不確実性のモデル化と予測

竹村彰通　合原一幸　駒木文保　青木敏　下川英敏　鈴木秀幸
情報理工学系研究科数理情報学専攻

概 要

現実の現象をモデル化する際には，不確実性のモデル化が避けられない．しかしながら現象のどの部分を確定的に扱い，どの部分を統計モデルなどにより不確実性として扱うかの切り分けはあきらかではない．不確実性のモデル化と予測グループではロバスト性という観点からモデル化の方法論を確立することをめざしている．

1　はじめに

数理モデルにおける確定的な部分と非確定的な部分の切り分け，さらに非確定的な部分の扱い，のバリエーションを考慮すれば，現象のモデル化において様々なアプローチが可能である．実際多くの競合する方法論について，それらのメリットが個々に主張されそれぞれに研究されているのが現状である．このような中で重要な研究目標は，ロバスト性の観点，すなわち与えられた現象へのモデルの安定的な適合と予測の観点から，多くの方法論を統一的に比較しすぐれたモデルを選びだす指針を与えることである．このような指針を与える基礎研究として，個々の手法の一層の深化が必要であることは言うまでもない．

以下では，以上のような目標を念頭におきながら，複雑システムモデリングの理論と応用，個票データベースの安全な利用法，統計モデルの情報幾何とベイズ理論，マルコフ連鎖モンテカルロ法による離散データ解析について，この時点での研究成果と今後の研究の展望について述べる．

2　複雑システムモデリング

複雑システムモデリングの理論と応用に関して，今年度は，遺伝子・蛋白質ネットワーク [1] や部分放電現象 [2] などの，実在する複雑システムの数理モデルを構築すると共に，その複雑現象の背後に存在するゆらぎ特性や放電頻度特性の数理構造を理論的に明らかにした．また，2004 年 12 月 6 日より 8 日に，国際シンポジウム “International Symposium on Complexity Modelling and its Applications” を，東京大学本郷キャンパス工学部 6 号館において，科学技術振興機構 ERATO 合原複雑数理モデルプロジェクトと共催で開催し，内外の研究者を集めて，複雑システムモデリングの理論と応用に関して集中的に議論を行なった．

3　個票データベースの利用と安全管理の手法

官庁統計や社会調査で得られる統計データは，伝統的には統計表の形に整理された上で分析されてきた．最近では，統計調査の際に得られる個々の回答者のデータ (個票データ) は直接デジタルデータとして記録され，また統計パッケージ等の整備により，これらのデータを集計表以前の生の形で解析することが可能になってきた．この際に問題となり得るのは，回答者のプライバシーの問題である．統計調査で得られたデータからプライバシーが侵害されるようなことがあれば，統計調査そのものが成り立たなくなる可能性がある．

個票データの安全性の評価には，回答者が特定される確率のモデル化が必要となるが，この目的

のために集団遺伝学や計量言語学で発展してきた確率モデルを用いることができるのである．特に集団遺伝学での種の分布に関する確率モデルや，計量言語学における語彙の分布に関するモデルが，個票データの問題に直接的な関連を持っている．逆に，個票データの安全評価という観点からこれらのモデルを見なおすことにより，これらのモデルについてより深い理解が得られる．竹村および共同研究者の研究成果については，統計数理研究所の『統計数理』の 2003 年第 51 巻第 2 号で竹村をオーガナイザーとする「特集 個票開示問題の統計理論」([3]) が発行され，この分野におけるわが国の国際的貢献を含んで個票開示問題研究が多いに進展した．竹村自身がここで分野全体のサーベイ ([4]) を与えている．

4 統計モデルの情報幾何とベイズ理論

最近，ベイズ的手法の実用的な統計手法としての有効性が，広く統計科学をはじめとする情報分野の研究者に認識されてきている．本研究では，統計モデルに基づいた，性能の良いベイズ予測・ベイズ推測をシステマティックに構成する手法を開発することを目的としている．

従来，数理統計学におけるベイズ推測理論では，パラメータ推定について特に多くの研究がなされてきた．しかし，平均 2 乗誤差などのモデルの自然な構造と必ずしも合致していない人工的な損失関数にもとづいている研究が多い．本研究では，ダイバージェンスと情報幾何に基づいて問題を定式化し，パラメータの推定量よりも予測分布や事後分布そのものの評価をおこなうことにより問題の本質的な部分を取り扱った．

統計モデルの多様体が，大域的な幾何学的性質に関するある条件をみたす場合には，ジェフリーズ事前分布を利用した予測を優越する予測分布が構成できることが示し，さらに個別の重要なモデルに関して，許容性やミニマックス性などの最適性を備えた予測分布の構成法の研究をおこなった．統計モデルに対応する多様体の大域的な性質がベイズ統計理論と直接結びついていることを明確にし，多様体の大域的な微分幾何学的性質を調べることにより従来良いとされてきた予測（例えばジェフリーズ事前分布に基づくベイズ予測）を優越する予測を構成することを可能にした．

従来，情報幾何では局所的な性質を調べれば十分であったが，ベイズ予測分布の構成法を情報幾何学的な視点から考えると，モデル多様体の体積増大度などの大域的性質が本質的な役割を果たす．ベイズ推測理論と情報幾何の特に大域的な性質を本質的に結び付ける視点はほとんどなかったと思われる．

また，ベイズ予測分布の理論は，未知のデータに対する予測対数尤度にもとづいて理論を構成しているため，G. Schwartz による BIC の理論，J. Rissanen による MDL 理論，A. P. Dawid の提唱する prequential analysis の理論, A. R. Barron らによるベイズ符号化の理論などの統計科学・情報科学における重要な基礎理論と密接な関係がある．本研究での問題設定では，手持ちのデータを利用して将来のデータを予測するという問題の定式化をしている点がこれらの関連する理論とは本質的に異なるため．従来の予測尤度に基づく理論で取り扱えなかった多くの問題が取り扱うことが可能になるという特長がある．

具体的には, 予測分布の許容性，情報理論と予測分布，ブートストラップ法を利用した予測 [5], 統計的予測の漸近理論とモデル多様体の情報幾何学的構造，プロパーな事前分布に基づくベイズ予測分布で通常の無情報事前分布に基づくベイズ予測分布を優越するものの構成，についての研究をすすめた．

とくに，分散共分散行列既知の多変量正規分布，多変量ポアソン分布，2×2 のウィシャート分布などの基本的なモデルで有限サンプルに基づく厳密な理論の研究をおこない，予測の許容性やミニマックス性等についての結果を得た．多変量ポアソンモデルにおけるあるクラスのベイズ予測分布の許容性について証明し，予測分布と情報理論における MDL 理論との関係と相違点，またイギリスの Dawid らにより提案された prequential

likelihood の理論との関係について議論した [6].

また，プロパーな事前分布が無情報事前分布として利用できる具体例を求めた．無情報事前分布として利用できるプロパーな事前分布が存在するための幾何学的な条件を求め，有用性と限界について議論した．

さらに，階層ベイズモデル，ウェーブレットを利用したモデル，ネイマンスコット型のモデルなど，応用上重要で，高次元のパラメータをもつパラメトリック統計モデルの情報幾何学的構造を調べ，それらが幾何学的に非常に自然で簡単な構造を持つことを示した．

今年度得られた関連する研究成果について 2004 年 7 月にドイツで開催された国際会議で発表を行った [7].

5 マルコフ連鎖・モンテカルロ法による離散データ解析

集団を，性別や年齢，あるいは疾患の有無や生活習慣などの要因で多重分類し，それぞれの人数を表の形で表したものは分割表と呼ばれる．分割表に要約されたデータから，要因間のさまざまな関連を調べるための解析手法は，その，医学，疫学，工学，自然科学などのさまざまな分野における応用上の重要性もあり，発展してきたが，特に近年，計算機の進歩や，インターネットを利用した大規模なデータを取り扱う可能性などを背景に，サンプリングベースの統計量の計算手法が注目されている．類似の例では，遺伝性疾患の関連遺伝子の同定において，候補となる数十〜数百の部位を同時に解析しなければならない，というような状況がある．そのような場合に，研究上興味のない多くの不確実性（統計学的には局外母数に対応する）を，興味の対象となる不確実性と同列に扱うことは全くナンセンスであるため，局外母数の値によらない推定方式が必要となる．より具体的には，われわれが遭遇する多くの統計的推測の問題は，集約的には，ある統計量の条件付期待値の推定問題として定式化することができ，マルコフ連鎖・モンテカルロ法は，その数値的評価のためのひとつのアルゴリズムである．とくにそれは，単純なモンテカルロ積分が実行できない（直接的なサンプリングが不可能）というようなケースを想定しており，Importance Sampling 法とはその目的を共有するものである．

分割表データのような離散データの解析に，マルコフ連鎖･モンテカルロ法を適用するための理論的な困難は，そのために必要とされる，標本空間内の連結なマルコフ連鎖の構成が難しい，という点である．この問題に対する最初の解は，Diaconis and Sturmfels の 1998 年の論文の中で与えられた．彼らは，連結な連鎖を構成するために必要な基底 (マルコフ基底) が，多項式環のトーリック・イデアルの生成系に対応することを示し，代数アルゴリズムを利用したマルコフ基底の計算アルゴリズムを提案した．しかし一方で，この代数アルゴリズムを用いた基底の算出方法には問題点も多く，中でも，計算時間の問題と，得られる基底が極小でないという問題は重大である．計算時間の問題は，代数アルゴリズムの理論的な計算量が，変数の数の二重指数オーダーであることに起因しており，比較的小さなサイズの問題に対しても，実際の計算はすぐに破綻してしまう．また，基底が極小でないという問題は，代数アルゴリズムが変数間に項順序を与えて計算するものであるため，変数間の対称性が崩れる，ということに起因している．

これに対し青木，竹村は，これまでに，変数間の対称性に注目した手法により，代数アルゴリズムを用いずに直接極小な基底を算出する方法を提案し，比較的小さな分割表に対しては，代数アルゴリズムを用いるよりもはるかに効率的に極小基底が得られることを示した．特に，3 因子交互作用の検定問題として定式化される 3 元分割表の問題に対する結果は，計算機に頼るよりも，人間が「目で見て」判断し，膨大な場合の数を効率的に分類し，探索する，という接近法が，きわめて効率的であることを示す例であり，この分野の他の研究者に，大きな影響を与えたといえる．さらに，極小なマルコフ基底の性質と特徴付けに関する研究も，いくつかの成果を得た．現在投稿中である

論文 [8] では，分割表の各軸に関する水準の入れ換えの群の，分割表への作用を考えることにより，極小不変マルコフ基底の概念を提案し，その応用を示したが，現在はさらに，より一般的な群の作用に対する不変性の定義と，その応用例の研究を進めている．これらの一連の研究は，本来分割表が持つ，変数間の対称性という性質を，様々な問題設定で統一的に扱うための下地となる，本質的で重要な研究である．また，[9] では，別の概念として，標本空間の任意の 2 要素の間の「距離」を常に縮めることができる，という性質をもつマルコフ基底を，ノルム縮小マルコフ基底として定義し，その性質の解明を行なっている．マルコフ基底によりノルム縮小が常に可能，という性質は，その基底から誘導されるマルコフ連鎖の収束の速さの評価と関わっており，実際の数値計算や実データへの適用といった応用面でも重要な結果であると考えられる．これらの研究結果は，国内の学会，研究会のみならず，2004 年 12 月にインドで行なわれた国際学会 Eleventh International Conference on Interdisciplinary Mathematical and Statistical Techniques でも発表され，好評を得た．

参考文献

[1] Ryota Tomioka, Hidenori Kimura, Tetsuya J. Kobayashi, Kazuyuki Aihara: “Multivariate Analysis of Noise in Genetic Regulatory Networks”, *Journal of Theoretical Biology*, **229**, pp.501–521 (2004).

[2] Hideyuki Suzuki, Kazuyuki Aihara, Tatsuki Okamoto: “Complex Behaviour of a Simple Partial-discharge Model”, *Europhysics Letters*, **66**(1), pp.28–34 (2004).

[3] 『統計数理』「特集 個票開示問題の統計理論」2003 年第 51 巻第 2 号, 統計数理研究所, 竹村彰通:オーガナイザー．pp.181–387.

[4] 竹村彰通 (2003). 個票開示問題の研究の現状と課題.「特集 個票開示問題の統計理論」『統計数理』2003 年第 51 巻第 2 号, pp.241–260.

[5] Fushiki, T., Komaki, F., and Aihara, K. (2004). On parametric bootstrapping and Bayesian prediction, *Scandinavian Journal of Statistics*, vol. 31, no. 3, 403–416.

[6] Komaki, F. (2004). Simultaneous prediction of independent Poisson observables, *the Annals of Statistics*, vol. 32, no. 4, 1744–1769.

[7] Komaki, F. (2004). Noninformative priors for prediction based on group models, in R. Fischer, R. Preuss, and U. von Toussaint (eds.) *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Garching, Germany 2004, AIP Conference Proceedings, American Institute of Physics, Melville, 525-532.

[8] Aoki, S. and Takemura, A. (2003). Invariant minimal Markov basis for sampling contingency tables with fixed marginals. *METR Technical Report, 03-25.*

[9] Takemura, A. and Aoki, S. (2004). Distance reducing Markov bases for sampling from discrete sample space. *Bernoulli*, to appear.