２．１．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
新領域創成科学研究科複雑理工学専攻

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．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す．

1　はじめに

現実の様々な現象をモデル化する際に，現象のどの部分を確定的に扱い，どの部分を非確定的に扱うかという切り分けが重要である．複雑なシステムを扱う限りにおいて，様々な不確実性を避けることはできないから，純粋に確定的なモデルではロバスト性に欠ける．一方で，単に不確実性と言っても，不確実性には例えば現象に関する知識の不足等を含め，さまざまな側面があり，確率的なモデルを設定して統計的推論をおこなえばよい，といった単純なものではないことに注意する必要がある．例えば時系列解析一つをとりあげても，統計学で扱う時系列モデルの他にもカオスに代表される非線型システムに基づいた解析を用いることもできる．

このように，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，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　研究の現状と今後の展望

ここでは，メンバーの研究テーマにそって，確率場の理論と応用，個票データベースの安全な利用法，ベイズモデルの構築と予測，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述べる．これらはそれぞれ不確実性の扱いにおいて重要な研究課題である．

2.1　確率場の理論と応用

確率過程や時系列解析の理論は応用上の重要性もあり，大きく発展して来た．連続時間の確率過程の理論も，近年では例えば数理ファイナンスへの応用などの目的から，実際の現象の解析の道具として用いられはじめている．

さらに最近になって，地理的なデータの集積とともに空間統計データの解析の重要性が高まって来ている．空間統計データの数学的な基礎は確率場の理論によって与えられるものである．確率過程から確率場への一般化は，添字集合の１次元か

ら多次元への一般化であるが，添字集合が多次元化するとともに，これまでの確率論的な方法に加えて，幾何学的な手法を併用する必要が出てくる．例えば，さまざまな方向からの断面に関するデータが得られるような場合には，添字集合が球面やグラスマン多様体などの場合も扱わなければならない．このように確率場の理論では，確率論と幾何学の融合が必要であり，解決すべき問題が多く残されている．

観測される確率変数がガウス分布に従うガウス確率場の理論においては，最近になって「オイラー標数法」や「チューブ法」の性質があきらかになって来た．これらは，確率変数の最大値の分布に関して良好な近似を与える手法として 1980 年代に提案され，画像データの解析等に応用されてきたものであるが，数学的には近似の有効性が保証されているものではなかった．[1] では，ガウス確率場が有限な直交関数展開を有する場合について，モースの定理を拡張することにより，「オイラー標数法」と「チューブ法」の同等性を証明した．また近似の誤差限界を与え，近似の正当化をおこなった．この話題については，最近になって竹村と統計数理研究所の栗木哲氏，及びスタンフォード大学の Jonathan Taylor 氏との共同研究が急速に進展しており，直交関数展開の有限性の仮定が不要であることが確認された．また近似の誤差限界の評価も大幅に改善されつつある．この研究は，確率場の理論に幾何学の手法を応用するのみならず，幾何学的な諸概念を確率場の概念に対応させることにより，いわば無限次元の空間の幾何学の研究という側面を持っており，この研究で得られる諸結果は基礎的な重要性を持つものと考えられる．この問題についてはここ 1,2 年のうちに一連の研究成果を発表する予定である．

2.2 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来た．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来た．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．

統計的な個票データはあくまで統計的な分析が目的であり，名前や住所など個人を直接特定するような情報は不要である．従ってこれらの情報は削除される．しかしながら，たまたま高額所得者がデータに含まれるような場合を考えると,「高額所得」といった間接な情報から回答者が特定される危険がある．統計的な個票データについてはこのような問題点があり，個票データの安全性と有用性のバランスを確保することが必要である．この問題は，情報社会における個人情報の保護という一般的な観点からも重要な研究課題である．

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来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』の特集号が現在編集中である．

2.3 ベイズモデルの構築と予測

計算機の発達にともない，ベイズ的な統計手法を実用的な統計モデルに対して適用することが可能になり，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情報分野の研究者に認識されてきている．

ニューラルネットワークやサポートベクトルマシンなどの機械学習や情報理論等の分野では従来

の手法がベイズ理論の枠組みから自然に捉えなおすことが可能であることが示されたり，ベイズ理論に基づく新たな手法が数多く提案され利用されるようになってきている．

古典的なベイズ統計学では，パラメータに関する知識を事前分布の形で表現して，得られたデータをもとにベイズの定理によってパラメータ空間上の分布を更新してパラメータに関する推測を行う，という考えがとられてきた．この場合，どのようにして事前分布を構成するかということが常に問題になる．パラメータに関して何も情報がない状態を表現する無情報事前分布を，どのような目的に対しても適用が可能な様に構成するのは困難であることが認識されるようになっていた．

ベイズ予測の観点から事前分布を構成することにより，この問題に対するひとつの解決を与えることができる．ベイズ統計理論におけるパラメータ空間上の事前分布が，統計モデルに対応する多様体上の体積要素と見なせることを利用して，モデル多様体の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が多くの例で可能であることがわかってきている．このアプローチにより漸近理論，変換群モデル，個別のモデルなどさまざまなレベルで，多くの問題が理論的に解決できると考えられる．

情報幾何学的手法を用いた統計的推測に関する研究はいままでにも数多くなされてきた．そのほとんど全てが統計モデルの多様体の局所的な性質に基づく研究であり，情報幾何においては局所的な性質の研究で十分であると言われてきた．しかし，ベイズ理論ではモデル多様体の体積増大度などの大域的な性質が本質的な役割を果たすため，従来のモデル多様体の局所的な性質のみの研究では不十分になる．

漸近理論に基づく予測理論の研究により，統計モデルの幾何学的性質と予測分布の性能の関係についていくつかの結果が得られている．特に，モデル多様体がある微分幾何学的な性質をもつとき，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分布を構成できることがわかっている．

統計モデルが変換群構造を持つときには，より詳しい議論が可能になる．この場合にはモデル多様体は等質空間になり，モデル多様体上の不変測度とベイズ予測のために利用する事前分布の構成の問題との関係についていくつかの結果が得られている．また，ウェーブレット変換を利用したモデルなど，ある種の時系列モデルや空間統計学に現れるモデルは群構造と密接に関係している事が知られており，幾何学的なアプローチが有効であると考えられる．

さらに個別の重要なモデルに関して，有限サンプルに基づく厳密な議論が展開できる．多変数正規モデル，多変数ポアソンモデル，ウィシャートモデルなどの重要な統計モデルに関する有限サンプルに基づく厳密な予測分布の理論を，許容性・ミニマックス性などの最適性の性質に関する研究も含めて展開している [2].

今年度得られた関連する研究成果について 2002 年 12 月に韓国で開催された国際会議で発表を行い高い評価を得た．

2.4 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．このように，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．

離散データ解析におけるマルコフ連鎖・モンテカルロ法は，1998 年に Diaconis and Sturmfels によって提案され，注目を集めた．この方法は，マルコフ連鎖を構成するための基底を，代数アルゴリズムを用いて算出するものであり，理論上はこの方法で，任意の離散の条件付分布からのサンプリングが可能になった，という点で興味深いものである．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，変数間の対象性が崩れる，ということに起因している．

これに対し，青木，竹村は，変数間の対象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．これらの結果は，例えば [3] などで発表するとともに，2002 年 8 月にベルリンで行なわれた国際学会 Compstat 2002，および，2002 年 12 月にソウルで行なわれた国際学会 East Asian Symposium on Statistics において発表され，好評を得た．また，より大きな分割表や，さまざまな特殊な分割表に関する極小基底や，極小基底の性質，その一意性などの理論的な結果もあり，これらは現在，投稿中である．この研究で得られた諸結果は，理論的な重要性のみならず，応用上，非常に大きな意味を持つものであり，今後も研究を継続し，国内外で発表していく予定である．

参考文献

[1] Takemura, A. and Kuriki, S. (2002). On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. *Annals of Applied Probability*, **12**, 768–796.

[2] Komaki, F. (2002). Simultaneous prediction of independent Poisson observables, tentatively accepted for publication in *Annals of Statistics*.

[3] Aoki, S. and Takemura, A. (2002). Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. *Australian and New Zealand Journal of Statistics*, to appear.