様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月 25 日現在

機関番号：　８２１１８
研究種目：　基盤研究（C）
研究期間：　2006～2009
課題番号：　１８５００２２３
研究課題名（和文）　ベイズ統計によるデータ解析の研究

研究課題名（英文）　Study on data analysis method based on Bayesian statistics

研究代表者
柴田　章博（AKIHIRO SHIBATA）
大学共同利用機関法人高エネルギー加速器研究機構・計算科学センター・研究機関講師
研究者番号：３０２９０８５２

研究成果の概要（和文）：ベイズ統計に基づく実験データの解析法、特に、隠れ変数を伴うデータモデリングやモデルパラメータの推定法やそのアルゴリズムついて焦点をあてて研究を行なった。具体的には将来の線形加速器実験を対象として、統計モデルとパラメータ推定法の提唱やその改善について研究を行なった。

研究成果の概要（英文）：We study the Bayesian method for data analysis of experiment. We focus on the data modeling with hidden parameters and study the inference methods of model parameters and these algorithms.　In practice, we study with the experiment in the future linear accelerator. We proposed statistical models and inference methods and studied improvement of the method.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,900,000	0	1,900,000
2007 年度	700,000	210,000	910,000
2008 年度	500,000	150,000	650,000
2009 年度	500,000	150,000	650,000
年度			
総　計	3,600,000	510,000	4,110,000

研究分野：総合領域
科研費の分科・細目：情報学、統計科学
キーワード：数値統計、ベイズ統計、EM アルゴリズム、多変量解析、データマイニング、高エネルギー実験

１．研究開始当初の背景

データマイニングや情報処理に通知統計に基づく研究やその応用研究が様々な分野で行なわれている。例えば、高エネルギー実験におけるデータ解析は、データマイニングの過程であり、様々な検出器でとらえられた事象を蓄積し、その蓄積されたデータの中から興味あるデータの組(事象)を何段階かに分けて抽出し、データモデルの解析をとおして物理現象を理解する。デー

タ解析の各ステップにおいても、ニューラルネットワークをはじめとしてさまざまなデータマイニングの手法が使われてきた。近年の高エネルギー実験は、装置の複雑化・巨大化に伴って、解析すべきデータも複雑化・巨大化となるため、効率よい開発のためには、実験装置やデータ解析法開発が不可欠である。また、急速な計算機の発展は、データのデジタル処理を加速し、大規模かつ高速なデータ処理が必要とされる。

一方、ベイズ統計の枠組みは、データ及びモデルパラメータを確率変数として統一的に扱うことができ、また従来のパラメータ・フィットで使う最尤法を含むため系統的かつ厳密な解析を行うことができるが、一方で、膨大な計算量を必要としていた。近年の計算機の急速な発展と普及に伴い、データマイニング及び情報処理においてベイズ統計に基づく数値統計の研究が急速に発展を遂げ、その成果の様々な分野への応用が行われている。

２．研究の目的

本研究の主たる目的は、データマイニング及び数値シミュレーションを活用し、複雑な構造をもったデータの詳細分析法を開発することである。また、数値解析に必要なさまざまなアルゴリズムの開発をあわせて行い従来手法の改善を行う。

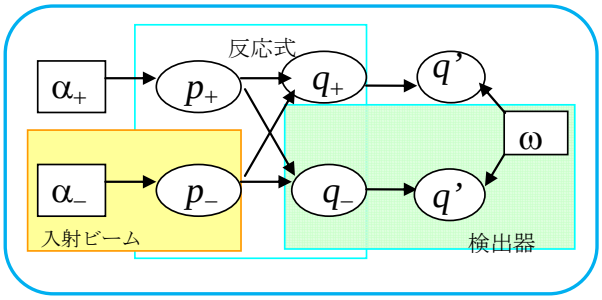
具体的なテーマとして、高エネルギー実験や構造解析などの加速器実験を題材として、基本的な手法として重要であり、また次世代の加速器実験で手法改善や開発が求められているテーマを取り上げ、実験解析に応用できるよう枠組みとなるように研究をすすめる。

ベイズ統計の特徴であるデータとパラメータを統一的に取り扱いや条件付確立は因果関係と対応付けたクラフィカルモデルを活用した解析を行うとともに、ベイジアン・ネットワークの学習（パラメータ推定）理論などの統計学のさまざまなテクニックを活用、及びそれに付随するアルゴリズムの研究を行う。

３．研究の方法

ベイズ統計の枠組みを用いることで、対象としている変数全体の関係を統一的に扱い、データやモデルパラメータにあるあるデータの因果律（物理のモデル法則）を対応させ分析する。ベイジアン・ネットワークに現れる条件付確率分布関数を連続変数における分布関数に拡張し、グラフィカルモデリングの手法を用いて、視覚的・直観的に複雑なデータ構造捉え統計モデルの分析を系統的に見通しよく行う。

高輝度電子・陽電子線形加速器の衝突実験におけるルミノシティー推定法を題材に、事象のスペクトラムの推定法を研究した。図は、粒子衝突実験のデータモデルの例を示す。（連続変数におけるベイジアンネットワークに対応）。



粒子加速、反応式、測定器のモデルに分解することができ、データ解析に登場する（隠れ変数を含む）変数全体について分析を行うことができる。ベイズ統計の特徴であるデータとパラメータを統一的に取り扱いによって、各プロセスに分解し統計モデルの検証を行うことができる。分解したプロセスごとのモデルの構築と評価をイベント生成器と連携して行った。

解析用のデータ生成は、素粒子反応のイベント生成器を活用した。模擬実験(シミュレーション)でデータ生成することで、事象スペクトラムやノイズの厳密にコントロールの下にモデル評価ができる。このとき、ある実験のベイジアンネットパラメータを与えれば、この実験を行ったときのデータ分布のシミュレーションの結果（順問題）を表現していることになり、ベイジアン・ネットワークのパラメータ推定を行うことは、測定器の誤差などを含めた計全体のモデルパラメータを決定する逆問題に対応する。

一方、統計解析の観点に立てば、さらには,ベイジアン・ネットワークはニューラルネットワークとも対応付けることができるので、実験解析をデータマイニングや多変量解析など情報科学におけるさまざまな学習（パラメータ決定）理論と応付け

て分析を進める。

４．研究成果

電子・陽電子弾性散乱の事象を用いたルミノシレィー測定モデルについて、グラフィカルモデル(連続変数を含むベイシアンネットワーク)との対応を行うことによって詳細な粒子反応のモデルとして記述した。ビーム生成、反応、測定の各プロセスに分解し因果律を考慮した、厳密な統計モデル構築を行った [1].

各プロセスの分解されたモデルは、直接観測されないパラメータを含むモデルの評価が必要とされ、EMアルゴリズムに基づくモデル評価を行いその有効性を確かめた [1][2] .

厳密な粒子反応の統計モデルを構築とベイズ統計に基づく推定を用いることで、従来の研究におけるパラメータ推定よりはるかに少ない(実用的な)データ量でパラメータ推定を行うことができることを示した。また、(加速器パラメータを非対象な設定にするなど)推定するパラメータ数を増やした環境下においても、精度よくパラメータ推定が可能であり、従来提唱の方法を大きく改善するものであることを示した。[1][3][6].

EMアルゴリズムによるEMベイジアンネットの評価は、高次元積分などCPU資源を必要とするため、事象の独立性を利用したMPIによる並列化アルゴリズムの実装し高い効率化が可能であることを確認した[3]。また、高次元積分の評価法の改善として、積分モジュールの並列化やモンテカルロ法のアルゴリズムの改善について検討を行った[8]

検出器を配置できないなどにより測定データに欠損が生じ、粒子反応の事象を再構成ができない場合や解析対象と酷似する事象が発生する場合について、推定法の検討及び系統的な誤差評価法についても検討を行った。(1)個々の測定器誤差が比較的大きな場合におけるパラメータ推定(2)データ欠損や解析事象と酷似するデータの混在する場合のフィルターリングとパラメータ推定法(モデルセレクション)など。擬似データ除去や欠損を伴うデータの取り扱いを確立することは、精密実験を行なう上で重要であり、現在も、統計モデルの解析を継続中である。

多数の隠れ変数を含むプロセスとして時系列のデータに対するパラメータ推定の適用について検討を行った. 多数の隠れ変数を含むモデルは、多次元積分の実行が必要とされ、統計力学の手法やモンテカルロ積分法を併用し評価する方法を提案した。[7]

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　4 件）

(1) 柴田章博、村上直、インターネット通信のサービス別DFA解析　交通流のシミュレーションシンポジウム(論文集) p65-68 2009 査読あり

(2) Shibata, S.Kawabata, J.Fujimoto, Y.Kurihara, T.Watanabe. "An inference method of luminosity spectrum in future e+e- linear collider." Physics letter B645. 12-18 (2007) 査読あり

(3) Akihiro Shibata, “An inference method of luminosity spectrum in a future high luminosity e+e- linear collider”, Published in PoS(ACAT)048, 2007.査読あり

(4) 柴田章博 "EM アルゴリズムによる高輝度電子陽電子加速器におけるルミノシティースペクトラムの推定", Proceedings of 9th Workshop on Information-Based Induction Science. 232-237 (2006) 査読あり

〔学会発表〕（計　4 件）

(1) 柴田章博 「モンテカルロ法による高次元積分の改善」 日本応用数理学会 2009 年度年会 2007 年 9 月 28 日～9 月 30 日 大阪大学豊中キャンパス

(2) 柴田章博 「速度最適化 (OV) 模型のパラメータの時系列データからの推定」 日本応用数理学会 2007 年度年会 2007 年 9 月 15 日～9 月 17 日 北海道大学

(3) Akihiro Shibata, “An inference method of luminosity spectrum in a future high

luminosity e+e- linear collider", XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2007) April 23-27, 2007 at Nikhef, Amsterdam, Netherlands

(4) 柴田章博 "EM アルゴリズムによる高輝度電子陽電子加速器におけるルミノシティースペクトラムの推定" 第 9 回情報論的学習理論ワークショップ(IBIS2006)、2006 年 10 月 31 日,～11 月 2 日(木)、大阪大学中之島センター佐治敬三メモリアルホール

６．研究組織

(1)研究代表者
柴田　章博（Akihiro Shibata）
大学共同利用機関法人高エネルギー加速器研究機構・計算科学センター・研究機関講師
研究者番号：30290852