「解析法」セッションのイントロダクション

統計数理研究所　伊庭幸人 [1]

1　「解析法」セッションの趣旨

位相空間における分布の形やその動的な変化をシミュレーションデータから記述・解析することは簡単ではない。従来、モンテカルロ法などの解析では、物理的な議論から考えられた少数の物理量を測定して、それらを解析するという手法が行われてきた。これに対して、多数の物理量を測定して、そこから半自動的にシステムの特徴を抽出する という方法が考えられる。具体的には、

- スピングラスのクラスターのシミュレーションデータからの検出
- 有限温度の準安定状態の自動検出
- シミュレーションデータからの近似的な緩和モードの計算

などがその例である。このセッションでは、この種の方法論についてまとまった発表と討議を行う。

2　高次元空間のものを表現する

高次元の空間のものを表現する一般的な方法は，それほど多いわけではない．

2.1　「空間」配置型のモデル

ひとつの大きなグループとしては，高次元のデータをある方向に射影したり，それらの間の距離を保つように配置したりする方法がある．広い意味の「三面図」や「地図」のようなものである．

このうち，多変量解析でもっとも基本的なのは**主成分分析** (Principle Component Analysis) の手法である．これは，分散共分散行列（または相関行列）の固有値分解を行い，大きい固有値に属する固有ベクトルの張る空間への射影でデータを表現する．物理量の共分散行列は感受率の行列ということになるので，この手法は,「シミュレーションで求めた感受率の行列を対角化して固有モードを求める」と表現できる．特に，固有モードが局在してい

[1] E-mail: iba@ism.ac.jp

る場合はある種の「クラスター」と解釈できる可能性がある．本研究会での**根本**氏の発表は，スピングラス研究においてこの種の方法がどのように応用できるかを調べたものである[2]．この種の手法はたんぱく質のシミュレーションでの有限温度での形状や運動の解析にもすでに用いられている [1].

「地図」を作る方法と考えた場合，データポイントの間のなんらかの「距離」を与えて，それをなるべく変えないように点を配置するというのが自然である．このような手法は**多次元尺度法** (Multi Dimensional Scaling) と総称される．この種の手法は社会科学や調査の解析などに古くから使われている．また，主成分分析は，この特殊例（特別な距離の与え方をした場合）と見られることが知られている [2]. 多次元尺度法は，「言語の創発」のシミュレーションの解析に使われた例がある [3]. 距離そのものではなく，距離の大小の順位のみに着目して，それをなるべく保存するような配置を考える手法は**非計量的**多次元尺度法といわれる．これも古くからあるが，最近，物理関係の人達が，生物の形態の分類・進化への応用に注目しているようである [4].

本研究会での**高野**氏の講演では，リウビル演算子あるいはマルコフ連鎖の遷移行列の固有状態を，シミュレーションから近似的に求める方法が議論された．これはある意味では根本氏らの方法の動的な方向への拡張ともみられる．本来，この固有状態は phase space（システムサイズ N について，イジングスピンの場合 2^N 次元）に存在するもので，もともと N 次元空間に住む感受率行列の固有モードとは質的に異なる存在であるともいえるが，高野氏らの方法では，変分近似を考える際に低い次元の空間に落として考えているので，実質的には似た面が出てくる（と筆者は理解している）.

2.2 グループ分けモデル[3]

もうひとつの代表的手法としては，統計学で**有限混合分布** (Finite Mixture) モデル [5] といわれているものがある．非階層的クラスター分析といわれる手法群のなかで，「しっかりした統計モデルにもとづく」ものは，有限混合分布のあてはめをベースにしている．また，最近では（主成分分析もそうであるが）ニューロの人たちの間で，unsupervised learning の基礎モデルとしても研究されている.

このモデルでは，データの１つ１つは複数の分布のどれかから来ているが，その由来は観測者には未知であると仮定される．より具体的にいえば，なんらかのテンプレート（プロトタイプ）があって，そのいずれかのまわりの平均場（引力）の中でゆらいでいる対象

[2] この研究の原型はそうとう以前に物理学会で発表されており，今後さかんになる（かもしれない）この分野の先駆的なものといえると思う．論文の発表が待たれる.

[3] この方向を「空間的」手法と対立する方向と考えるかどうかは，見方による．たとえば，主成分分析を一個の多変量ガウス分布の分散・共分散行列を扱うものと考えた場合，複数のガウス分布の混合 (Gaussian Mixture) によるグループ分けモデルはその拡張と考えられる．実際，文献 [6] では Gaussian Mixture に近いモデル（Jumping-Among-Minima Model）を主成分分析の拡張として導入している．また，多次元尺度法を利用することで，グループ分けモデルで求めたテンプレートの間の関係を空間的に表現するといったかたちでの「併用案」も考えられる.

をサンプルしたものがデータになっていると考えるわけである.

このモデルをあてはめる素朴な方法は，最初になんらかの方法でテンプレートの集合か各データの所属先のいずれかを決め，次に残りのほう（及び，テンプレートのまわりの平均場や分散・共分散）を求めるという方法である. より洗練された方法としては，ランダムなテンプレートから出発して，重みつきの分類とテンプレート・平均場の再定義をを交互に繰り返すことで，テンプレート・平均場とデータの所属（度）を同時に決める方法がある. これはEMアルゴリズム（EMはExpectationとMaximization）と呼ばれ，ある意味で「自己組織化」的なアルゴリズムといえる（少々大げさであるが，世の中で自己組織とか創発とか称しているものも，実態はこの程度のものが多い.）. 実はこれは有限混合分布モデルでパラメータ（テンプレート・平均場）の最尤推定を行うアルゴリズムになっている（本研究会の伊庭・福島の報告を参照）.

伊庭の発表（福島氏との共同研究）では有限混合分布とEMアルゴリズムを用いてスピングラスモデルの有限温度でのシミュレーションデータから準安定状態の組を構成することを試みた. この範囲でもいろいろ問題があるが，将来的にはタンパクのシミュレーションなどにも適用できればよいと思う.

2.3 階層的クラスター分析

スピングラスの局所的極小の配置などでは，もともとultrametricな構造が予想されていることもあって，配置間の距離を利用したtree状の表現がしばしば使われる[7]. データ解析の用語でいえば，階層的クラスター分析といわれるカテゴリーの方法である. さまざまなやり方があるが，基本的には手続き的・動的に定義されていて，有限混合分布モデルのような基盤となる統計モデルがあるわけではない. 階層的な統計モデルを作ることも不可能ではないし，それ自身，挑戦的な課題であるが，モデル選択の規準・アルゴリズムなどかなり大変だと思う.

2.4 リヤプノフ・ベクトル

大自由度の決定論的な力学系の解析に使われる汎用的な手法として，リヤプノフ・スペクトラム，リヤプノフ・ベクトルの方法がある. これは，位相空間での固有値・固有状態に相当するものを扱うという点では，高野らの方法と似ているが,「軌道の離れ方」「離れる方向」に着目している点では異なっている. 分野間の交流をはかるという意図もあって，今回は，**小西**氏に依頼して，リヤプノフ解析について，経験談をまじえて話して頂いた. なお，この分野について最近行われたワークショップ「リアプノフ解析の逆襲」の予稿や内容が，`http://www.kuamp.kyoto-u.ac.jp/~yyama/Workshop/` で見られる. 興味のある向きは参照されたい.

2.5 さらに

高自由度の系を理解するためには,(後述の「個別か普遍か」という問題はもちろんあるが),もっともっと多様な表現が望まれるのだと思う.研究会では,そこらへんのことを考えて,笹井氏に水、タンパクなど,複雑な系とその表現について話して頂いた.

3 良い表現とは何か

蛇足であるが,一般に「良い表現」とは何かということを考えてみたい.たとえば,次の表現は,閉区間上の滑らかな周期関数を一様収束の意味でいくらでもよく近似できる.

- フーリエ展開
- ルジャンドル展開
- 多項式近似(次数順)

しかし,これらは,どれもそれだけでは恒等式あるいは座標変換にすぎない.「モデル」「表現」としての価値は,むしろそれを途中で打ち切ったものがどれだけ良くもとのものを表しているかにある.つまり,「途中で打ち切ったフーリエ展開」「途中で打ち切ったルジャンドル展開」がモデルなのである.この意味で,「ルジャンドル展開」と「多項式近似(次数順)」は,どっちも無限に続けば同じであるけれども,違う表現であるといえる.しばしば,統計的情報処理についていわれる「情報を捨てることが情報処理」であるというのはそういった意味である.

もちろん,表現というのは人間が見て理解するものである.また,物理の問題では,物理学の体系からくる意味,また,実際の実験の説明ということからくる意味づけも,いうまでもなく重要である.従って,いま考えているような問題では「予測」とか「情報圧縮」のような情報科学・統計科学の概念だけで「良い表現」とは何かを考えていくわけにはいかない.しかし,ある極限で『すべての情報を含んでいるから』良い表現であるという論理は一般に成り立たない,ということは注意しておく価値がある.そもそも,もとのハミルトニアンやリウビル演算子がすでにすべての情報を含んでいるわけであるから,それで十分なら,もとの式を眺めていれば良いわけである!

4 位相空間の動的解剖学

基研「複雑系」の初期の段階で既に「位相空間の動的解剖学」というようなことがテーマのひとつとして掲げられていた.その後も,「複雑系」のセッションとして,また,ハミルトン系の研究会などのテーマとして,位相空間の複雑さの解析ということが何度かとりあげられてきた.一方で,「動的解剖学」の「動的」については「対象が動的」はよいとし

ても「表現が動的」というのは理解しがたいという伊庭のコメントに応じて，(少なくとも金子邦彦氏は) この部分は撤回したのではないかと思う．

ここでは，あえて，「動的」の真意は何かということを再考してみたい．これは本研究会の趣旨に即しても重要な問題だと思う．伊庭の個人的意見としては次の２つがあると思う．

★ 一般的方法 対 個別的方法

これがなんで「動的」かと不審に思う人もいるかもしれないが，非線形動力学の人たちの間には，ダイナミクスを研究する人たちは精神が生き生きとしてダイナミックであり，熱平衡系の研究者は精神も熱平衡状態であるという考えがあるように伊庭には思える．これが本当であるかどうかはさておき，問題ごとに「柔軟かつ創造的に考える」ことを強調して「動的」といっているということは十分ありうると思う．

もちろん，一般論を使いこなしつつ，柔軟かつ創造的に考えられたほうが，そうでないよりも良いに決まっているし，その内容を最終的に一般化して一般論を豊富にできればもっと良いにきまっている．しかし，研究会を組織する際にどちらに重点をおくかは問題である．本研究会では，(結果はともかく) 一般論に重点を置こうと試みた．「複雑系」などではむしろ，各人の努力の様子をまず見て参考にするといった方針があったかもしれない．これは「概念」か「方法」(あるいは「手法」) かという問題にも関係してくるだろう．

★ global な表現 対 local な表現

上に微妙に関連した問題として，global か local か，というのがある．「リウビル演算子の固有ベクトル」「リヤプノフベクトル」のような量は時間軸・位相空間全体について平均化された global な量である．平均化された量で，位相空間の構造の多様性がとらえられるのか，というのがひとつの問題である．逆にいえば，そうでないような local な量について一般論がありうるのか，という問題もあり，ここで上の問題がからんでくる．これは計算物理というより，人工知能・統計科学一般にも共通する問題なのかも知れないが．「複雑系」の人たちからみれば，local な良い表現を得ようとするなら，「構成的方法」によって系をじっくり観察せねばいけない，という議論もあるであろう．また，「モード分解＝還元主義」という複雑系の人による批判は，歴史的にはいろいろな背景があるが，ここでいう global な表現への批判とも受け取れる．

実際に具体的な複雑なシステムを解析するとなれば，いろいろな方法や視点を併用するのが当然であって，ここで述べたような対立を設定する意味はあまりないのかもしれない．しかし，お互いに漫然と違和感をもっているような状態はよくない．意識の違いをはっきりさせるためには，未整理で抽象的であるけれども，こうした議論も何かの役に立つかもしれない．差異は常に意識化され，言語化されなくてはならない．「精神の解剖学」もまた必要なのである．

5　研究会の実態と展望

本セッションと拡張アンサンブルセッションの部分とをあわせて,「方法論中心の研究会」という企画であったが，実際にはどちらのセッションも，対象となる系の話がかなりの部分を占めた．それはそれで面白くもあり，また自然でもあったが,「スピングラス・たんぱく・ポリマー・水 *etc.* の教育的入門研究会」だという向きもなきにしもあらずであった．しかし,（計算物理の）方法論中心での顔合わせというのは，もう少し追求してみる価値があると思う．1999年度には後続する研究会は予定されていないが，もしできれば，次の年度には何か企画したいと個人的には考えている．その際には，クラスターモンテカルロ法やMDの技法など，ダイナミクスに関係した面も含めたい．

参考文献

[1] 手元にあるものではたとえば以下の２つ．北尾氏には是非この研究会に参加していただきたかったが，滞米中ということで断念した．２番目の論文の著者の中村氏は本研究会の拡張アンサンブルのセッションで講演されたが，その際には主成分分析を用いた解析についても触れられた．★ 北尾 彰朗，蛋白質・水系の階層的ダイナミクス,「物性研究」 60-3 (1993-6) 239- . ★ H. Shirai, N. Nakajima, J. Higo, A. Kidera, H. Nakamura, Journal of Molecular Biology (1998), 278, 481-496.

文献 [6] は，Gaussian Mixture と似たモデルを主成分分析の拡張として導入したものとみなせるが，主成分分析やそれ以外の手法によるたんぱく質モデルの位相空間の解析に関する文献を多数含んでいる（現在勉強中）．

[2] たとえば，以下参照（なんかミもフタもない実用的レファレンスですまないが，わかりやすい.），★ 統計解析ハンドブック 多変量解析 田中豊，垂水共之 編，共立出版，16 章（主座標分析）．

[3] 橋本敬氏（北陸先端大）による．★ 橋本 敬,「認知科学」1999 年第 6 巻第 1 号 (特集号「複雑系から見た知能創発」). ★ Hashimoto. T., *Multi-agent systems and Agent-Based Simulation*, Sichman, Conte and Gilbert (eds), LNAI series, volume 1534, 124–139, Berlin, Springer-Verlag, (1998).

[4] 田口善弘氏，大野克嗣氏らはこの目的のために新しいアルゴリズムを開発した．

[5] ★ D. M. Titterington, A. F. M Smith, and U. E. Makov: *Statistical Analysis of Finite Mixture Distributions* (John Wiley and Sons, Chichester, 1985). 本研究会報告の伊庭・福島の報告の文献も参照．

[6] 研究会のあとになって，北尾氏・郷氏ほかによる以下の文献を郷氏に教えて頂いた．そこでタンパク質の準安定状態の解析のために導入されているモデル（Jumping-Among-Minima Model）は，統計学の用語でいうガウス分布を成分とする有限混合分布モデルに近い．ただし，EM アルゴリズムのような分類と準安定状態の同定を同時に iterative に行う方法は利用されていないと思われる．★ Kitao, Hayward and Go, PROTEINS, 33:496 (1998).

[7] １つの例は，★ Nemoto, K., J. Phys. A, 21, L287 (1988).