基盤研究(S)(代表：谷口正信)

「広汎な観測に対する因果性の導入とその最適統計推測論の革新」の目指すもの

本研究では、Granger 因果性などを含む高度な因果指標を極めて一般的な乖離度から導入して、データ科学における今まで捉えられなかった潜在要因の統一的指標を提案する。観測対象も従来の統計データだけでなく、高次元時空間過程、グラフ・ネットワーク、遺伝子、トポロジカルデータ等にも適用する。この統一的指標を以下、一般化因果性指標と呼ぶことにする。また前述の観測対象も、一般化観測データと呼ぶ。本研究の主題は、一般化観測データからの一般因果性指標の統計的推測理論の構築とその広汎な分野への新しい潜在要因抽出法の提案である。具体的には局所漸近正規性(LAN)に基づいた一般化データに対して一般化因果性指標の統計的最適推測論の構築を基礎とする。推測法としては、尤度原理だけでなく、経験尤度推測、高次元推測等、膨大な手法提案、検証を行い、高次元データ、生体・遺伝子データ、グラフィカル・トポロジカルデータ等の観測に対して我々の構築する最適統計推測法を適用し、広汎な分野の現象に対する新たな潜在指標を洗い出し、それにより予知、要因分析、コントロール、リスク管理に貢献する。

時系列（定常確率過程）｛X(t)｝の過去から未来時点への平均２乗予測誤差と、別の時系列（定常確率過程）｛ Y(t)｝の情報を{X(t)｝に加えた対応する平均２乗予測誤差を比較して後者が前者より小さくなるとき、Y(t) から X(t) へ因果性があるという。この因果性のコンセプトは、ノーベル経済学賞受賞者の Granger が 1969年に導入した。当初は計量経済学の分野での話で、２次定常過程に対する限定的なものであった。それが、今やグラフやネットワーク、遺伝子まで　極めて広汎に応用されてきている。過去、谷口は、金融解析の分野で提唱された非線形時系列モデルを癲癇患者の脳波と筋電波に適用し意味ある相関を見出した。従って経済・金融の分野から自然科学や異質の分野への大きなインパクトを実感、実証してきている。近年、金融時系列解析では、多くの金融時系列は裾の重い分布を持つと言う実証分析があり、２次モーメントを持たないと想定される。従って平均２乗誤差で定義されるGranger 因果性は定義できない。そこで、２次モーメントを持たないデータに因果性は導入できるか？　そこで、本研究では、このようなデータに対しても記述できる極めて一般的な乖離度を導入して、これに基づく因果性指標を定義する。また近年、諸分野で高次元のデータ観測がなされており、高次元データの統計解析では、従来の手法が有効でないことが知られている。高次元観測に最適因果性推測が可能であろうか？

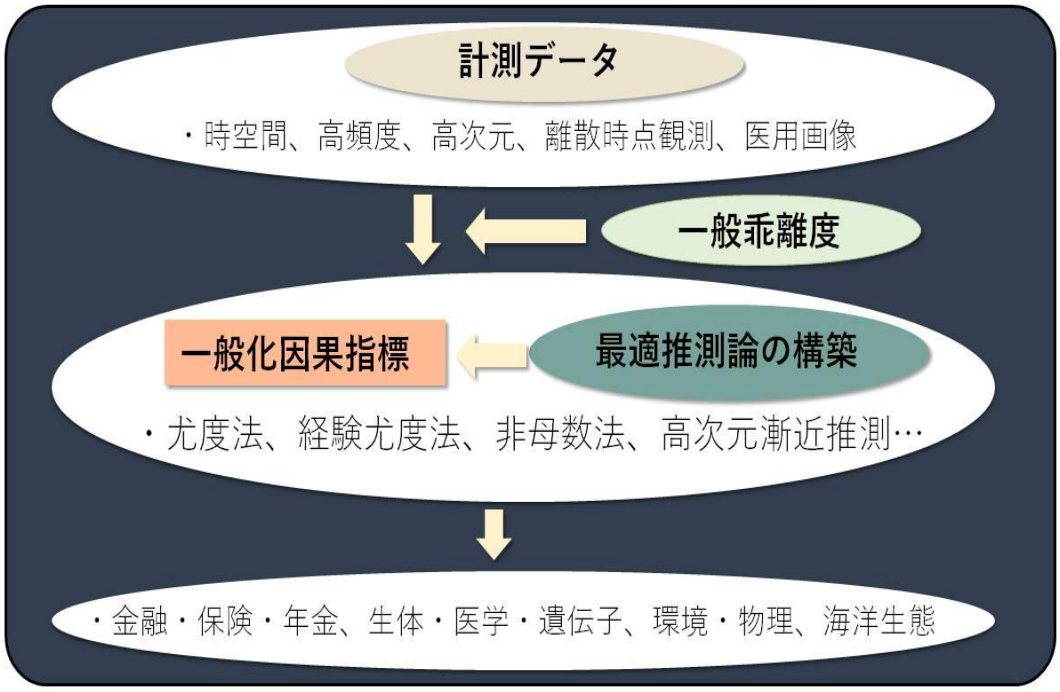
そこで本研究では、高次元確率過程に対して一般化乖離度から高次元確率過程観測の間の一般化因果性を導入する。この推測は、高次元統計推測になるので、従来の統計推測とは異なってくる。この状況で、最適統計推測理論を構築する。以上の理論貢献は統計数理分野でも、全く新奇なものであり種々の新しい結果が得られると思われる。本研究では、Granger 因果性などを含む 高度な因果指標、潜在指標を極めて一般的な乖離度から導入して、データ科学における潜在要因の統合的、統一的指標を提案する。観測対象も従来の統計データだけでなく、高次元時空間過程、グラフ・ネットワーク、カテゴリカルデータ、トポロジカルデータ等に適用することが可能か？　例えば医用画像を時空間データと捉えれば我々の乖離度は因果性をスペクトル構造やトポロジカルな構造から見出しうると思われ、結果として疾病の予知等が可能となろう。また、遺伝子データ等に見られる多くの高次元データは、変数間の関係がスパースではなく、非常に強い相関関係がもつことが、理論的に突き止められている。これは、高次元データの背後にはデータ変動を制御する巨大な潜在因子があると認識されていて、その因子を余すことなく抽出することが、応用上非常に重要だということである。高次元データでは巨大なノイズが入り乱れ、これから余すことなく潜在因子を抽出することができるか？　そこで、潜在因子を壊さずに巨大なノイズだけを除去する新たな非正則推測理論を展開し、多様なビッグデータに対応した新たな潜在因子分析法と高次元漸近理論を整備する。統計数理研究所では、他の研究機関にはない重要なデータを利用可能な状態で保有している。例えば、非公開データのデータを入手した統合データベース、インターネット上の金融に関するデータを自動取得したビッグデータ等である。独自の統計モデルを作成し、方法論の開発と実証研究を同時に推進し、企業データを用いて企業のデフォルトと、企業の成長に関する我々の新しい因果構造を明らかにすることが可能か？が求められているが、これは前述の研究の流れより十分可能である。

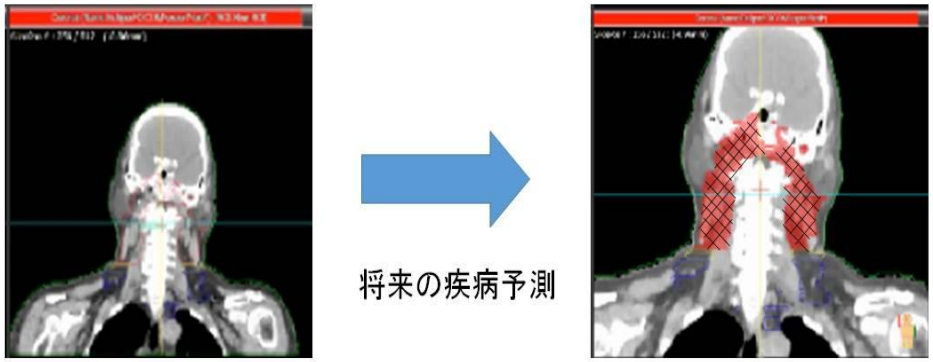
本研究では、Grangerによって提案された因果性などを含む 高度な因果指標、潜在指標を極めて一般的な乖離度から導入して、データ科学における今まで捉えられなかった潜在要因の統一的指標を提案する。観測対象も従来の統計データだけでなく、高次元時空間データ、グラフ・ネットワーク、遺伝子、医用画像、海洋生態、トポロジカルデータ等に適用することを狙いとする。この統一的指標を以下、一般化因果性指標と呼ぶことにする。また前述の観測対象も、一般化観測データと呼ぶ。本研究の目的は、一般化観測データからの一般因果性指標の統計的推測理論の構築とその広汎な分野への応用である。具体的には一般化データに対して局所漸近正規性(LAN)に基づいた一般化因果性指標の統計的最適推測理論の構築を基礎とする。推測法としては、尤度原理だけでなく、非母数的推測、経験尤度推測、ベイズ推測、高次元推測等、膨大な手法提案、検証を行う。さらに、遺伝子データなどの背後には、データ変動を制御する巨大な潜在要因（因子）がある。高次元データ特有の巨大なノイズに埋もれた因子を余すことなく抽出するためノンパラメトリックな非正則推測理論を展開し、多様なビッグデータに対応した新たな潜在因子分析法と漸近

理論を整備する。また、経済事象等に現れる一般の連続時間型確率過程モデルに対して、現実的なデータ計測状況を考慮した「離散観測に基づく推測論」を展開し、高度な数理モデルを実際に応用可能にするための方法論を構築する。上記で理論的に構築された方法論やモデルについて、実データを用いて有効性実証を行う。統計数理研究所では金融時系列データや企業の財務データ、デフォルト情報データ、政府調査の個票など、他の大学・研究機関、データベンダーより充実したデータベースを保有している。これらのデータを利用することにより、本研究で得た知見があるデータのみに有効なのか、普遍的に有効なのかを確認する。さらには、従来のデータとは異なる形状のグラフィカル・トポロジカルデータ等の計測法を検証し、我々の構築する最適統計推測法を適用し、一般化因果性指標の推測を行い従来見出せなかった現象に対する潜在指標を洗い出し、予知、計画、コントロール、リスク管理に貢献する。従って従来にない独自性と創造性を持った研究推進である。上記研究成果から得られる諸分野へのインパクトとしては、

（i）まず、統計数理において、このような一般化因果性指標の提案は、従来の因果性が定義出来なかった場合への成果も含み、それ自体新しいもので、それを一般化データから、LAN 性に基づいての統計的最適推測理論の構築は、統計数理理論への貢献として大変 innovative である。　この理論成果の応用としては、例えば、（ii）医用画像解析においては時空間の因果性をスペクトル構造やトポロジカルな構造から我々の因果性指標を推測し、脳機能の因果性、癌やアルツファイマー病等、将来の疾病予測も行う。（iii）我が国の巨大な年金積立金のポートフォリオは、国債、国内株式、外債、外国株式の４資産で運用されているが、この背後に社会、環境、テクノロジー、国外要因が当然影響しており、年金ポートフォリオへの潜在的因果性を抽出することは、国民生活へのリスク軽減につながると思われる。（iv）遺伝子データを含む多様な高次元データは、元来、巨大なノイズが入り乱れている。そのような巨大なノイズを除去し、潜在要因を浮き彫りにすることで、多様で大規模なデータにおいても、潜在要因の推測を可能にする。（v）また、統計研究所にある非公開の膨大なビッグデータからの新たな潜在要因抽出も行い社会リスクの予測や政策提言につなげる。（vi）保険数理におけるリスク評価の問題は、近年の新ソルベンシー規制の導入があって緊急の話題であり、複数の保険資産のモデルとして連続時間の高次元確率過程モデルの因果推測論は重要である。これにより、統合的リスク管理のツールとして従来より精密なリスク評価の提案が期待できる。以下は本研究目的と医用画像の解析のイメージ図である（医用画像は東京女子医大西尾教授による）。

イメージ図


計測データ
・時空間、高頻度、高次元、離散時点観測、医用画像
一般乖離度
一般化因果指標
最適推測論の構築
・尤度法、経験尤度法、非母数法、高次元漸近推測…
・金融・保険・年金、生体・医学・遺伝子、環境・物理、海洋生態


将来の疾病予測