データから知識へ：多変量情報流による潜在的機構の推定
From Data To Knowledge: Estimating Latent Mechanisms Through Multivariate Information Flows

日高　昇平 1)
HIDAKA Shohei[1)]

1)北陸先端科学技術大学院大学
1) Japan Advanced Institute of Science and Technology,

【要約】科学的活動の一つの目的は、観測された現象に対し、それを近似的に再現するより簡潔な表現形式(理論・モデル)を構築する事である。多くの理論構築は、現象の観察・経験を通じ、直観に沿う仮説を立てる事で始まる。この直観的な理論構築過程は暗黙的であり、この解明が知識獲得・創造を理解するための鍵である。理論構築過程の一つの定式化として、ある多変量時系列を付与のデータとし、それを生成する潜在的な概念モデルの推定問題を考える。本研究では、情報理論を一般化した多変量双方向情報理論を用いる事で、現象に関する事前知識性を必要とせずに、データのみから一般的な潜在的な概念モデルを推定できる事を示す。

【キーワード】知識獲得，仮説生成，情報理論，非線形力学系，時系列解析

1. 知識獲得過程における観察者の暗黙的計算過程のモデル化

科学的活動の一つの目的は、観察・観測された現象そのもの(データ)から、それを生成・近似する簡潔な表現形式(理論・モデル)を構築する事である。もし構築されたモデルが、何らかの意味で一般性をもち、有意義であるならば、それはコミュニティ(研究グループ・学会・社会など)の多くのメンバーに共有される知識となりえる。このようなモデル化・理論化の過程は、多くの場合、現象・データの観察・経験を通じ、観察者の“直観”に沿った仮説を立てることから始まる。我々が経験的にとるアプローチの一つは、概念モデルの構築である。概念モデルは、ある現象におけるいくつかの概念の関係に関する我々の直観を反映した関係グラフによって表される。概念モデルは、複雑な対象において、それの間の関係性を整理する一つの形式であり、多くの研究分野において、数理的モデルを構築する前段階として不可欠である。

現象の観測から数理モデルの構築までの知識獲得過程を図 1 の模式図に示す。現象の観測そのものから、それを分節化・符号化することにより、ある形式のデータとして現象を記述し、データに基づき、概念モデルあるいはより厳密な記述形式である数理モデルの構築へと進む。この模式図において、構築されたモデルあるいは理論は形式的(形式知)であるのに対し､それをデータに基づき生成する観察者の直観的過程は暗黙的(あるいは暗黙知)である。知識の獲得過程を理解するうえで、データを情報へ、さらに情報を形式知へと変換する暗黙的な認知過程の解明が重要である。

概念モデル(関係グラフ)の構築は、多くの場合、観察・分節化を通じた初めての言語化・形式化である。本研究では、データの観察から知識を構築する際の鍵となる、概念モデルの生成過程を説明する理論体系を提案する。一般的な定式化として、ある多変量データが与えられた場合に、それを生成する可能性の高いモデルを推定する問題を考える。図 2(a)は、 1 例として、3 つの変数がある場合に可能な因果関係の組み合わせ 16 通りを示している。図 2(a)の各グラフでは、変数を頂点とし、頂点間の方向つきの矢印である変数から他の変数への相互依存関係を示している(②→③は、変数 3 が変数 2 の状態によって(部分的に)決まる事を意味する)。組み合わせ 1 は 3 変数が独立であり、組み合わせ 16 は全ての変数が全ての変数と相互関係を持つ。この定式化では、ある現象に関する変数群の観測データから、その変数間の相互依存関係 (e.g., 組み合わせ 1～16)を特定することが、その現象に関する概念モデルを構築に相当する。本研究では、あるいくつかの変数に関する時間的変化を観測したとき、それを生成する可能性の高い潜在的な関係グラフを特定する問題を考える。これは、任意の N 変量時系列から、その N 変量の間の 2 項関係を N^2 の方向つき情報流として定量化する問題として定式化できる。こうして得られる情報流ネットワークは、観察された要因間の関連性の度合いに関する直観、あるいは「概念モデル的」な対象の特性付けと考える事ができる。

このような関係グラフによるデータの要約を、ある知識獲得過程における「概念モデルの生成過程」として考えるには、少なくとも以下の２つの要件を満たすべきである。

（１）少事前知識性：観察事象に対する最小限の事前知識で概念モデル的仮説を構成可能である。
（２）汎用性：特定の機構の生成するデータだけでなく、多様なデータクラスに適用可能である。
条件（１）は、概念モデルが、他のモデル・理論に関する事前知識に拠らず、データの観測のみで生成可能である事を要請する。観測されるデータの一貫性(i.e., データが途中で異なる観測データと混合していない、十分に密に記録しているなど)に関する最小限の事前知識のみを要するためこれを少事前知識性と呼ぶ。条件（２）は、このような概念モデル形成が、特定の問題のみならず、様々な対象について応用可能である事を要請する。これらの条件は、概念モデルの構築に先立つものが多くの場合現象の観測のみであり、また分野を問わず幅広い対象に対して概念モデルが提案されている経験的な事実を反映している。この２つの条件に対し、本研究では、（１）与えられた時系列データの観測変数の一貫性・同期性のみの前提に基づき、（２）多様な確率的・力学的なデータ生成過程に対し、その潜在する情報的メカニズムを推定可能な理論的枠組みを示す。具体的には、情報理論(Shannon, 1948)あるいはその拡張である双方向情報理論(Marko, 1973; Schreiber, 2000)を、さらに多変量へと一般化した双方向情報理論(Hidaka, 2012)を提案する。

多変量双方向情報理論の計算過程は、典型的な科学的方法論を抽象化したものとみなせる。多くの場合、科学者はある２つの因子を、他の因子の影響をなるべく統制・固定した上で調べ、その２因子間の関係性を見極める。これと同様に、多変量双方向情報理論では、ある変数 X から変数 Y への情報の流れを、その他の変数 Z の影響を差し引いた上で(条件つきで)推定する(Hidaka, 2012)。一方、(二変量)双方向情報理論(Marko, 1973)では、第三の変数 Z が存在していても、その影響を統制せずに、ある変数 X から変数 Y への情報の流れを(Z の影響を含みながら)推定する。従って、多変量双方向情報理論によって、単に情報量の算出というよりも、複数変数間の関係性を考慮した上でのデータの持つ情報的特性を定量化する事ができる。本研究では、知的営みにおいて重要な過程の一つである概念モデルの構築は、こうした変数間の関係性を考慮した上での情報量の推定とみなせると仮説をたて、これを検討する。次節以降では、まず双方向多変量情報理論の概略を示し、次にそれを用いたシミュレーションおよび実データの解析の結果を示す。最後に、双方向多変量情報理論の分析手法としての技術的実用性および、知識獲得過程の計算過程について考察を行う。

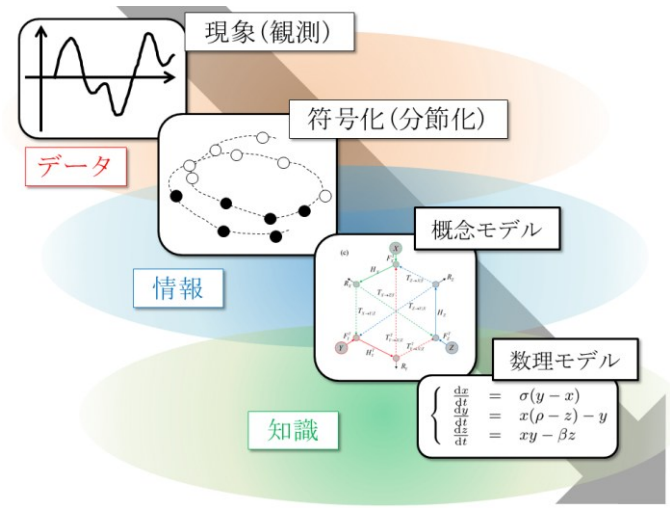

図 1: 知識獲得過程の概念モデル

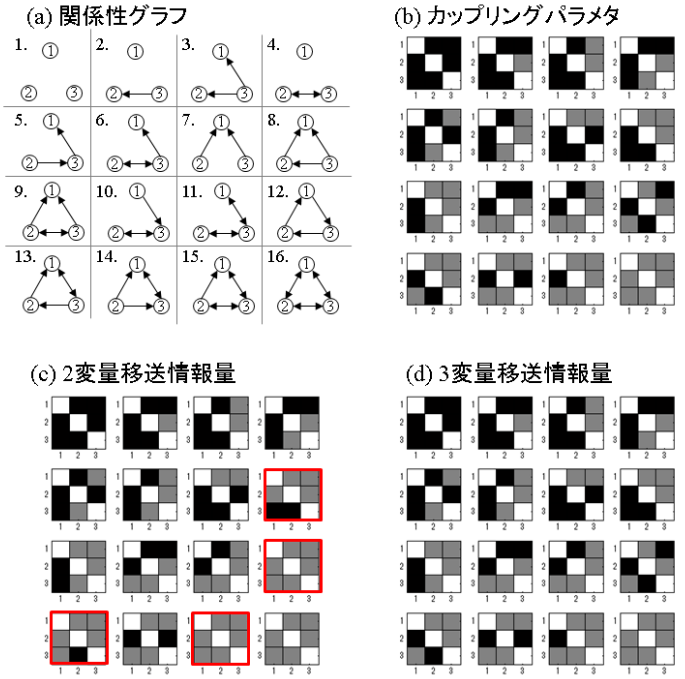

図 2: (a) 関係性グラフ(3 変数の場合、全 16 組み合わせ), (b) 結合写像格子の結合パラメータ, (c) 二変量移送情報量の推定値(非対角要素), (d) 三変量移送情報量の推定値(非対角要素)

2. 情報理論から多変量双方向情報理論までの概説

本節では、情報理論(Shannon, 1948)および双方向情報理論(Marko, 1973)を一般化した多変量双方向情報理論(Hidaka, 2012)について概説する(詳細は、Hidaka (2012)を参照)。情報理論は、Shannon (1948)の定式化後、通信分野のみならず、物理学、数学、統計学など様々な分野に広く用いられてきた。その骨子は、ある事象の生起に関する確率的なばらつきを定量化する(情報理論的)エントロピーと、二つの事象の確率的な関連性を定量化する相互情報量の二つに集約される。エントロピーは、ある確率分布に従って生起する記号列をもれなく符号化するのに必要な最小の符号長である。記号列がある規則的に従っているとき短く符号化でき(e.g. “00000000”を”0×8”と短く符号化)、逆に記号列に規則性が無いときに最も長い符号が必要である(e.g., “00101110”は短いルールで符号化しにくい)。従って、エントロピーは、情報圧縮、暗号理論などにおいて、符号化可能性の限界を表す。一方、相互情報量は、２つの記号列 X と Y の間で、一方の記号列が他方に対して有する情報を定量化する。相互情報量は記号列 X と Y が同一であるとき最大であり(X が与えられると Y が完全に予測できる)、記号列 X と Y が確率的独立であるときに最小の 0 をとる(X が与えられても Y の予測に全く寄与しない)。相互情報量は、ある確率分布に従って作成されるメッセージ(記号列 X)を、ある通信手段を使って伝送したときに得られるメッセージ(記号列 Y)の間の誤差の最小限界を示しており、情報通信の分野において重要な役割を果たす。

Shannon の定式化した情報理論は、フィードバックの存在しない一方向のみへの情報を扱う。一方、Shannon 自身はフィードバックのある双方向の情報理論への拡張の重要性を述べていた(Massey, 1990)。このような背景を踏まえ、Marko (1973)は、２つの情報源 X と Y が相互に影響を受ける双方向通信問題を定式化し、双方向情報理論を提唱した。図 3(a)は Shannon の情報理論を情報の流れのネットワークとして図示したものである。情報源 X のもつエントロピー(H(X))がある通信路を通って、Y に変換された場合、そのエントロピーH(Y)と X が与えられたときの条件つきエントロピーH(Y|X)の差 I(X; Y)が相互情報量(通信可能な情報の限界)となる。双方向情報量は、相互情報量を X から Y へあるいは Y から X への情報の流れに分解したものである(図 3(b))。特に、双方情報理論(図 3(b))において、X から Y(Y から X)に伝わる情報量を X から Y への移送情報量(Transfer Entropy: TE, 図 3(b)中の $T_{X\to Y}$)と呼ぶ(Schreiber, 2000)。二方向の移送情報量の和が相互情報量となる事から、移送情報量は方向を明示して両者の依存関係を定量化したものと捉えることが出来る。

三変量以上では、二変量関係の重ね合わせに帰着できない高次の従属関係が存在する。従って、二変量間の双方向情報量を、多変量関係に応用する事は理論的に適切ではない。そこで、Hidaka (2012)は任意な数の変数の間の双方向情報量を定量化する多変量双方向情報理論を提案した。図 3c は三変量の場合の双方向情報ネットワークを示している。三変量双方向情報量 $T_{X\to Y|Z}$ は、Z をある条件で固定したときの X から Y への移

送情報量を表し、X、Y の二変量関係に関して第三変量の効果を割り引いたものと解釈できる。多変量移送エントロピーはこの考えを任意の数の変量に拡張したもの(図 3d)で、理論上、高次情報量(Watanabe, 1960; Garner, 1962; Studeny & Vejnarova, 1999)を非負の情報の流れとして分解したものと解釈できる。多変量双方向情報理論は情報理論(Shannon, 1948)、(二変量)双方向情報理論(Marko, 1973)を特殊な場合として含み、それらの一般化理論である。

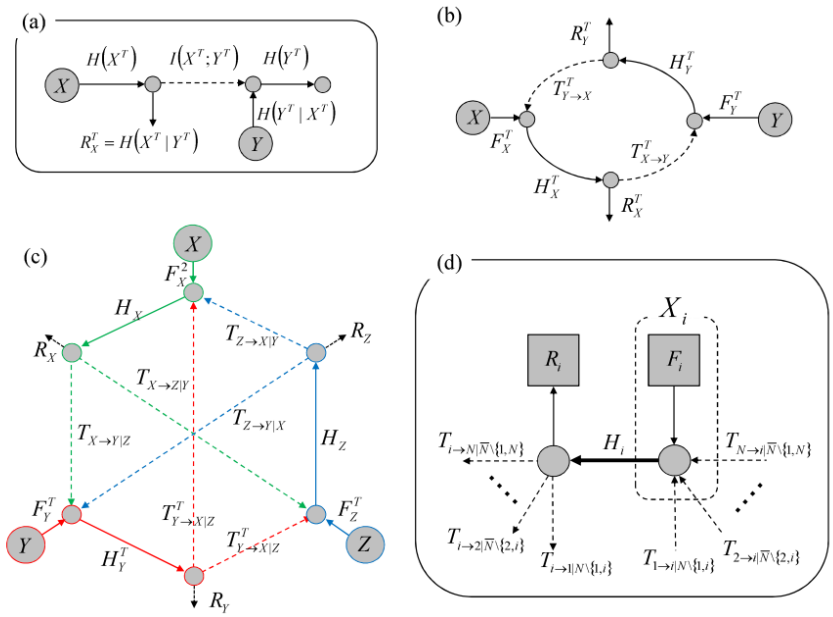

図 3: (a) Shannon の情報理論，(b)Marko の双方向情報理論，(c)3 変量に一般化した双方向情報理論(Hidaka, 2013), (d)任意数の変数の双方向情報理論の一部(Hidaka, 2013)。

2.1 多変量双方向情報理論の数学的定式化

T の長さをもつ N 変量の時系列(変数の添え字の集合を $\eta=\{1,2,\cdots,N\}$、時間ステップの添え字集合を $\tau=\{1,2,\cdots,T\}$ とする)に対し、$X_\eta^\tau=\{X_1^1,X_2^1,\cdots,X_N^1,X_1^2,X_2^2,\cdots,X_N^2,X_1^T,X_2^T,\cdots,X_N^T\}$ を各時点の変量の確率変数の集合とする。このとき、変数 j から i への移送情報量 $T_{j\to i|\eta\setminus\{i,j\}}$ は以下の式で定義される。

$$T_{j\to i|\eta\setminus\{i,j\}}\equiv\sum_{t=1}^{T}I\left(X_i^t;X_j^{\bar{t}\setminus t}\mid X_{\eta\setminus j}^{\bar{t}\setminus t}\right)\quad(\text{式 1})$$

ただし、$\bar{t}=\{1,2,\cdots t\}$ は t までの累積の添え字集合、$\bar{t}\setminus t$ は t を除く(t-1)までの累積の添え字集合、$I(X;Y|Z)=I(X|Z)-I(X|Y,Z)$ は条件付相互情報量である。移送情報量は以下に示す分解性・ネットワーク性という 2 つの性質を持つ(Hidaka, 2013)。

2.1.1 高次情報量の分解定理

N 変数の全ての(方向つき)対に関して移送情報量を和をとると、N 変数高次情報量以下になる。すなわち、

$$C\left(X_\eta^\pi\right)=\sum_{i=1}^{N}\sum_{j=1}^{i-1}G_{ij}\quad(\text{式 2})$$

ただし、$C\left(X_\eta^\pi\right)=\sum_{i=1}^{N}H\left(X_i^\tau\right)-H\left(X_\eta^\tau\right)$ は total correlation (Watanabe, 1960; Garner, 1962; Studeny & Vejnarova, 1999)と呼ばれ、N 変数の高次情報量を表し、$G_{ij}=T_{i\to j|\eta\setminus\{i,j\}}+T_{j\to i|\eta\setminus\{i,j\}}+R_{ij}^T$ は、変数 i から j, j から i への移送情報量と非負の余剰エントロピー $R_{ij}^T=\sum_{t=1}^{T}I\left(X_i^t;X_i^t\mid X_{\mathrm{N}\setminus\{i,j\}}^t\right)$ の和を表す。式2 両辺の個々の項は非負なので、明らかに分解定理が成り立つ。

2.1.2 ネットワーク性定理

移送情報量は、N 変量間の情報の流れとみなせる。すなわち、図 3(d)に示す N 変数ネットワーク上の全ての流れは非負であり、各頂点において入力の総和は出力の総和に等しい(キルヒホッフ則)。具体的には任意の変数 i に対する入力の和は $H_i=F_i+\sum_{j\in\eta\setminus i}T_{j\to i|\eta\setminus\{i,j\}}$ であり、出力の和は $H_i=H\left(X_i^\tau\mid X_{\eta\setminus i}^\tau\right)+\sum_{j\in\eta\setminus i}G_{ij}$ である。ただし、$H_i^T\equiv\sum_{t=1}^{T}H\left(X_i^t\mid X_i^{\bar{t}\setminus t}\right)$ は過去の状態($\bar{t}\setminus t$)から時点 t までのエントロピーレート(Cover & Thomas, 1991)で，$F_i^T\equiv\sum_{t=1}^{T}H\left(X_i^t\mid X_{\mathrm{N}}^{\bar{t}\setminus t}\right)$ はフリーエントロピー(Marko, 1973)である。

3. 多変量情報理論による情報論的な関係グラフの推定：非線形力学系の場合

本節では、具体的に多変量情報理論をデータ分析に応用する事で、推定される関係グラフを調べる。その一つの方法として、まず因果関係を操作することの出来る理論的な対象に対し、その潜在的な因果関係が未知な場合に双方向多変量情報量のみでどの程度正確に関係グラフが推定可能であるかシミュレーションを行った。因果関係が決定論的に決まり、しかし、長期的な予測が困難な複雑な対象として、非線形力学系が挙げられる。非線形力学系では、ある時点から次の時点への状態変化が、ある非線形方程式(連続系では微分方程式)によって与えられる。従って、変数間の関係はある方程式によって決定的に与えられるが、多くの場合、その挙動は複雑でカオスと呼ばれる。このシミュレーションでは、ある複数変数のカオス的挙動の系列を観測した場合に、その経験的な時系列から、系に固有の変数間の関係を情報の流れとして記述する。これは、まさに複雑な多変量関係を観測したときに、その変数間の関係性を特徴づける概念モデルの構築に対応する。シミュレーションでは、“正しい”概念モデルを非線形力学系のパラメータとして操作し、多変量情報理論によってその潜在的な関係性をどの程度正確に推定できるか評価した。

具体的な系として、結合写像格子(Coupled Map Lattice: CML)を用いた。CML は比較的単純な 1 次元力学系を結合し、個々の要素の関係性を任意に操作可能であり、かつ個々の要素だけでは起きない創発的な特性を全体として示す(Kaneko, 1992)。この性質のため CML は、地震(Ceva, 1995), ニューロンの形態(Sakaguchi & Ohtaki, 1999), 交通(Yukawa & Kikuchi, 1995), 対流(Yanagita & Kaneko, 1993), 細胞遺伝子間相互作用(Bignone, 1993), 癲癇性発作（Larter, Speelman, & Worth, 1999）などの諸現象の抽象モデルとして用いられてきた。本シミュレーションでは、3 から 5 個のテント写像を結合した結合写像格子を用いて、時系列データを生成し、それに多変量情報流を適用する事で、元の力学系における変数間の関係性を特定した。具体的に、以下の式で表される結合テント写像格子を用いた。

$$x_i^{t+1} = f\left(\frac{x_i^t + \varepsilon\sum_{j\neq i}\delta_{ij}x_j^t + \eta_i^t}{1+\varepsilon\sum_{j\neq i}\delta_{ij} + \eta_i^t}\right) \qquad \text{(式 3)}$$

式 3 において、$0<x_i^t<1$ は変数 i の時間 t における状態を表す実数値, δ_{ij} は二値(0 または 1)で変数 j から変数 i への結合の有無、εは非負実数値で結合の強さを表すパラメタで、$f(x)=2x$ ($x<1/2$)または$f(x)=2-2x$ ($x\geqq 1/2$)はテント写像ある。ただし全ての変数 i について$\delta_{ij}=1$ に固定した。結合 i→j は x の過去の状態が y の未来の状態への影響を及ぼす事を意味する。図 2(b)は 3 変数の場合の変数間の方向つき結合を表す行列を図 2(a)の各関係グラフに対応して示している。図 2(b)において、各セルの色が結合の強さを表し、それぞれ白色=1, 灰色=ε, 黒色=0 である。従って、結合あるいは結合の強さを設定することで、CML では変数間の相互関係の強さを系統的に操作可能である。こうして、あるパラメタを設定された式 3 を真の因果関係とし、それから生成された多変量時系列をデータとして得られたと想定する。このとき観測データに対して、どのような推論過程が働けば、変数間の関係性グラフ(概念モデル)を構築する事が可能であるか検討する。具体的に、本シミュレーションでは、多変量移送情報量の計算が、関係性グラフの構築に対応すると仮説を立てた。従って、潜在的な真の因果関係(式 3 およびそのパラメタ)に対し、データから算出される多変量移送情報量が、どの程度精度よくそれを推定可能か分析した。比較の対象として、変数対関係のみを考慮する二変量移送情報量も同様にその推定精度を検討した。

3.1 シミュレーションの手続き

0 から 1 までの一様擬似乱数$(0, 1)^N$により発生させた初期値を用い、式 3 の CML により 10^5 点の N 変量時系列に対し、N 変量移送情報量(式 1)の推定を行った。推定では、各変数の値を x>1/2 で二値化し、4 次マルコフ性を仮定した上で、その遷移確率分布の N 変量移送情報量を算出した。比較のため、高次依存関係を考慮しない二変量移送情報量も同様の方法で算出を行った。もし真の移送情報量が 0 である場合、有限データから推定された移送情報量は、サンプルサイズ・変数の数に依存したパラメタを持つガンマ分布に従う(Goebel, Dawy, Hagenauer, & Mueller, 2005)。これを利用し、推定された移送情報量を 0 とする帰無仮説に対し統計検定を行うことで、α = 0.01 水準で有意に 0 より大きい移送情報量を求め、これを関係グラフにおける情報の流れとして定義した。

3.2 結果・考察

まず、3 変数 CML において結合強度をε=0.2 とし、式 3 により可能なすべての結合$\{\delta_{ij}\}$の 16 の組み合わせ(図 2(b))について時系列データを生成した。この時系列データに対して三変量移送情報量を推定した結果を図 2(d)に示す。図 2(c)には、比較対象として、第 3 変数を考慮せずに変数対の情報量を推定する二変量移送情報量の結果を示す。推定結果は、データに潜在する真の因果関係(図 2(b))に近いほど良い。ここでは、移送情報量 0 に対する統計検定(α = 0.01 水準)を行い、真の関係グラフにおいて正である関係を 0 と推定した場合、あるいは真の関係グラフにおいて 0 である関係を 0 より有意に大きいと推定した場合を､推定誤りと定義する。16 の各関係グラフについて推定結果のそれぞれについて、推定誤りを 1 つでも含むケースを図 2c,d で赤枠で示している。この結果は、多変量移送情報量によって、真の関係グラフを誤り無く推定できた事を示している。一方、比較基準として行った二変量移送情報量を用いた場合、16 のケースのうち、4 ケースで関係性の推定に誤りが見られた。この結果は、第３変量の効果を考慮して変数対の従属性を定量化する多変量移送情報量によって、高精度の関係グラフ構築が可能である事を示唆している。

次に、3 変数の場合と同様の手続きに従って、変数の数を 4,5 変数に増やしてシミュレーションを行った。変数の数が増えると、潜在的に可能な関係グラフの組み合わせは指数的(より正確には変数 N の $O(\exp(N^2)$のオーダ)に増える。従って、その組み合わせ爆発に伴って、関係グラフの推定は非常に困難になる。表 1 は、３，４，５変数の間で可能な方向つき依存関係の組み合わせ数(変数の交換に対し対称なものを除く)を示している。対称性を考慮した依存関係は 3, 4, 5 変数と増えると、16, 218, 9605 と膨大に増え、また可能なグラフはの組み合わせは 2^6, 2^{12}, 2^{20} と増加する。これは、例えば、5 変数の関係グラフをでたらめに推量した場合、それが偶然正しい確率は 2^{-20} である事を意味している。変数の増加に伴い、一般に推定はより不正確になる。従って、現実的なデータに対してある程度の有用性を保障するためには、変数の数を増やした場合の推定精度の低下を見積もる必要がある。

表 1 は、変数の数を 4,5 と増やした場合に、全ての組み合わせ(それぞれ 218, 9608 ケース)の変数対ごと、そしてケースごとの正解率を示している。ケースごとの正答率は個々のケースで 1 つでも推定誤りが含まれる場合は不正解としているため、変数対毎の正答率より厳しい基準となっている。この結果は、多変量移送情報量はケースごとで 90.78%(4 変数), 81.48%(5 変数)、変数対ベースで 97.78%(4 変数), 95.41%(5 変数)と比較的高い推定精度を保つ事を示している。一方、比較基準となる二変量移送情報量はケースごとで 9.22%(4 変数), 1.24%(5 変数)、変数対ベースで 70.67%(4 変数), 71.23%(5 変数)と、変数が増えると急激に推定精度が下がる事を示している。この結果は、変数の増加に対し、多変量移送情報量は比較的高い推定精度を保つ事を示し、より高次な依存関係に対する堅牢性を示唆している。これに対し、二変量移送情報量がこの堅牢性を持たない事、また両者の理論的性質の違いを考慮すると、関係グラフの推定において、（1）情報量の推定に加えて（2）その高次の従属関係(交絡関係)を明確に分離する(条件わけする事)の２点が要点として挙げられる。

表 1：CML 多変量データに基づく関係グラフの推定結果

問題設定				正解率 多変量		正解率 二変量	
変数の数	関係対組合わせ	ケース数	関係対の数	ケースベース	変数対ベース	ケースベース	変数対ベース
3	2^6	16	96	100%	100%	75.0%	93.75%
4	2^{12}	218	2,616	90.78%	97.78%	9.22%	70.67%
5	2^{20}	9,608	192,160	81.48%	95.41%	1.24%	71.23%

4. 実データへの応用：身体運動の情報流ネットワーク

次に、実際の経験的なデータに対する多変量双方向情報理論の適用を行い、概念モデルの形成過程を検討する。複雑な要因間の相互作用のある対象の一つとして、身体運動の分析を行った。身体運動は環境、身体、神経活動の三者の間の複雑な相互作用によって成立する。複雑な多関節運動の場合、身体各部はそれぞれに協調しながら一つの運動を生成すると考えられている(Yamamoto, Ishikawa, & Fujinami, 2006)。ある一つの運動の習得は、必ずしも身体部位間の静的な関係性の獲得を意味しない。むしろ、熟達に伴って、運動はより効率的に、より効果的に変化すると考えられる。こうした熟達した運動の獲得をスキル学習と呼ぶ(Leonard, 2002)。しかし、スキルと呼ばれる複雑な運動に関して、身体運動の熟達をどのように捉えるべきかは個別の課題に特化してケーススタディとして行われており、未だに統一的な理論的解釈は与えられていない。たとえば、球技(e.g., バスケットボール)のように、明確に運動の標的が定義されている特定の運動(ゴールにボール投げ込む)の場合、その成功率として習熟を一つの尺度で表す事

が可能である。しかし、こうした明確な標的が定まらない、たとえば演劇、歌唱や音楽演奏、描画など運動においても、人はその運動の質(あるいはその結果としてのパフォーマンス)をある程度一貫して評価する事が出来る。こうした背景を踏まえ、複数の身体部位間の協調関係を反映する概念モデルをデータから推定し、熟達化に伴う概念モデルの変化を検討する。

4.1 分析手続き

対象としたのは、サンバの演奏運動中の身体動作をキャプチャしたデータである(Yamamoto, Ishikawa, & Fujinami, 2006)。この実験では全身の 16 の特徴点と楽器(シェイカー)の両端２箇所の計 18 箇所にマーカーをつけた演奏家 5 名が、5 つのテンポ(60, 75, 90, 105, 120 拍毎分(BPM))で演奏を数分間を行った。こうして得られたデータのうち、本研究では、特に熟達に顕著な差があると見られる３名(40 年以上の経験を持つ上級者, 5 年の経験をもつ上・中級者、2 年の経験をもつ中級者)の演奏に直接関わる右腕(肘・手首)と楽器(両端 2 箇所)の 4 箇所の動きを分析した。各演奏者につき、4 箇所の身体動作は 3 次元の空間上の点列として 86.1Hz のレートでサンプルされたが、計測に伴う高周波ノイズのため、その半分のデータ点(43.05Hz に相当)を分析に用いられた。各時系列データは、演奏者、演奏リズムごとに記号的最近傍法(Buhl & Kennel, 2005)を用いて、符号化をした後、4 次マルコフ性を仮定した上で、その遷移確率分布の N 変量移送情報量を算出した。前述したシミュレーションの分析と同様に、推定された移送情報量を 0 とする帰無仮説に対し統計検定を行うことで、$\alpha = 0.01$ 水準で有意に 0 より大きい移送情報量を求め、これを関係グラフにおける情報の流れとして定義した。

4.2 結果・考察

経験年数の違いから期待される熟達の度合いに応じて、上級者、中上級者、中級者とした。リズム運動の多変量時系列(楽器 1, 2, 手首、肘の 4 箇所)から推定された情報流グラフを図 4 に示す。それぞれの演奏家に対し、5 つの演奏テンポ条件、12 関係対について、有意に 0 より大きい情報流の割合 (12 ペア・5 条件、のべ 60 ペア中)をバーグラフに、その条件間の標準偏差をエラーバーとして表示している。熟達の度合いの高いほどに、演奏に関わる右腕および楽器に間における有意な情報流の割合が高い事が分かった。また、バーグラフの上部には、それぞれの演奏家で 5 条件すべてで有意に 0 より大きい身体部位間の関係を方向つきの関係グラフとして示している。推定された関係グラフは、上級者ではほとんど全ての身体部位間で情報の流れがあるのに対し、中級者は手首と楽器の一部のみに有意な情報がある事を意味している。また、中上級者では、中級者と上級者と中間程度の情報の流れが見られた。この中上級者の関係グラフでは、肘から手首、手首から楽器への情報の流れがあり、これは腕のしなりによるむち運動で楽器に動きが伝達する状態を表していると考えられる。これに加えて上級者では楽器から手首・肘への情報の流れも見られ、楽器の状況によって手首・肘の運動を変化させていると解釈できる。以上の結果は、上級者ほど、全ての関連する身体・楽器の間で双方向に情報伝達が起こり、腕から楽器に運動が伝わるだけではなく、楽器から腕への情報がその制御に重要である事が示された。

こうした結果は、予想された通り、熟達によって身体運動が巧妙になる事を示すだけでなく、多変量情報理論によってその具体的な関係性を構築できた事を意味する。本実験で調べた複雑な動きの精妙な制御は、未だに解明すべき部分が多い。数十ミリ秒の精度で多数の間接・筋肉が制御され、作り出された一つの洗練された動きに、我々はそれをどのように、熟達を感じ取るのだろうか。本分析は、このような問いに対して、一つの可能性を示している。つまり、我々が他者の身体運動の観察を通じて、情報的な関係性を推定し、その情報流ネットワーク上の密度として、運動の精妙さを感じ取るのではないだろうか。本研究は、ひとつの事例研究に過ぎないが、多変量双方向情報理論はこうした複数の要因間の複雑な時間的なパタンから、それを整理する有用な関係グラフを構築できることを示している。

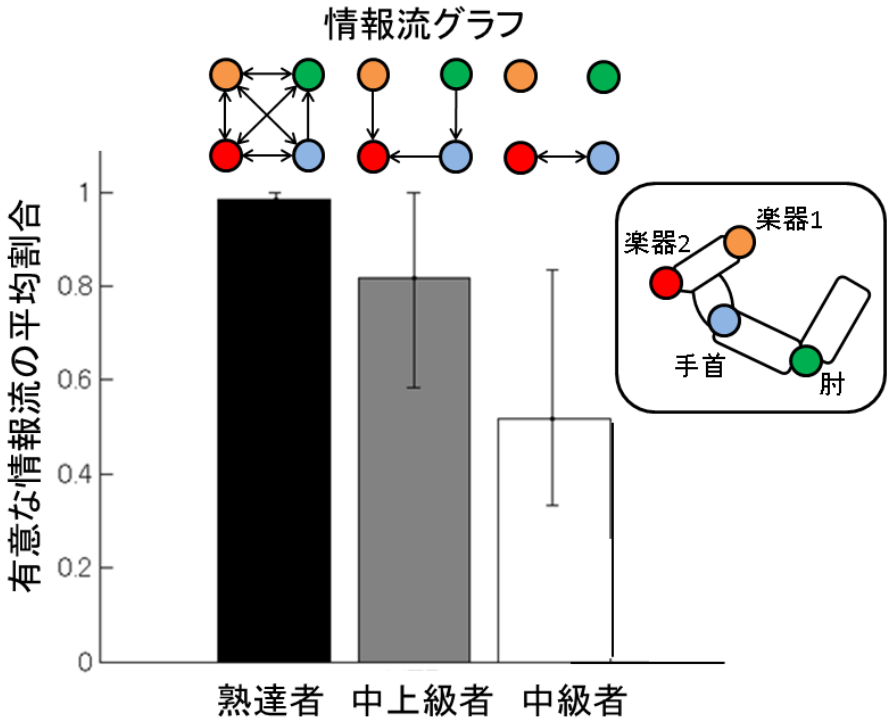

図 4: リズム運動の多変量時系列(楽器 1, 2, 手首、肘の 4 箇所)の情報流グラフ

5. 総合討議

本研究では、データから知識を獲得する過程において、最も初期に行われる概念モデルの形成について理論的検討を行った。概念モデルでは、ある現象に関してそれを適切に代表するいくつかの状態の分節化、そしてそれらの間の関係性の記述を行う。この概念モデルはそれに続くより詳細な数理モデルなどの基礎となり、あるいは社会科学など数理的なモデル化が困難な複雑な現象の説明においては中心的な役割を果たす。一方、概念モデルの形成過程そのものは、データ観察者の暗黙知であり、最終的な表現である関係グラフを除いてその内省的メカニズムは不明な点が多い。本研究では、多変量時系列データの観察から、その変量間の情報論的な関係グラフを推定する過程と概念モデルの構築とみなし、これを最小限の事前知識により行う理論体系を提案した。

本研究で検討した多変量双方向情報理論(MTE)は、Shannon の(一方向)情報理論、Marko の二変量双方向情報理論を拡張した一般化理論である(Hidaka, 2013)。MTE では、複数の変量間の関係(高次情報量)を方向つきの情報の流れに分解して表現する。MTE の計算過程は、典型的な科学的な思考法と同様に、他の交絡因子を統制した上で、二変数 X と Y の関係を特定する事を情報理論の拡張として表現したとみなせる。

理論モデルを用いてデータを生成したシミュレーションでは、第三の交絡因子の関与を考慮しない PTE よりも、MTE を用いた場合に高精度で関係グラフの構築が可能である事が示された。これに続く身体運動データの分析では、熟達者の運動と中級者の運動の違いが、身体部位間の依存関係の密度という形で、定量化できることが明らかになった。分析では、一切の身体部位間に関する事前知識を与えず、データだけから具体的な関係グラフの図式化したにも関わらず、熟達者の運動がより精密かつ複雑な構造を持つ、という我々の直観に合致した関係グラフが得られた。こうした一連の結果は、MTE による関係グラフ構築が、概念モデルの構築に類似した性質を持っていることを示唆している。これは、序論で提示した少ない事前知識の下で関係グラフを構築という条件（1）を満たすことを意味する。さらに、3,4,5 変数の全ての可能な関係グラフの組み合わせを分析したシミュレーションの結果と、さらにノイズなど必ずしも理想的ではない条件下で得られた身体運動データの分析結果を考慮すると、MTE は少なくとも一定程度の汎用性（条件（2））を有すると考えられる。Hidaka (2012)は、本研究で検討した 2 つのケースに加えて、さらに、理論力学系の一つである Lorenz 系への適用、また生理学的データ分析への応用も示している。この二条件については、今後も MTE による分析を理論的・経験的な側面から検証を重ねる事で、どの程度の事前知識が必要であるか、また汎用性を持つかを調べていく必要がある。

なぜ MTE による情報流ネットワークは、その背景にある(数理的)モデルを推定せずに、変数の間の従属性を定量化できるのだろうか。通常はある数理的モデルを立てた上で、そのモデルの当てはめを行った上で、関係性を特定する事を考えれば、潜在する数理メカニズムの特定前に、その関係性を定量化できるのは不思議ではある。このように数理モデルを特定せずに、変数間の情報量を特定できるのは、情報理論の持つモデルフリー性にある。情報理論において、モデルはある現象の符号化に関わる。ある空間を十分に高い精度で符号化を行った場合に、どのような座標系で符号化しても情報量が不変になる場合に、モデルフリーと言う。多変量情報理論の特殊な場合である二変量移送情報量に関して、このモデルフリー性が成り立つ事が示されている(Kaiser & Schreiber, 2002)。この理論的性質により、ある現象をどのように符号化するか、という詳細に立ち入らずに、多変量双方向情報理論は対象の従属性を定量化することができる。つまり、少事前知識性・汎用性を同時に満たしながら、データから概念モデル的な構造を推定しており、これは概念モデル形成過程を説明するモデルになりうる。

5.1 結語

本研究は、知識獲得の最初期に行われる、現象の観察から関連する変数の洗い出しおよびその関係性の特定という関係グラフ構築の段階についての理論的なモデルを提案した。モデルフリー性を持ち、第三変数の関与を統制しながら二変数間の双方向性の関係を定量化する多変量双方向情報量は、少事前知識性を満たしながら汎用的な問題の関係グラフを構築できることを示した。今後はこうした理論上好ましい性質を持つ多変量双方向情報理論が、実際の人の知識獲得過程をどの程度説明可能かか心理実験的手法などにより検討をしていく。

謝辞

本稿第 4 節の実データ分析では、藤波努氏の好意により、身体運動のデータを提供していただいた。ここに感謝いたします。本研究の一部は科学研究費補助金(No.23300099)、人工知能研究振興財団、ニューロクリアティブ研究会研究助成の補助を受けた。

参考文献

Bignone, F. A. (1993). Cells-gene interactions simulation on a coupled map lattice. Journal of Theoretical Biology, 161(2), 231 - 249.

Buhl, M. and Kennel, M. B. (2005). Statistically relaxing to generating partitions for observed time-series data., Phys. Rev. E 71, 046213.

Ceva, H. (1995). Influence of defects in a coupled map lattice modeling earthquakes. Phys. Rev. E, 52, 154–158.

Cover, T. M., & Thomas, J. A. (1991). Elements of Information Theory. Wiley-Interscience.

Garner, W. R. (1962). Uncertainty and Structure as Psychological Concepts. New York: NY: JohnWiley & Sons.

Goebel, B, Dawy, Z, Hagenauer, J, & Mueller, J. (2005). An approximation to the distribution of finite sample size mutual information estimates. Vol. 2, Communications, 2005. ICC 2005. 2005 IEEE International Conference on pp. 1102 – 1106.

Hidaka, S. (2012). Characterizing Multivariate Information Flows, eprint arXiv:1212.5449

Kaiser, A. & Schreiber, T. (2002). Information transfer in continuous processes. Physica D 166: 42-62.

Kaneko, K. (1992). Overview of coupled map lattices. Chaos, 2, 279-282.

Kantz, H., & Schreiber, T. (1997). Nonlinear time series analysis. Cambridge, UK: Cambridge University Press.

Larter, R., Speelman, B., & Worth, R. M. (1999). A coupled ordinary differential equation lattice model for the simulation of epileptic seizures. Chaos, 9, 795-804.

Leonard, C. T. (1997). The Neuroscience of Human Movement. Mosby: St. Louis, MO (松村道一，森谷敏夫，小田伸午(訳) ヒトの動きの神経科学).

Marko, H. (1973). The bidiretional communication theory - a generalization of information theory. IEEE Transaction on Communication, COM-21(12), 1345-1351.

Massey, J. L. (1990). Causality, feedback and directed information. In Proceedings of the international symposium on information theory and its applications.

Sakaguchi, H., & Ohtaki, M. (1999). A coupled map lattice model for dendritic patterns. Physica A: Statistical Mechanics and its Applications, 272(3-4), 300-313.

Schreiber, T. (2000, Jul). Measuring information transfer. Phys. Rev. Lett., 85(2), 461–464.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423, 623-656.

Studen´y, J., M & Vejnarov´a. (1999). The multiinformation function as a tool for measuring stochastic dependence. Cambridge, MA: MIT Press.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. IBM J. Res. Dev., 4(1), 66–82.

Yamamoto, Y., Ishikawa, K., & Fujinami, T. (2006). Developmental stages of musical skill of samba. Journal of biomechanics, 39, S555.

Yanagita, T., & Kaneko, K. (1993). Coupled map lattice model for convection. Physics Letters A, 175(6), 415-420.

Yukawa, S., & Kikuchi, M. (1995). Coupled-map modeling of one-dimensional traffic flow. Journal of the Physical Society of Japan, 64(1), 35-38.

連絡先
住所：〒923-1292 石川県能美市旭台 1-1 北陸先端科学技術大学院大学
名前：日高　昇平
E-mail：shhidaka@jaist.ac.jp