倒的に優れており、まさにブレイクスルーと言うべき手法であった。数千もの変数を扱うのは古典手法では夢の世界で、それが（2007 年当時の計算機ですら）ものの 10 秒とかで解けるという報告に、衝撃を受けた研究者も多かったはずだ。

4　グラフィカルモデルをどう使うか

ガウス型グラフィカルモデルの疎構造学習の手法は、その成り立ちからして、変数間の隠れた関係を発見するために使うことができる。実例として、先に示した図 2 の右側のデータについて計算したグラフィカルモデルを図 3 に示す。計算には R の `glasso` というパッケージを使った。ここでは、**偏相関係数**（partial correlation coefficient）という量

$$r^{i,j} \equiv -\frac{\Lambda_{i,j}}{\sqrt{\Lambda_{i,i}\Lambda_{j,j}}} \tag{22}$$

を可視化してある。点線は $r^{i,j} < 0$、実線は $r^{i,j} > 0$ を表し、線が太いほど $r^{i,j}$ の絶対値が大きくなるようにしてある。$\rho = 0.3, 0.6, 0.9$ と変わるにつれ、グラフが疎になる傾向があることがわかる。西欧の大国ドイツ（DEM）に「フラン経済圏」とでも言うべき国々を加えた西欧の大陸諸国の関係が深いことが読み取れる。日本（JPY）、オーストラリア（AUD）、カナダ（CAD）はそれらと独立性が高い。特に日本は、オーストラリアと負の偏相関係数を持つなど独特である。

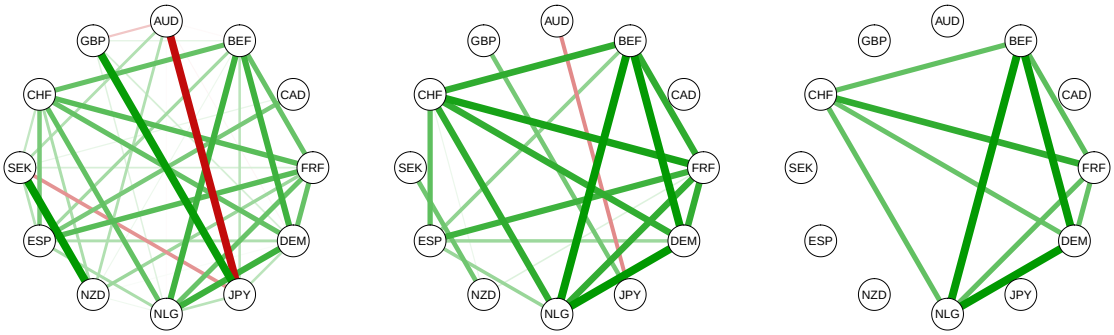

図 3　`Actual spot rates` データでのガウス型グラフィカルモデル。図 2 の右側のデータに対応。左から $\rho = 0.3, 0.6, 0.9$ での結果。

しかし一方で、ρ によりこのようにグラフの見かけが激変するのなら、信ずべきなのはいったいどれなのかという当然の疑問が出てくる。実はこれは、グラフィカルモデリング一般に宿命的な問題である。実世界のデータ解析に疎構造学習を使う場合に注意すべき点は、ほとんど常に正解データがなく、構造学習それ自体のみでは、正則化項の係数 ρ を決める手段が乏しいことである。それこそ、桶屋が儲かったのは、本当に風が吹いたためだったかもしれないが、真実は誰も知らない。

この点に対処するための考え方はおおよそ二つある。ひとつは、モデルが正しいことを前提に、**尤度交差検証**（likelihood cross-validation）を行うことである。手順は次の通りである。ρ の候補をいくつか用意しておく。データ $\mathcal{D}$ を分割して学習用と検証用に分け、学習用データを使い ρ の候補のそれぞれについてガウス型グラフィカルモデルを学習する。次に、検証用のデータに対し、尤度 (6) を計算し、尤度が最も高くなるような ρ を選択する。交差検証の代わりに、拡張ベイズ情報量基準 [4] などの情報量基準を使いたい人もいるかもしれない。しかし実用上の多くの場合、これらの方法だけで ρ を決めるのはリスクが高い。問題は、そもそも、単一の多変量正規分布で表されるほど実世界のデータは甘くないという点である。モデルがデータに合っていないのなら、尤度の観点で最善を尽くしてもあまり意味はない。

二つ目の考え方は、学習された疎構造を別のタスクのための入力として使い、その別のタスクの交差検証を元に ρ などの未知パラメターを決めるというものである。実応用上の多くの場合、こちらの方が合理的である。この場合、得られた疎構造は、ガウス型グラフィカルモデルという「めがね」をかけて見たときの実世界の表現、といった程度の意味であり、仮にめがねを通して見た像がゆがんでいたとしても（単一の多変量正規分布という仮定が密度推定の意味で最善でなかったとしても）、最終的に自分が解きたいタスクにおいて最善の性能を発揮できればそれでいいということである。

そのようなものの例として、異常検知への応用がある [7, 10]。シナリオとしては、外れ値検出と、2 標本検定（データセット比較）という 2 つがある。いずれにしても、実世界のデータ解析で重要なことは、変数ごとに異常度が計算できなければならないという点である。実業務においては、異常検知の後に常に診断や回復といったステップが必要になる。たとえば 100 個のセンサーで

監視しているシステムで、「今異常が起きたっぽいです」とだけ言われても、現場のエンジニアは困ってしまう。

今、異常標本がまったく含まれていないか、いたとしても圧倒的少数であると信じられるデータ $\mathcal{D} = \{\boldsymbol{x}^{(n)} \mid n = 1, \ldots, N\}$ があるとする。疎構造学習を行うためには、まず、標本平均 $\hat{\boldsymbol{\mu}} = \boldsymbol{m} \equiv \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}^{(n)}$ を求め、新たなデータを $\mathcal{D}' = \{\boldsymbol{x}^{(n)} - \boldsymbol{m} \mid n = 1, \ldots, N\}$ のよう作っておき、$\mathcal{D}'$ に対して式 (9) を解く。その結果、ある ρ の値に対して、Λ という疎な精度行列が得られたとする。これを使うと、系から新たに出てくる標本 $\boldsymbol{x}$ の予測分布が、$\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \Lambda^{-1})$ のように与えられる。

第 i 番目の変数に異常度を定義するひとつの方法は、他の変数を与えたときのこの変数の条件つき分布 $p(x_i|\boldsymbol{x}_{-i})$ を使うことである。たとえば $i = 1$ なら、$\boldsymbol{x}_{-1}$ というのは $(x_2, \ldots, x_M)^\top$ の意味で、M は標本の次元である。正規分布の場合この条件つき分布を求めるのは簡単である。式 (5) と同様に、条件つき分布の定義から、$p(x_1|\boldsymbol{x}_{-1})$ は $p(x_1, \ldots, x_M)$ に比例している。正規分布の定義 (4) において、$\boldsymbol{\mu} = \boldsymbol{m}$ の下、x_1 に関係する項だけを拾うと

$$p(x_1 \mid \boldsymbol{x}_{-1}) \propto \exp\left\{-\frac{1}{2}\Lambda_{1,1}(x_1 - m_1)^2 - (x_1 - m_1)\sum_{j=2}^{M}\Lambda_{1,j}(x_j - m_j)\right\} \tag{23}$$

であるから、括弧の中身を平方完成して正規分布の定義と見比べることで、結局

$$p(x_1 \mid \boldsymbol{x}_{-1}) = \mathcal{N}\left(x_1 \mid u_1(\boldsymbol{x}_{-1}), \Lambda_{1,1}^{-1}\right) \tag{24}$$

$$u_1(\boldsymbol{x}_{-1}) = m_1 - \frac{1}{\Lambda_{1,1}}\sum_{j=2}^{M}\Lambda_{1,j}(x_j - m_j) \tag{25}$$

を得る。これを使うと、新たに観測した標本 $\boldsymbol{x}$ の、第 i 番目の変数の異常度（外れ値度）を

$$a_i(\boldsymbol{x}) \equiv -\ln p(x_i \mid \boldsymbol{x}_{-i}) \tag{26}$$

$$= \frac{1}{2}\Lambda_{i,i}\left\{x_i - u_i(\boldsymbol{x}_{-i})\right\}^2 - \frac{1}{2}\ln\frac{\Lambda_{i,i}}{2\pi} \tag{27}$$

のように定義できる。これが、古典的なホテリング T^2 スコア [9] の拡張になっていることは式の形から明らかであろう。異常度に適当な閾値を付せば、異常か正常かの判別ができる。もし既知の異常標本が含まれる検証用データが別にあれば、異常標本と正常標本それぞれに当たりと外れの個数がわかるので、異常検知器としての性能を最大化するように ρ（と閾値）を決めることもできる。

グラフィカル lasso に基づく異常検知の手法は、相当無茶なデータを食わせてもおかしな結果を出しにくいという意味において実用性が高く、逆に、よく「落ちる」ことで有名な古典手法としてのホテリングの T^2 法の有力な代替手法となりえる、と個人的には思う[*5]。

ただ、いくらグラフィカル lasso 自体が有用であったとしても、どうしようもない場合もある。典型的なのが、データが多峰性を持つ場合である。図 4 にそういう例を挙げる。これはある海洋石

[*5] ホテリングの T^2 法が普及している分野に、半導体製造装置の監視がある。10 年以上前、もちろんグラフィカル lasso の発明の前だが、IBM の半導体エンジニアから T^2 法についての怨嗟の声を山ほど聞いたことがある。

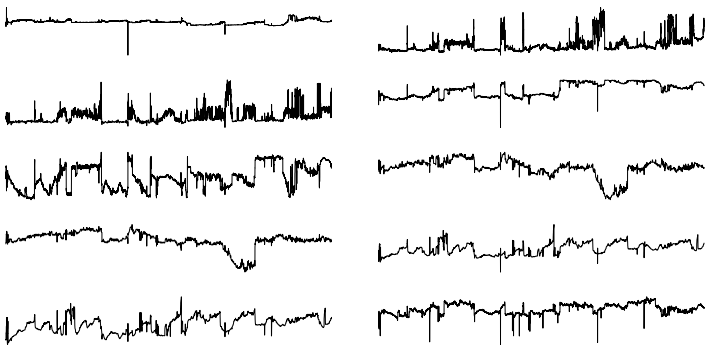

図 4　海洋石油採掘基地のセンサーデータ（シミュレーション）。

油採掘基地の圧縮機から得たデータであり、それぞれの測定値は、加速度は圧力といった物理量の測定値である。海底油田から採掘される石油は、水や砂などの混じった非一様の流体であり、絶え間ない圧力変動にさらされている。孤立した突発的な外れ値はスムージングで取り除けるとしても、問題は、系の動作が区分的に異なるように見えることである。これは単一の正規分布では捉えるのが難しい。

そこでひとつの考え方は、グラフィカル lasso を混合モデルに拡張することである。ただし、その代償としていくつか新たに解くべき技術課題が生ずる。ひとつは、混合モデルの混合数をどう決めるかという問題である。もうひとつは、混合モデルの学習は一般に凸最適化問題にはならないため、ある初期値から出発した反復算法を導入するのが通例であるが、その場合の初期化をどうするかという問題である。また、混合モデルにすると、どのように変数ごとの異常度を計算するかが単一ガウス分布とは違って複雑になる。これらの点についての詳細は最近の論文 [6] を参照されたい。

参考文献

[1] Onureena Banerjee, Laurent El Ghaoui, and Georges Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proc. Intl. Conf. Machine Learning*, pp. 89–96, 2006.

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

[3] A. P. Dempster. Covariance selection. *Biometrics*, Vol. 28, No. 1, pp. 157–175, 1972.

[4] Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pp. 604–612, 2010.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, Vol. 9, No. 3, pp. 432–441, 2008.

[6] Tsuyoshi Idé, Ankush Khandelwal, and Jayant Kalagnanam. Sparse gaussian markov ran-

dom field mixtures for anomaly detection. In *Proceedings of the 2016 IEEE International Conference on Data Mining (ICDM 16)*, p. TBD, 2016.

[7] Tsuyoshi Idé, Aurelie C. Lozano, Naoki Abe, and Yan Liu. Proximity-based anomaly detection using sparse structure learning. In *Proc. of 2009 SIAM International Conference on Data Mining (SDM 09)*, pp. 97–108, 2009.

[8] Tsuyoshi Idé, Dzung T. Phan, and Jayant Kalagnanam. Change detection using directional statistics. In *Proceedings of the Twenty-Fifrth International Joint Conference on Artificial Intelligence (IJCAI 16)*, pp. 1613–1619, 2016.

[9] 井手剛. 入門 機械学習による異常検知 – R による実践ガイド –. コロナ社, 2015.

[10] 井手剛, 杉山将. 異常検知と変化検知. 機械学習プロフェッショナルシリーズ. 講談社サイエンティフィク, 2015.

[11] 朝野熙彦, 鈴木督久, 小島隆矢. 入門 共分散構造分析の実際. KS 理工学専門書. 講談社サイエンティフィク, 2005.