

博士論文審査報告書

論 文 題 目

統計的モデリングに基づく
マハラノビス・タグチ・システム
The Mahalanobis-Taguchi system
based on statistical modeling

申 請 者

大久保	豪人
Masato	OHKUBO

経営システム工学専攻 統計科学研究

2018年2月

田口玄一博士が開発し、世界中で用いられているタグチメソッドの代表的な方法論の一つにマハラノビス・タグチ (MT) システムがある。これは我が国の製造業を中心に広く普及している。しかし、MT システムを現行プロセスに則って実問題に適用する際、適切な分析ができない場合がある。そこで本研究では、実問題を適切に分析するため、統計的モデリングの枠組みに則った MT システムの新たな解析プロセスを提案している。

ある個体が正常な個体か異常な個体かを判定する問題を考える。正常な個体は同一のパターンであることを想定できる一方で、異常な個体には様々なパターンが存在するものとする。このとき、元の問題を「ある個体が正常な群に属するか否かを判定する問題」として再定義する。これを“異常検知問題”と呼ぶ。この定式化により、未知パターンの異常個体も検知可能となる。

田口博士は異常検知のための実用的な方法論として MT システムを提唱した。MT システムはホテリングの多変量管理図に独自のアイデアを加えて発展させた方法論である。MT システムとホテリングの多変量管理図は、マハラノビス距離とその派生距離に基づく異常検知方式である点で共通している。一方、両者を比較すると、MT システムは母集団の分布に対して全く仮定をおいていない等、汎用性が高い解析プロセスと考えられる。

しかし、実問題に対して MT システムを適用する際、必ずしも適切な分析ができるとは限らない。現行プロセスは統計的モデリングの観点から言えば、(1)モデル設定、(2)母数の推定、(3)モデル評価の各ステップに問題がある。

第 1 に「(1) モデル設定」では、変数の組合せに対応した統計モデルを設定しているため、未知パターンの異常が検知できない危険性がある。一般に異常検知問題では、未知パターンの異常検知も目的となるため、変数選択は保守的に行った方がよい。すなわち、事前には異常検知に役立つか不明である変数も未知パターンの異常検知には役立つ可能性がある。そのため、異常検知に役立たないことが自明な場合を除いて変数選択は実行しない方がよい。

第 2 に「(2) 母数の推定」についてはデータの特徴を考慮せず常に同じ推定法を使用することが問題である。現行プロセスでは標本マハラノビス距離の計算に標本平均ベクトルおよび標本共分散行列を用いている。すなわち、常に経験分布に基づくモーメント法によってパラメータを推定している。

第 3 に「(3) モデル評価」については解析の目的によらず常に予測力を規準としてモデル評価を実行していることが問題である。現行プロセスでは信号対雑音 (SN) 比に基づく変数選択を実行することで、予測精度が向上するような変数の組合せを選び出している。

以上を踏まえて本研究では、統計的モデリングの枠組みに基づく MT システムの新たな解析プロセスを提案している。具体的には、統計的モデリングを通して統計モデルを推定後、そのパラメータの推定量を用いて標本マハラノビス距離を計算するものである。このようなプロセスの導入により、解析目的に合致した適切な統計モデルの評価およびデータがもつ特徴を考慮した

統計モデルのパラメータ推定が実行できる．そして，様々な実問題に対して解析目的やデータの特徴を考慮した適切な分析が実現できるようになる．

本研究では，MT システムを小標本データ，高次元データ，汚染データに適用する場合を想定した統計的モデリングの枠組みに基づく新たな解析プロセスを提案している．そして，提案プロセスの有用性を実データ解析および数理的な検討を踏まえたモンテカルロ・シミュレーションを通して確認している．第 1 に小標本データの場合，提案プロセスにより異常検知性能が向上するだけでなく，異常原因の特定に有益な知見を獲得できることを示している．第 2 に高次元データの場合，提案プロセスにより高次元データの特徴を活かした異常検知を実現できることを示している．第 3 に汚染データの場合，提案プロセスによりミスラベルのデータが大量に混入する状況でさえも安定的な異常検出を実現できることを示している．

第 1 の小標本データを対象とした提案プロセスの概要は次の通りである．一般に小標本データ解析では，学習データへの過剰適合が発生するため，推定すべきパラメータの削減が必須である．加えて，異常検知問題では，未知パターンの異常検知性能の維持も重要となる．そこで本研究では，未知パターンの異常検知性能を考慮した小標本データ解析プロセスを提案している．具体的には，確率的主成分分析モデルあるいはガウシアン・グラフィカル・モデルをモデル族に設定し，その母数推定とモデル評価を行っている．このとき，多重共線性が頻発するため，母数の推定法にはその問題を回避する方法を導入している．またモデル評価には，異常原因の追究の知見獲得の観点重視したベイズ情報量規準 (BIC) タイプの評価規準を使用している．

第 2 の高次元データを対象とした提案プロセスの概要は次の通りである．高次元データ解析では，サンプル数 n よりも変数の次元 p が遥かに多いデータ ($p \gg n$) を取り扱う必要がある．しかし， $p \gg n$ の状況では，現行プロセスのままでは解析不能となるため，新しい解析プロセスや異常検知方式自体の改良が必要となる．そこで本研究では， $p \gg n$ の状況で計算可能な 2 種類の距離を取り上げたうえで，その距離を用いた新たな解析プロセスを提案している．具体的には，高次元データにおいて慣例的に想定される母共分散行列に対する固有値モデルをモデル族に設定したうえで，その母数の推定とモデル評価している．このとき， $p \gg n$ の状況でも精度よく推定するため，スパース主成分分析に基づく推定法を使用している．またモデル評価には BIC タイプの評価規準を使用している．

第 3 の汚染データを対象とした提案プロセスの概要は次の通りである．汚染データ解析では，十分なサンプル数が確保されている場合でさえ，現行プロセスは良好な異常検知性能を発揮できない可能性がある．単位空間に混入したミスラベルのデータの影響を受け，母数の推定が困難となるためである．そこで本研究では，統計的モデリングのステップの中でも母数の推定に焦点をあて，ミスラベル・データの混入に対してロバストな推定法の導入効果に

ついて検証している．そして実務的な動向を考慮して，ミスラベルのデータが大量に混入する場合でも安定的な異常検出を実現する解析プロセスを提案している．具体的には，多変量正規分布をモデル族に設定したうえで，ガンマ・ダイバージェンスに基づくロバスト推定法を母数の推定に使用している．

近年，情報通信技術の進展に伴って，様々な産業分野においてセンサーやスマートデバイスから膨大なデータが取得・蓄積されている．特に製造業では，このように蓄積された履歴データを活用し，設備機器の異常検出を実現しようという動きがある．本論文で提案された解析プロセスを用いると，様々な実問題に対して適切な分析ができるため，MT システムの適用対象は格段に広がり，実用的な異常検出の実現につながることを期待できる．

本論文は7つの章から構成されている．第1章では，本研究の背景と目的を述べている．第2章では，MT システムについて概説している．第3章では，統計的モデリングの枠組みに基づくMT システムの新たな解析プロセスの概要を示している．続く第4章，第5章，第6章では，小標本データ，高次元データ，汚染データの各々に対してMT システムを適用する場合を想定した新たな解析プロセスを提案している．そして，提案プロセスの有用性を実データ解析およびモンテカルロ・シミュレーションを通して評価している．最後に第7章では，結論を述べている．

本研究で提案された解析プロセスは，異常検出問題に対して新たな方向性を与えるものである．また，経営工学分野の多くの応用事例への適用が可能である．さらに，異常検出問題に関する研究のさらなる発展も期待できる．以上より，本論文は現代的な統計科学の発展に大きく寄与する研究であり，博士（工学）早稲田大学の学位論文として価値あるものと認める．

2018年2月

審査員

主査 早稲田大学教授 工学博士（大阪大学） 永田 靖

早稲田大学教授 博士（工学）早稲田大学 後藤 正幸

早稲田大学教授 博士（工学）早稲田大学 椎名 孝之

早稲田大学准教授 博士（情報科学）大阪大学 蓮池 隆