ベイズモデリングによる実世界問題チャレンジ

2010年7月9日　統計数理研究所　オープンハウス

樋口　知之　　モデリング研究系　教授

1．ベイズモデリングとは？

■動的かつ複雑な対象を理解する：対象から得られるデータは，さまざまな計測・観測条件に応じて時間的にも空間的にも多様な様相を呈する。この各状況の特性に即しつつデータもうまく説明できる表現方法をもし手にしたのなら，予測や制御といった次のステップの作業が見通しよく完遂できる。通常我々はこの要求に，データ$\boldsymbol{y}$の生成メカニズムを，潜在変数ベクトル$\boldsymbol{x}$を持つ統計モデル$p(\boldsymbol{y}|\boldsymbol{x})$でもって近似することで解決を図る。

■事前分布：従来の統計的モデリングでは$\boldsymbol{x}$の次元をなるべく小さくすることが王道，さらにいうと美徳とされてきた。潜在変数を増やせば増やすほど統計モデルの記述能力は向上するが，汎化能力と呼ぶ将来のデータの予測能力が減少する。この問題への対策として，$\boldsymbol{x}$についても統計モデル$p(\boldsymbol{x})$を想定するのがベイズモデルである。この$p(\boldsymbol{x})$をベイズ統計では事前分布と呼ぶ。

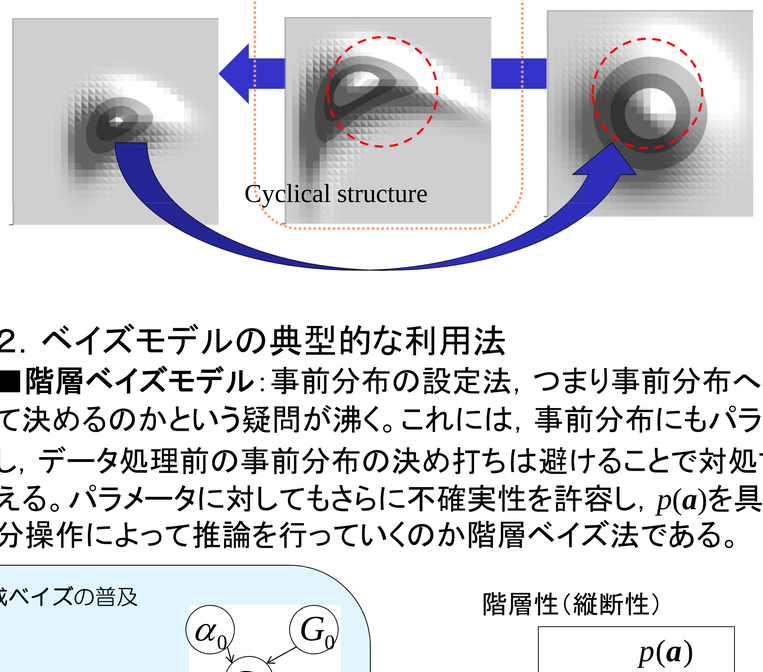
■ベイズの定理：事前分布の導入により以下のベイズの定理でもって，想定した事前分布$p(\boldsymbol{x})$がどのように修正されるのか，つまり$\boldsymbol{x}$に関する不確実性がデータによりどの程度修正されたかを観察するのである。この$p(\boldsymbol{x}|\boldsymbol{y})$を事後分布と呼ぶ。この仕組みにより多数のパラメータも安定して推論できるようになり，結果として高い予測能力とデータ記述能力を同時に持つ，総合的な統計モデルが構成できる。ここでは、この一連のモデル化行為をベイズモデリングと呼んでいる。

$$p(\boldsymbol{x}\mid\boldsymbol{y})=\frac{p(\boldsymbol{y}\mid\boldsymbol{x})\cdot p(\boldsymbol{x})}{p(\boldsymbol{y})}\propto p(\boldsymbol{y}\mid\boldsymbol{x})\cdot p(\boldsymbol{x})$$

Posterior
Improved knowledge about values of $\boldsymbol{x}$

Likelihood
Feasibility of realization of $\boldsymbol{y}$ for given $\boldsymbol{x}$

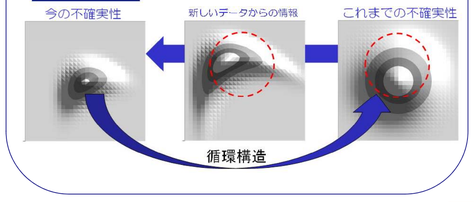
Prior
Belief about values of $\boldsymbol{x}$



事後分布の逐次的更新

$$p(\boldsymbol{\theta}\mid\boldsymbol{y}_1,\boldsymbol{y}_2)\propto p(\boldsymbol{y}_2\mid\boldsymbol{\theta})\cdot p(\boldsymbol{\theta}\mid\boldsymbol{y}_1)$$

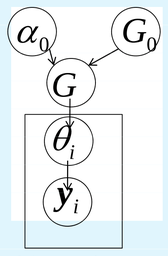
新しいデータ$\boldsymbol{y}_2$を含めたすべてのデータにもとづく$\boldsymbol{\theta}$の事後分布　新しいデータ$\boldsymbol{y}_2$の尤度　既存データ$\boldsymbol{y}_1$にもとづく$\boldsymbol{\theta}$の事前分布



2．ベイズモデルの典型的な利用法

■階層ベイズモデル：事前分布の設定法，つまり事前分布への信念の置き具合いはどのようにして決めるのかという疑問が沸く。これには，事前分布にもパラメータ$\boldsymbol{a}$を導入することで自由度を残し，データ処理前の事前分布の決め打ちは避けることで対処する。つまり，事前分布を$p(\boldsymbol{x}|\boldsymbol{a})$で与える。パラメータに対してもさらに不確実性を許容し，$p(\boldsymbol{a})$を具体的に計算に導入し，さまざまな積分操作によって推論を行っていくのか階層ベイズ法である。

連成ベイズの普及

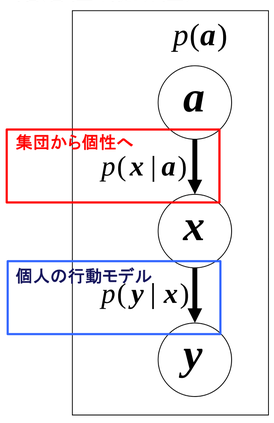


■ベイズモデルの階層がますます高層化(HDP: 階層ディリクレプロセス)

■ノンパラメトリックベイズ法の浸透
・ディリクレプロセスの利用が自然言語の分野ですすんだ。
・モデル選択からモデル混合へ

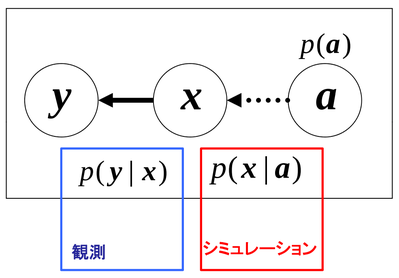
■ミクロからマクロへ生成モデルの連続結合（“連成”）による、いわばフォワード推論がベイズ計算の主流

階層性（縦断性）



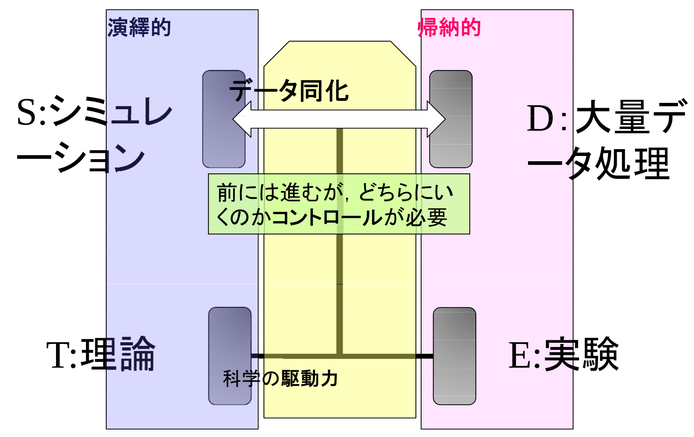
$$p(\mathbf{x},\mathbf{a}\mid\mathbf{y})\propto p(\mathbf{y}\mid\mathbf{x})\,p(\mathbf{x}\mid\mathbf{a})\,p(\mathbf{a})$$

横断性



■対象の構造を自然にモデル化：上図の左側を使って、各個人の購買にまでモデル化を試みるマイクロマーケティングを例に階層ベイズモデルの利用法を考察する。最下層では個人の購買行動といったミクロ単位の確率モデルを採用し，階層があがるにつれ，個人の特性，層別化された集団の特性，地域の特性，そして時代効果といったふうに，各demographicが階層ベイズモデルの階層に対応していているのが分かる。このように，マイクロマーケティングと階層ベイズモデルは親和性が高い。

■異種情報をつなぐ：階層構造を横にすることで，階層ベイズモデルは異種情報を統合するプラットホームにもなり得る。このことを模式的に示したのが図の右側である。今，一番左端にあるように変数$\boldsymbol{y}$と$\boldsymbol{x}$の間の関係が$p(\boldsymbol{y}|\boldsymbol{x})$で，また$\boldsymbol{a}$の情報の不確実性を$p(\boldsymbol{a})$でモデル化されている状況を想定する。このとき，$\boldsymbol{x}$と$\boldsymbol{a}$の間の関係をもし何らかの形で確率的に表現し，それを統計モデル$p(\boldsymbol{x}|\boldsymbol{a})$で与えることができたならば，データ$\boldsymbol{y}$に基づいた$\boldsymbol{a}$の不確実性の評価が可能になる。図では，点線で示されている$p(\boldsymbol{x}|\boldsymbol{a})$の関係を明示的に与えたならば（実線にしたならば），いままで分離していた$\boldsymbol{y}$に関する情報と，$\boldsymbol{a}$に関する情報を統合することができる。この仕組みを利用した例がデータ同化（下図）やゲノムデータ解析である。



両者をつなぐのは禁手とされてきた！

さまざまな科学の領域において計算機シミュレーションは，研究対象の複雑な現象の解明の手段として、実験・理論と並ぶ自然科学の第三の研究手法として確立している。この関係を左図に模式的に示した。後輪が、科学の駆動力である、実験と理論である。また左前輪が，第三の科学と呼ばれるシミュレーションである。一方、データ解析の基盤となる統計科学においては、研究対象の理解のために、現象を支配している規則、関係式といった経験則を観測や計測データから推定していく、すなわち帰納的推論を行う。統計科学は，データマイニングや機械学習といった，大量データ処理を行う研究領域と相互に関連しながら，図の右前輪に対応する分野を構成している。研究推進にあたって、右側が帰納的アプローチに，また左側が演繹的アプローチになるように描いてある。データ同化は，シミュレーション科学のような演繹的な推論と、統計科学に代表される帰納的な推論を融合するためのプラットフォームである。つまり、前輪をつなぐ軸となる数理技術である。前輪の両輪が軸として繋がることで、制御や予測がうまくいくことが期待されるのである。もしうまくいかない場合でも，その問題をつきとめることで、シミュレーションモデルに内在する不確かさを推測・検討できる。

3．データ同化

■手法：我々はデータ同化手法を統計科学の枠組みで正確に定義し、統計科学や情報科学の分野において蓄積されてきたアルゴリズムやモデリングに関する研究成果を利用しながら、これからの計算機インフラ※を視野に入れつつ、実装が平易かつ広い分野に適用できる逐次データ同化手法の開発を行っている。

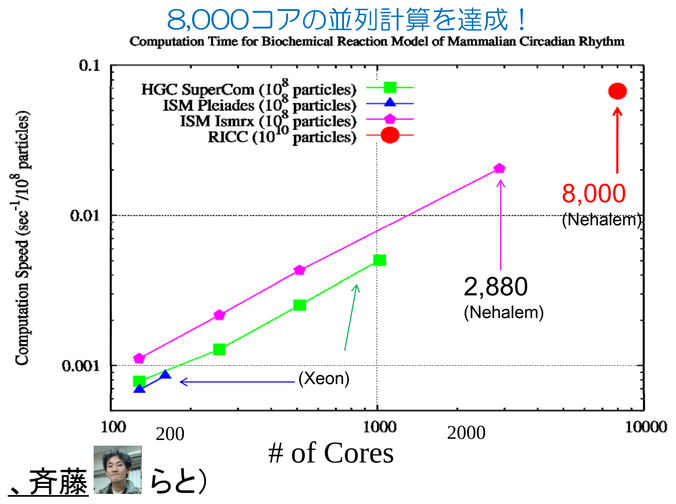
※1：次世代スーパーコンピュータプロジェクトへ参加　（吉田、中野（慎）、長尾、斉藤らと。情報・システム研究機構では我々のみ国家プロジェクトに参加！）

汎用CPUを超並列につないだ分散メモリ型にHPCのプラットフォームは確実に移っていく。このHPC環境の中で、計算機の集約的利用に基づく統計計算、つまり計算集約型統計計算の計算効率を高めるためには、HPCのハード特性を最初から意識した手法の開発が肝要となるであろう。考慮すべきハード特性としては、階層性とネットワーク構造の二つが特に重要である。

データ同化Gが利用している（予定の）スパコン

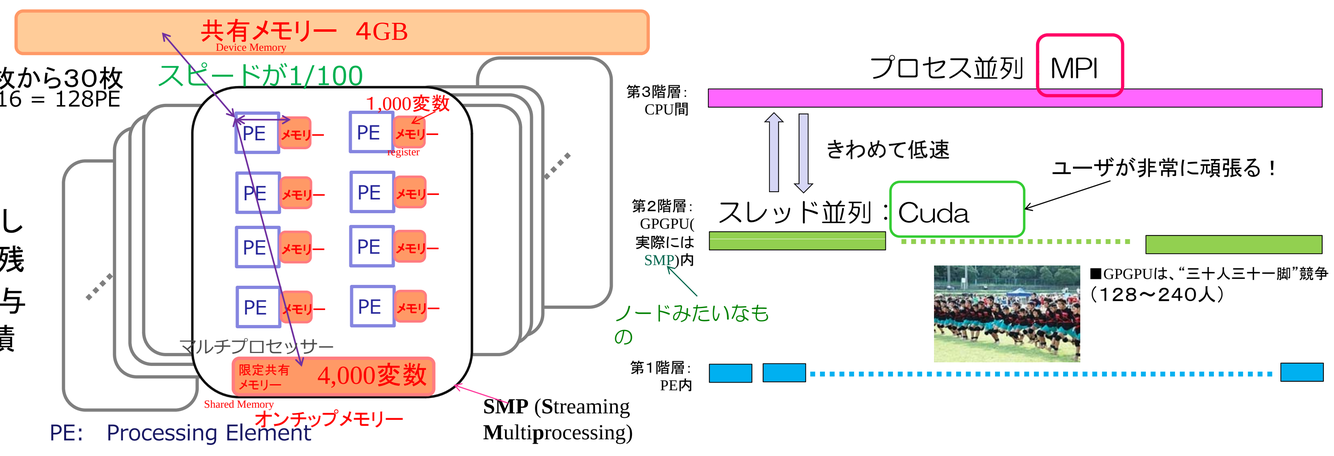
System	CPU	Clock frequency	# of nodes	# of cores	Memory
ismrx (ISM)	Intel Xeon X5570	2.93 GHz	360 nodes	2,880 cores	32GB/node
pleiades (Higuchi Lab)	Intel Xeon E5440	2.83 GHz	24 nodes	192 cores	32GB/node
sheep (Higuchi Lab)	Intel Xeon E5550	2.66 GHz	13 nodes	104 cores	24GB/node
Type-B (HGC/Univ. Tokyo)	Intel Xeon E5450	3.00 GHz	768 nodes	6,144 cores	32GB/node
RICC (RIKEN)	Intel Xeon X5570	2.93 GHz	1,024 nodes	8,192 cores	12GB/node
PetaCom* (RIKEN)	SPARC64 VIIIfx	2.00 GHz	> 80,000 nodes	> 640,000 cores	16GB/node
M System (JAXA)	SPARC64 VII	2.50 GHz	3,008 nodes	12,032 cores	32GB/node

8,000コアの並列計算を達成！

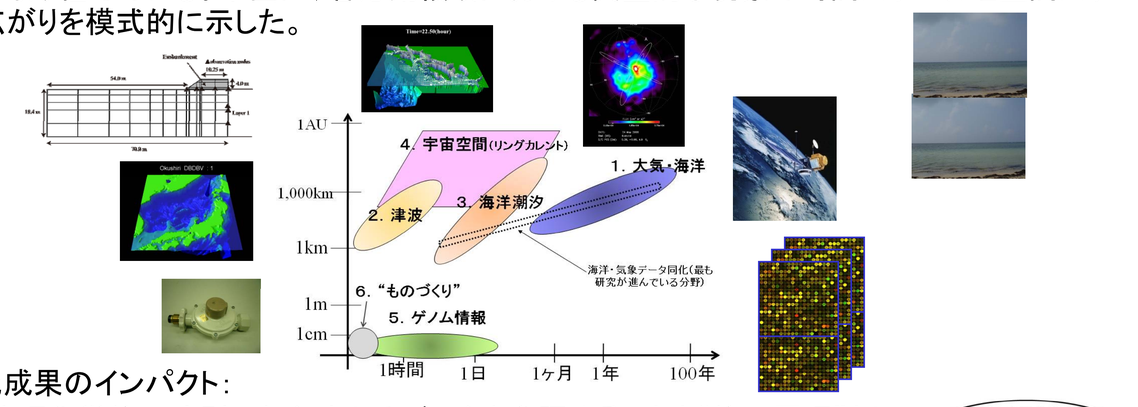


※2：GPGPUを用いた革新的高速化　（林、斉藤らと）

GPGPUとはGeneral-purpose computing on graphics processing unitsの略で、画像処理を専門とする補助演算装置であるGPUを画像処理以外の一般的な計算に用いる技術の総称である。GPUのクロック数自体はCPUと比較すると見劣りする（1.3GHz程度。CPUの約1/2～1/3）が、一つのインストラクションを多数の異なるデータ（統計でのデータの意味でなく、計算機科学でのデータの意味）に同時に適用する操作、つまりSIMD型タイプとしては著しく性能が良い。この特性をいかした統計計算の高速化に取り組んでいる。



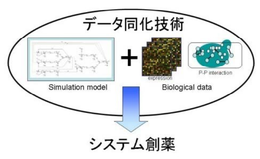
■適用分野：データ同化手法の研究を行うには、具体的なシミュレーションモデルとデータセットの二つ、つまり具体的テーマの選定が必要である。我々は主に、大気・海洋、津波、海洋潮汐、宇宙空間（リングカレント）、ゲノム情報、ものづくり、環境経済の7つの領域における新しいデータ同化実験に取り組んでいる。下図に、各応用領域における典型的な現象の時間スケールと空間スケールの広がりを模式的に示した。



■研究成果のインパクト：

1）物理化学等の法則でなく、むしろデータを参照することを“第一原理”とした、初期条件・境界条件・パラメータの最適化と計算資源の集中化の研究を行うことで、逆問題解析の革新的アルゴリズムの開発に貢献。

2）多品種少量生産を基本とする製品開発現場でのステップの簡略化や、患者一人一人に合った治療サービスの提供と期間の短縮化等の、シミュレーション技術の個人化技術への発展的進化に貢献。



4．マイクロマーケティング

■個人化技術：個人化技術の実現に必要な基本的要素技術は、ケース数Nよりも属性変数の数Pが圧倒的に大きい状況、つまり『新NP問題』の克服法である。この困難を減じるためには、ある特徴量（属性）で似た値をとるものは、他の特徴量でも似た値をとることが期待できるといったようなさまざま先見的知識を活用し、スパースな情報空間を確率的に埋めていくことが肝である。

■応用例：スーパーマーケットのPOSデータの解析と消費者行動のモデル化を通して、個人に特化したサービスが提供できるシステム構築の研究を行っている。