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

2018 年　2 月

大久保　豪人

Masato OHKUBO

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

2018 年　2 月

早稲田大学大学院　創造理工学研究科

経営システム工学専攻　統計科学研究

大久保　豪人

Masato OHKUBO

目次

第 1 章　序論

タグチメソッドの分野において Mahalanobis-Taguchi（MT）システムと呼ばれる方法論が提唱され，我が国の製造業を中心に広く普及している．MT システムは従来の多変量解析法にタグチメソッドの概念を導入した一連のデータ分析法である．しかしながら，MT システムを現行プロセスに則って実問題に適用する場合，実務的な需要に対応した適切な分析ができるとは限らない．そこで本研究では，MT システムへの実務的な需要を踏まえたうえで，実問題を適切に分析するための新たな解析プロセスを提案する．

本論文の内容は大久保・永田 (2012)，大久保・永田 (2015)，大久保・永田 (2017)，Ohkubo and Nagata (2017a)，Ohkubo and Nagata (2017b)に基づく．

1.1.　本研究の背景

タグチメソッドの代表的な方法論である MT システムは，品質管理，医療，ビジネス等の様々な分野において重要な役割を担っている．例えば，品質管理の分野では，製品の外観検査（間ヶ部ら (1998)），設備機器の状態監視（福島ら (2009)）等に応用されている．医療の分野では，健康診断（兼高 (1987)）や肝疾患の病態評価（中島ら (2004)）等に応用された事例がある．さらにビジネスの分野では，経営状態評価（芝野・安永 (1999)）や作業者の能力評価（高橋ら (2003)）等に応用された事例がある．立林(2013) によれば，2012 年時点での MT システムに関する事例発表数は 250 件を超える．また，近年ではセンサーデータを利用した異常検知システム（高濱・三上 (2012)）に MT システムが使用され，実務家の注目を集めている．

MT システムには分析目的に対応した様々な分析方式が数多く存在する．その中でも主要な方式として，田口玄一博士が提案した Mahalanobis-Taguchi（MT）法（例えば，Taguchi and Jugulum (2002)），田口 (2002)の Mahalanobis-Taguchi Adjoint（MTA）法，田口 (2006)の Recognition Taguchi（RT）法，田口 (2005)の Taguchi（T）法が挙げられる．また，各方式を改良した方式が提案された主要な文献として，宮川・永田 (2003)，永田・土居 (2009)，稲生ら (2012)，大久保・永田 (2012)が挙げられる．なお，原方式が抱える問題および一連の改良方式については，永田 (2013, 2017)に詳しい．

このように MT システムにはいくつかの改良方式が既に提案されているものの，実問題に対して適用する場合，必ずしも適切な分析ができるとは限らない．言い換えれば，MT システムを使用する場合でも，一般的な多変量解析法と同様に，解析者は分析の目的や対象となるデータに合わせた適切な解析プロセスについて慎重に検討する必要が

ある．しかしながら，次章以降で述べるように MT システムを実問題に適用する場合，解析プロセスについての慎重な検討がなされることなく，分析が実行されてしまうケースも少なからず存在するといえる．

1.2. 本研究の目的

本研究の目的は，MT システムへの実務的な需要を踏まえたうえで，実問題を適切に分析するための新たな解析プロセスを提案することである．具体的には，統計的モデリングの枠組みを MT システムに導入する．次章以降で述べるように現行プロセスは統計的モデリングの観点から言えば，解析目的によらず予測精度を唯一の評価規準として統計モデルを評価しているといえる．また，統計モデルのパラメータ推定法もデータの特徴を考慮せずに常に同じ推定法を使用していると解釈できる．一方，提案プロセスでは，解析目的に合致した評価規準を選択したうえで統計モデルを評価する．加えて，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を実行する．言い換えれば，統計的モデリングの枠組みを MT システムに導入することで，様々な実問題において解析目的やデータの特徴を考慮した適切な分析が実行できるようになるといえる．

1.3. 本論文の構成

本論文の構成は次の通りである．第 2 章では，MT システムについて概説する．また，関連する異常検知方式との比較を通して MT システムの概念について説明する．第 3 章では，統計的モデリングに基づく MT システムの新たな解析プロセスの概要を示す．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，汚染データに対して MT システムを適用する場合の統計的モデリングに基づく新たな解析プロセスを提案する．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する．最後に第 7 章では，結論と今後の課題を述べる．

第 2 章　MT システムによる異常検知

田口玄一博士は異常検知のための実用的な方法論として MT 法，MTA 法，RT 法等のデータ分析方式を提案した．そして，それら一連のデータ分析方式は MT システムと呼ばれるようになった．本章では，MT システムの概要を示したうえで，MT 法による異常検知プロセスについて説明する．また，MT システムに共通する概念を MT 法と関連する分析方式との比較を通して説明する．

2.1.　MT システムの概要

田口玄一博士は当初, 次節で示す異常検知問題を解くための分析方式として MT システム（後の MT 法）を提案した．ところが，その後，MT システムの改良方式である MTA 法や，高次元バイナリ・データを対象とした異常検知方式である RT 法，回帰問題を扱う T 法等が同博士によって提案された．そのため，タグチメソッドの分野では，当初の MT システムを MT 法, 一連の分析方式を総じて MT システムと呼ぶようになった（例えば，立林ら (2008)を参照されたい）．

本研究でも田口玄一博士によって最初に提案された異常検知方式を“MT 法”と呼ぶことにする．一方，本研究では異常検知問題を扱う分析方式について議論するため，特に断りのない限り，“MT システム”という用語は異常検知問題を扱うための分析方式の総称として使用する．具体的な分析方式としては，前述の MT 法，MTA 法，RT 法等が挙げられる．また，それらの改良方式も MT システムに含めるものとする．

2.2.　異常検知問題

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．ただし，正常な個体はほぼ同一のパターンであることが想定できる一方で, 異常な個体には多種多様なパターンが存在するものとする. 言い換えれば, 正常な個体は群をなすと仮定できる一方で, 異常な個体は群をなすとは仮定できない．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. このような問題を“異常検知問題”と呼ぶことにする．異常検知問題として定式化することで，未知のパターンが発生した場合でも，異常な個体を検知することが可能となる．

2.3. MT 法による異常検知

MT 法は異常検知問題を解くための分析方式の一つである．ここで，正常な群が形成する均質な母集団を MT システムでは“単位空間”と呼ぶ．MT 法は単位空間からの逸脱度をマハラノビス距離に基づいて定量化する．そして，マハラノビス距離が大きな値をとるほど，判定対象となる個体の異常度が高いとみなす．以下，MT 法における単位空間上のマハラノビス距離の算出プロセスを中心に説明する．他の解析プロセスについては Taguchi and Jugulum (2002)，立林ら (2008)等を参照されたい．

いま p 次元変数 $\boldsymbol{x}$ が母集団で観測され，母集団からの大きさ n の無作為標本を $\boldsymbol{x}_i$ (i = 1,2, ..., n)とする．また，$\boldsymbol{x}$ の母平均ベクトルおよび母共分散行列を $\boldsymbol{\mu}$ および $\boldsymbol{\Sigma}$ として，その推定量を $\hat{\boldsymbol{\mu}}$ および $\hat{\boldsymbol{\Sigma}}$ と表記する（以降，同様に母数 $\boldsymbol{\theta}$ の推定量を $\hat{\boldsymbol{\theta}}$ と記す）．このとき，母マハラノビス距離 $\Delta(\boldsymbol{x})$および標本マハラノビス距離 $D(\boldsymbol{x}_i)$を

$$\Delta^2(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{2.1}$$

$$D^2(\boldsymbol{x}_i) = (\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}) \tag{2.2}$$

と定義する．ここで，ベクトルあるいは行列の転置を T とした．なお，本論文中のベクトルはすべて縦ベクトルであり，その転置を横ベクトルとして扱う．

同様に各変数の平均を 0，標準偏差を 1 に基準化するとき，p 次元変数 $\boldsymbol{u}$ および大きさ n の無作為標本 $\boldsymbol{u}_i$ (i = 1,2, ..., n)に対する母マハラノビス距離 $\Delta(\boldsymbol{u})$および標本マハラノビス距離 $D(\boldsymbol{u}_i)$は母相関係数行列を $\boldsymbol{\Pi}$ として次式で定義される．

$$\Delta^2(\boldsymbol{u}) = \boldsymbol{u}^T \boldsymbol{\Pi}^{-1} \boldsymbol{u} \tag{2.3}$$

$$D^2(\boldsymbol{u}_i) = \boldsymbol{u}_i^T \hat{\boldsymbol{\Pi}}^{-1} \boldsymbol{u}_i \tag{2.4}$$

MT 法では，単位空間に属する個体を n 個サンプリングし，(2.4)式の基準化後の標本マハラノビス距離を求める．ただし，単位空間に属する n 個の個体から求めた標本平均および標本標準偏差を用いて各変数を基準化した後，$\hat{\boldsymbol{\Pi}}$ として標本相関係数行列を用いて標本マハラノビス距離を計算する．また，判定対象となる新たな個体に対して標本マハラノビス距離を計算する際も，単位空間から既に計算した母数の推定量を用いる．そして，判定対象となる個体から求めた標本マハラノビス距離が，事前に定めた閾値よりも小さいならば単位空間に属する個体（正常な個体）とみなし，大きいならば単位空間に属さない個体（異常な個体）と判定する．

2.4. 関連研究との比較

近年，異常検知問題を扱う数多くの方法論が提案されており，様々な分野へ応用されている．代表的な方法論として，Hotelling (1947)の T^2 管理図を端緒とする種々の多変量管理図（以降，単にホテリング多変量管理図と呼ぶ），Breunig et al. (2000)の Local Outlier Factor（LOF），Schölkopf et al. (2001)の One Class Support Vector Machine（OCSVM）等が挙げられる（例えば，山西 (2009)，井手 (2015)，井手・杉山 (2015)に詳しい）．そして，MT システムは田口玄一博士がホテリング多変量管理図に Signal-to-Noise（SN）比に基づく変数選択や異常原因追及のための変数診断等のアイディアを加えて発展させた実用的な方法論であるといえる．本節では MT システムに共通する概念を MT 法と T^2 管理図との比較によって説明する．

まず，Hotelling (1947)の T^2 管理図では，母集団の分布に対して多変量正規分布を仮定したうえで，母マハラノビス距離を推定する．また，このような厳密な仮定をおくことで，標本マハラノビス距離の標本分布に関する精緻な議論を可能としている．例えば，Tracy et al. (1992)は標本マハラノビス距離が F 分布を定数倍した分布に従うことを示した．この結果を用いれば，検定論の枠組みに基づいて異常検知問題を解くことができる．すなわち，標本マハラノビス距離を検定統計量とみなすとき，ある棄却限界を超えたか否かで異常か正常かを決定する方式が採用できる．

一方，MT 法では，母集団分布に対する仮定をおかないまま，母マハラノビス距離を推定する（Jugulum et al. (2003)）．そのため，標本マハラノビス距離の標本分布に対する精緻な議論の展開は必ずしも容易ではない．実際，閾値設定を例にとっても，田口玄一博士は標本マハラノビス距離の標本分布については言及せず，誤判別に伴うコストを最小化するように決定すると述べている（田口 (1995)）．ゆえに，MT 法はコスト最小化の観点，すなわち予測力の観点から異常検知問題を解くといえる．

第 3 章　統計的モデリングに基づく解析プロセスの提案

本章では, 実問題を適切に分析するために MT システムの新たな解析プロセスを提案する．具体的には，統計的モデリングの枠組みを MT システムに導入する．統計的モデリングの観点から言えば, 現行プロセスは予測精度を唯一の評価規準として統計モデルを評価しているといえる．また，統計モデルのパラメータを常に同じ推定法で推定していると解釈できる．一方，提案プロセスでは，解析目的に合致した評価規準を選択したうえで統計モデルを評価する．加えて，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を実行する．したがって，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可能になるといえる．

3.1.　統計的モデリングの概要

いま母集団の真の分布を $g(\boldsymbol{x})$とするとき，$g(\boldsymbol{x})$を近似する統計モデル $f(\boldsymbol{x})$を推定する問題を考える．ここで, $f(\boldsymbol{x})$がパラメトリックなモデルであることを強調する場合, $f(\boldsymbol{x}|\boldsymbol{\theta})$あるいは$f_{\theta}(\boldsymbol{x})$と記す．また，有限個のモデル候補 $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_H$ のもとでの統計モデルを各々$f(\boldsymbol{x}|\boldsymbol{\theta}_1), f(\boldsymbol{x}|\boldsymbol{\theta}_2), \ldots, f(\boldsymbol{x}|\boldsymbol{\theta}_H)$と記す．このとき，次のような手順でパラメトリックな統計モデルを推定する統計的モデリング法を“モデル選択”と呼ぶ（統計的モデリング全般に関する解説書として，小西・北川 (2004)，安道 (2014)を挙げておく）．

手順(1) モデル設定. モデル候補 $\mathcal{M}_h$ $(h = 1,2, ..., H)$に対応したパラメトリックな統計モデル $f(\boldsymbol{x}|\boldsymbol{\theta}_h)$ $(h = 1,2, ..., H)$を設定する．

手順(2) 母数の推定. 各統計モデルにおいて最適なパラメータ $\boldsymbol{\theta}_h$ を各々推定する. なお, パラメータ推定法は最尤推定法，罰則付き最尤推定法，ロバスト推定法等の中から適切な方法を選択する．

手順(3) モデル評価. パラメータ $\boldsymbol{\theta}_h$ の推定量 $\hat{\boldsymbol{\theta}}_h$ に基づく統計モデル $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_h)$ $(h = 1,2, ..., H)$を事前に定めたモデル評価規準のもとで各モデルを評価する．そして，該当のモデル評価規準が最良となる統計モデルを選択する．なお，モデル評価規準は AIC，BIC，Cross-Validation 等の中から適切な規準を選択する．

3.2.　現行プロセスの問題点

統計的モデリングの観点から言えば，現行プロセスには少なくとも「(2) 母数の推定」

と「(3) モデル評価」のステップに問題があるといえる．ただし，変数選択がモデル選択の一例であって，2^p-1 通りの変数の組合せをモデル候補として，それに対応した統計モデルを設定できることに注意する．

まず，「(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といえる．第 2 章で述べたように，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．ここで，単位空間の分布に対して厳密な仮定をおいていないことに注意すると，現行プロセスは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推定を行っているといえる．

次に，「(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といえる．第 2 章で述べたように，現行プロセスでは 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．

さらに，「(1) モデル設定」についても異常検知問題では未知パターンの異常が検知できなくなる危険性がある．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，可能な限り多くの変数を残しておいた方がよいといえる．

3.3. 提案プロセスの概要

本研究では，統計的モデリング法に基づく MT システムの新たな解析プロセスを提案する．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．このような解析プロセスを実行することで，解析目的に合致したモデル評価規準をモデル評価に使用できるようになる．加えて，データの特徴に合わせたモデルの設定や母数の推定を実行することで，母マハラノビス距離の推定精度が改善され，異常検知性能の向上も期待できる．実際，次章以降で示すように，提案プロセスに則った解析を実行すれば，現行プロセスよりも高い異常検知性能を実現できる．以下，次章以降で提案する解析プロセスの概要を述べる．

第 4 章では，小標本データを対象とした場合の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する．一般に小標本データ解析では，学習データへの過剰適合の問題が発生するため，推定すべきパラメータを削減する方法論が必要となる．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する．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・モデルを仮定したうえで，その母数の推定とモデル評価を実行する．このとき，異常検知問題では観測変数間に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できる方法を導入する．またモデル評価についても，予測精度だけでなく，異常原因の追究のための知見獲得の観点も重視した BIC タイプの評価規準を使用する．

第 5 章では，高次元データを対象とした場合の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する．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 が遥に多いデータ（$p >> n$）を取り扱う必要がある．しかしながら，$p >> n$ の状況では，そもそも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，確率的主成分分析モデルに基づくマハラノビス距離および RT-PC 法で用いる距離を対象として，新たな解析プロセスを提案する．具体的には，高次元データにおいて慣例的に想定されるモデルを仮定したうえで，その母数の推定とモデル評価を実行する．このとき，$p >> n$ の状況でも精度よく母数を推定するため，スパース主成分分析に基づく推定法を使用する．またモデル評価は第 4 章と同様の理由で BIC タイプの評価規準を使用する．

第 6 章では，汚染データを対象とした場合の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する．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るため，母数の推定が困難となるからである．そこで本研究では，母数の推定法に焦点をあて，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する．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知を実現するための解析プロセスを提案する．具体的には，多変量正規分布を仮定したうえで，γ 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する．

第 4 章　小標本データ解析プロセスの提案

本章では，MT システムを小標本データに適用する場合の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する．変数の次元に比べてサンプル数が小さなデータを小標本データと呼ぶ．一般に小標本データを解析する場合，学習データへの過剰適合が発生してしまう．そのため，推定すべきパラメータ数の削減が有用となるものの，これまでの対策では未知パターンの異常が検知できなくなる危険性がある．そこで本研究では，未知パターンの異常検知を考慮した小標本データの解析プロセスを提案する．提案プロセスでは，統計的モデリングの枠組みに基づいて統計モデルを推定することによって，自然に推定すべきパラメータ数を削減できる．そして，提案プロセスは予測と診断の両面で有用であることを数値実験によって示す．

本章の内容は大久保・永田 (2015)および大久保・永田 (2017)に基づく．

4.1.　はじめに

MT 法では単位空間を設計する際，各個体が典型的なサンプルであるか慎重に検討する必要がある．言い換えれば，学習データの真のラベルが正常であるかを技術的に判断していく作業が必須となる．このようなラベリングの作業は高コストであるため，サンプル数に予算制約が存在する場合がある．また，データ自体の希少性が高いため，物理的に入手できない場合も発生し得る．しかしながら，現行プロセスでは，十分なサンプル数が確保されなければ，適切な解析ができるとは限らない．

第 2 章で述べたように，MT 法はマハラノビス距離に基づく異常検知方式である．また，MT 法では母マハラノビス距離の推定量として，標本平均ベクトルおよび標本共分散行列に基づく標本マハラノビス距離を使用することも述べた．この標本マハラノビス距離を用いることの問題の一つは，サンプル数が小さい場合，大きな予測バイアスをもつことである．例えば宮川ら (2007)は，母集団分布に多変量正規分布を仮定したもとで，標本マハラノビス距離の予測バイアスについて検証した．その結果，変数の次元に比べサンプル数が 10 倍以上なければ，標本マハラノビス距離は大きな予測バイアスをもつことを示した．さらに，標本マハラノビス距離が大きな予測バイアスをもつ場合，予測の観点からも問題が発生するといえる．すなわち，本章の数値実験で示すように，異常データの誤判別率が増大してしまう可能性がある．

このような単位空間のサンプル数に起因する問題に対する標準的なアプローチは，変数選択を実行することである．例えば，現行プロセスにおいても SN 比に基づく変数選

択が実施される．また，宮川 (2000)は McCabe (1984)の主変数選択に基づく変数選択を提案している．前述のように，現行プロセスにおける標本マハラノビス距離の予測バイアスは変数の次元に比べてサンプル数が十分でない場合に発生するため，変数選択は本質的な解決策となる．したがって，解析者は異常検知に役立たないことが明らかである変数がないか慎重に検討し，まずはそれを除外すべきであるといえる．

しかしながら，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，可能な限り多くの変数を残しておいた方がよいといえる．

そこで本研究では，未知パターンの異常検知を考慮した新たな小標本データの解析プロセスを提案する．具体的には，第 3 章で示した統計的モデリングの枠組みに基づいて真の分布を近似する統計モデルを次の 3 ステップで推定した後，そのパラメータ推定量を用いて標本マハラノビス距離を計算する．

(1) 第 1 に，真の分布を近似する統計モデルとして，Tipping and Bishop (1999)の Probabilistic Principal Component Analysis（PPCA）モデルあるいは Lauritzen (1996) の Gaussian graphical（GG）モデルを設定する．
(2) 第 2 に，PPCA モデルの場合は最尤推定法，GG モデルの場合は Friedman et al. (2008)の Graphical Least Absolute Shrinkage and Selection Operator（Glasso）等の罰則付き最尤推定法を用いてパラメータを推定する．
(3) 第 3 に，Chen and Chen (2008)の Extended Bayes Information Criteria（EBIC）によるモデル評価を行う．また，その他の規準も併用して技術的な考察を加えた後，最良となる統計モデルを選択する．

このようなプロセスの導入によって，小標本データにおける母マハラノビス距離の推定精度の向上が期待できる．また，提案プロセスは本章の数値実験で示すように，MT 法の目的である予測と診断の両面において有用であるといえる．

4.2. 現行プロセスの問題点

前節で述べたように MT 法を小標本データに適用する場合，母マハラノビス距離の推定精度に関して問題があることが指摘されている（例えば，宮川・永田 (2003)，宮川ら (2007)，永田 (2013)，永田 (2017)に詳しい）．また，該当変数が異常検知に役立たないことが技術的に明らかな場合を除いて，変数選択は避けるべきとの見解も前節で述べ

た．本節では，パラメータ数を削減するための簡便な方法として，Principal Component Analysis（PCA）による変数集約を利用する方法（例えば，松田ら (2002)を参照されたい）を取り上げ，その問題について考察する．

いま母相関係数行列 $\mathbf{\Pi}$ の固有値を $\lambda_1, \lambda_2, \cdots, \lambda_p$ とする．また，各固有値に対応する長さ 1 の固有ベクトルを $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p$ とする．このとき，母マハラノビス距離は

$$\Delta^2(\boldsymbol{u}) = \boldsymbol{u}^T \mathbf{\Pi}^{-1} \boldsymbol{u} = \sum_{j=1}^{p} \frac{\left(\boldsymbol{\xi}_j{}^T \boldsymbol{u}\right)^2}{\lambda_j} \tag{4.1}$$

と変形できる．ここで，標本相関係数行列から出発した PCA に基づく固有値を $\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_p$，各固有値に対応する長さ 1 の固有ベクトルを $\tilde{\boldsymbol{\xi}}_1, \tilde{\boldsymbol{\xi}}_2, \cdots, \tilde{\boldsymbol{\xi}}_p$ とする．このとき，母マハラノビス距離の推定量として

$$D^2(\tilde{\boldsymbol{u}}_i) = \sum_{j=1}^{p} \frac{\left(\tilde{\boldsymbol{\xi}}_j{}^T \tilde{\boldsymbol{u}}_i\right)^2}{\tilde{\lambda}_j} \tag{4.2}$$

を考えることができる．ただし，母集団からの大きさ n の無作為標本 $\boldsymbol{x}_i$ (i = 1,2, ..., n)の全変数を標本平均および標本標準偏差で基準化した個体を $\tilde{\boldsymbol{u}}_i$ とする．このとき，$\tilde{\boldsymbol{\xi}}_j{}^T \tilde{\boldsymbol{u}}_i$ が第 j 主成分得点に対応するため，PCA に基づく母マハラノビス距離の推定量は，主成分得点を対応する固有値の逆数で重みづけた和と解釈できる．

一方，高々q ($q < p$)個の主成分を利用して，標本マハラノビス距離を計算する場合，

$$D_{(q)}{}^2(\tilde{\boldsymbol{u}}_i) = \sum_{j=1}^{q} \frac{\left(\tilde{\boldsymbol{\xi}}_j{}^T \tilde{\boldsymbol{u}}_i\right)^2}{\tilde{\lambda}_j} \tag{4.3}$$

の形で表現できる．(4.2)式と比較すると，第 q+1 以降の主成分得点に関する項が存在しないとわかる．すなわち，第 q+1 以降の主成分軸上に発生した異常は全く検知できなくなる．したがって，PCA による変数集約を用いた場合でも，未知パターンの異常検知に問題が発生する危険性が高い．

4.3. 提案プロセス

本研究では，MT 法における未知パターンの異常検知を考慮した新たな小標本データの解析プロセスを提案する．提案プロセスでは，4.1 節で示したような統計的モデリングを通じて統計モデルを推定後，そのパラメータ推定量を用いて標本マハラノビス距離を計算する．本節では，統計的モデリングにおける各ステップの詳細を示す．

4.3.1. PPCA モデルの設定と母数の推定

いま単位空間の真の分布が多変量正規分布であることが想定できるものとする. 加えて，p 次元ある観測変数 $\boldsymbol{x}$ が高々q ($q < p$)次元の変数で説明可能との事前情報があるとする．例えば，PCA を実行するとき，観測変数が高々q 個の主成分で説明可能と技術的に判断できる場合が本パターンの一例である．本節では，このような事前情報がある場合のモデルの設定と母数の推定について考える．

まず，統計モデルとして Tipping and Bishop (1999)の PPCA モデルを設定する．PPCA モデルとは，p 次元観測変数 $\boldsymbol{x}$ が高々m ($m < p$)次元の潜在変数 $\boldsymbol{z}$ の線形変換と p 次元ノイズ $\boldsymbol{\varepsilon}$ から生成されることを仮定した次のようなモデルである．

$$\boldsymbol{x} = \boldsymbol{\Gamma z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{4.4}$$

ここで，$\boldsymbol{\Gamma}$ は $p \times m$ 行列，$\boldsymbol{\mu}$ は p 次元ベクトルである．また m 次元潜在変数 $\boldsymbol{z}$ は母平均ベクトルがゼロベクトル，母共分散行列が単位行列 $\mathbf{I}_{(m)}$の多変量正規分布に従う．p 次元ノイズベクトル $\boldsymbol{\varepsilon}$ は母平均ベクトルがゼロベクトル, 母共分散行列が $\sigma^2\mathbf{I}_{(p)}$の多変量正規分布に従う．そして，$\boldsymbol{z}$ と $\boldsymbol{\varepsilon}$ は独立な確率変数であるとする．なお，本節では $\boldsymbol{x}$ の各変数は母平均が 0，母標準偏差が 1 に基準化されているとする．

この PPCA モデルのもとでは，多変量正規分布は再生性をもつため，$\boldsymbol{x}$ は多変量正規分布に従う．そして，母平均ベクトルは $\boldsymbol{\mu}$ である．一方，母共分散行列は

$$\boldsymbol{\Sigma}_{\mathbf{ppca}} = \boldsymbol{\Gamma\Gamma}^T + \sigma^2\mathbf{I}_{(p)} \tag{4.5}$$

で定義される．ここで，$\sigma^2\mathbf{I}_{(p)}$は p 次元ノイズベクトル $\boldsymbol{\varepsilon}$ の母共分散行列であることに注意する．すなわち，PPCA モデルにおける母共分散行列は，本質的な共分散構造 $\boldsymbol{\Gamma\Gamma}^T$ とノイズがもつ共分散構造の和となる．なお，本章ではこの関係に着目して，PPCA モデルは多変量正規分布の共分散構造を仮定するモデルとみなして議論を進める．

次に，PPCA モデルにおける母数の推定について考える．本研究では，現行プロセスの自然な拡張となるように PPCA モデルのパラメータを最尤推定法によって推定するものとする．ここで，PPCA モデルのパラメータ $\boldsymbol{\mu}$ の最尤推定量 $\boldsymbol{\mu}_{ML}$ は標本平均ベクトルである．一方，$\boldsymbol{\Gamma}$ および σ^2 の最尤推定量 $\boldsymbol{\Gamma}_{\mathbf{ML}}$ および σ^2_{ML} は

$$\boldsymbol{\Gamma}_{\mathbf{ML}} = \widetilde{\boldsymbol{\Xi}}_{(m)}\left(\widetilde{\boldsymbol{\Lambda}}_{(m)} - \sigma^2_{\mathrm{ML}}\mathbf{I}_{(m)}\right)^{1/2} \tag{4.6}$$

$$\widetilde{\boldsymbol{\Xi}}_{(m)} = \left(\widetilde{\boldsymbol{\xi}}_1, \widetilde{\boldsymbol{\xi}}_2, \cdots, \widetilde{\boldsymbol{\xi}}_m\right) \tag{4.7}$$

$$\widetilde{\boldsymbol{\Lambda}}_{(m)} = \mathrm{diag}\left(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \cdots, \widetilde{\lambda}_m\right) \tag{4.8}$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{p-m}\sum_{j=m+1}^{p}\tilde{\lambda}_j \tag{4.9}$$

で与えられる（Tipping and Bishop (1999)）．このとき，(4.5)式のパラメータを最尤推定量で置き換えた推定量は $m = p-1$ の場合，標本共分散行列をスペクトル分解した形式に書き直すことができる．すなわち，$m = p-1$ の場合，PPCA モデルのもとでの母共分散行列の最尤推定量は標本共分散行列の値と一致する．したがって，PPCA モデルのもとでパラメータを最尤推定する場合，現行プロセスの自然な拡張になるといえる．

4.3.2. GG モデルの設定と母数の推定

いま単位空間の真の分布が多変量正規分布であることが想定できるものとする. 加えて，母共分散行列あるいは母相関係数行列の逆行列（以降，両者の区別なく“母精度行列”と呼ぶ）の非対角要素の大半がゼロの値をとるとの事前情報がある場合，パラメータ数を精度行列の非ゼロ要素数で制御可能な統計モデルを考えることができる. 例えば, グラフィカル・モデリングを実行するとき，疎なグラフで変数間の関係を記述可能と技術的に判断できる場合が本パターンの一例である．本節では，このような事前情報がある場合のモデルの設定と母数の推定について考える．

Lauritzen (1996)は母平均ベクトル $\boldsymbol{\mu}$ および母精度行列 $\boldsymbol{\Omega}$ をパラメータにもつ多変量正規分布を $N(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$とするとき，$\boldsymbol{\Omega}$ の非対角要素の大半がゼロであるようなモデルを GG モデルと呼んだ. この GG モデルは Dempster (1972)が提案した共分散選択を端緒とするグラフィカル・モデリングを多変量正規分布の仮定のもとで考察するために導入したモデルであるといえる．また，GG モデルに基づいてグラフィカル・モデリングを実行する場合，特に Gaussian graphical（GG）モデリングと呼ぶことがある．

本研究では，統計モデルとして Lauritzen (1996)の GG モデルを設定する．ただし，この GG モデルのもとで，統計的モデリングを実行する場合，GG モデリングのアルゴリズムを積極的に活用した方がよい．ここで，GG モデルには $2^{p(p-1)/2}$ 通りの膨大なモデル候補が存在することに注意すると, 効率よく最良のモデルの候補となるモデルを設定できるアルゴリズムが必須であるといえる．また，アルゴリズムの多くは，モデル設定と同時に母数の推定も実行する. そこで本研究では, 小標本データへの適用を前提として, モデル設定と母数の推定の両面で優れたアルゴリズムを選定する．

GG モデリングのアルゴリズムには様々な種類が存在するものの，本研究では Friedman et al. (2008)の GLasso を使用する．Glasso では多変量正規分布の尤度関数に L_1 正則化項を加えた次の最適化問題を解く．そして，その最適解を GG モデルのもとでの母精度行列 $\boldsymbol{\Omega}_{\mathrm{GG}}$ の推定量とするアルゴリズムである．

$$\max_{\boldsymbol{\Omega}_{GG}}\left\{\ell\left(\boldsymbol{\Omega}_{GG};\mathbf{S}\right)-\zeta\left\|\boldsymbol{\Omega}_{GG}\right\|_1\right\} \tag{4.10}$$

$$\ell\left(\boldsymbol{\Omega}_{GG};\mathbf{S}\right)\equiv\ln\left|\boldsymbol{\Omega}_{GG}\right|-\mathrm{tr}\left(\mathbf{S}\boldsymbol{\Omega}_{GG}\right) \tag{4.11}$$

ここで，|・|は行列式，tr(・)は行列のトレース，‖・‖$_1$ は行列の L_1 ノルム（行列の全要素の絶対値の和）を示す．また，**S** は標本共分散行列，ζ は正則化項の重み（非負の定数）である．ここで，(4.11)式の多変量正規分布の対数尤度関数は，母平均ベクトル $\boldsymbol{\mu}$ の最尤推定量である標本平均ベクトルを代入した形で定義している．なお，本節でも $\boldsymbol{x}$ の各変数は母平均が 0，母標準偏差が 1 に基準化されているとする．

この Glasso アルゴリズムを使用する利点は次の通りである．第 1 に，モデル設定を効率よく実施できる．Glasso は凸最適化問題である（Banerjee et al. (2008)）ため，全通りを探索するよりも遥に効率的である．また，局所的最適解に陥らず，大域的最適解に到達できる．第 2 に，小標本データを対象とする場合，母精度行列を精度よく推定できる可能性がある．Glasso では母精度行列の非ゼロ要素が縮小推定されるからである．第 3 に，多重共線性の問題を回避できる．異常検知問題を扱う場合，多重共線性の問題が発生することが多いものの，この場合にも Glasso は安定的に母精度行列を推定できる（Ide et al.(2009)）．また，Glasso は正定値に保ちながら精度行列を学習できることも保証されている（Mazumder and Hastie(2012)）．

4.3.3. モデル評価

モデル評価は Akaike(1973)の Akaike Information Criterion（AIC），Schwarz(1978)の Bayesian information criterion（BIC），Efron(1982)の Cross-Validation 等の各種規準および技術的な考察に基づいて総合的に判断する．ただし，本章で想定した統計モデルがすべて共分散構造に対する仮定であることを踏まえると，モデルの解釈容易性を重視することが有用である．すなわち，本質的な共分散構造を的確に把握することで，異常原因の特定に有益な知見の獲得を重視したモデル選択を実施する．そのためには，BIC タイプの情報量規準を利用するのがよいといえる．

しかしながら，BIC はパラメータを最尤推定することを前提とした場合のモデル評価規準であることに注意する必要がある．すなわち，GLasso のような罰則項付き最尤推定の場合，BIC は必ずしもよい評価規準とはならない可能性がある．そこで本研究では，次式で定義される Chen and Chen (2008)の EBIC を使用する．

$$EBIC=-2\ln L_n\left(\hat{\boldsymbol{\theta}}\right)+\nu\ln n+2\omega\ln\tau(\boldsymbol{\Psi})\ \ (0\le\omega\le1) \tag{4.12}$$

ここで，右辺の第 1 項は n 個のデータに対する対数尤度を示す．また，ν は該当モデルのパラメータ数，$\tau(\boldsymbol{\Psi})$は該当のモデルが属するクラスの大きさを示す．なお，ω が 0 のとき，EBIC は通常の BIC に一致する．一方，ω が正の値をとるときは，モデルが属するクラスの大きさに対するペナルティが発生する．

最後に，EBIC の使用方法について補足しておく．Foygel and Drton(2010)は GLasso を使用する場合，$p >> n$ の状況では 0.5 を採用することを推奨している．ただし，$p < n$ の状況における ω の設定方法は示されていない. そこで本章の実験では, 次のように EBIC を使用するものとする．まず ω を 0 から 1.0 まで 0.1 刻みで設定し，各 ω の値に対応した EBIC でモデルを評価する．次に，各 EBIC 規準で最良のモデルの中から，最も異常検知性能が高いモデルを選択する．

4.4. 実データ解析

前節で述べたように提案プロセスに則って解析する場合, 母集団分布に関する事前情報が重要となる．本節では，母集団分布が GG モデルで近似できることが想定される実データを取り上げ，実際に解析を行う．また，MT 法の解析目的の一つである診断の観点から GG モデルを設定した場合の提案プロセスの有用性を評価する．

4.4.1. データセットの概要

本実験では，University of California, Irvine （UCI） 機械学習レポジトリ（http://archive.ics.uci.edu/ml）の Ionosphere データセットを使用する．Ionosphere データセットは，レーダーが受信した電波を電離層から跳ね返ってきたパルスか否か（“Good” か “Bad”）に分類したデータセットである．変数は 34 次元であり，“Good” に属するデータは 225 個，“Bad” に属するデータは 126 個である．

なお，本実験では “Good” を正常ラベル，“Bad” を異常ラベルとする．また，全 34 変数中の 2 変数は全正常データから求めた標準偏差がゼロの値をとるため使用しない．単位空間は全正常データから無作為に N_0 (N_0 = 40, 200)個のサンプルを抽出して定める．テストには全正常データ 225 個および全異常データ 126 個を用いる．

4.4.2. 評価方法

本実験では次項で示す評価対象の判別性能を比較する．ここで，判定性能の評価方法は次の通りである．

判定性能．判定性能の評価指標は，テスト用異常データの正判別率（以降，単に正判別率と呼ぶ）を用いる．正判別率は，0 から 100 までの値をとり，100 に近いほどよい．判別のための閾値は，テスト用正常データの誤判別率（以降，単に誤判別率と呼ぶ）が 1%となるように設定する．ここで，誤判別率によって閾値を設定するのは，誤判別率が異なると手法間の正確な性能の比較ができないからである．誤判別率を等しくすることで，正判別率のみでの評価が可能になる．

4.4.3. 評価対象

本実験では，母共分散行列がフルモデル，PPCA モデル，GG モデル，ナルモデルの場合に対応した解析プロセスを MT 法に導入した場合の判定性能を比較する．ここで，各 MT 法の略称を MT(FULL)，MT(PPCA)，MT(GG)，MT(NULL)とする．ただし，フルモデルとは，母共分散行列に対する仮定がなく，全パラメータを推定する必要があるモデルである．このときの解析プロセスが現行プロセスに対応している．また，ナルモデルとは，母共分散行列の非対角要素がすべてゼロであると仮定したモデルである．このとき，標本マハラノビス距離はユークリッド距離に一致する．なお，MT(PPCA)および MT(GG)の解析プロセスは 4.3 節で示した通りである．

4.4.4. 実験結果

実験結果を図 4.1 に示す．

図 4.1 は各手法の判定性能を単位空間サンプル数の条件ごとに示したグラフである．グラフの縦軸は正判別率である．ここで，各手法の判定性能は単位空間およびテストデータを 100 組用意して解析を行った結果の平均である．なお，各手法において削減されたパラメータ数は，単位空間サンプル数が 200 個の場合，MT(FULL)は 0 個，MT(PPCA)は 0 個，MT(GG)は平均 220 個，MT(NULL)は 496 個である．また，単位空間サンプル数が 40 個の場合，MT(FULL)は 0 個，MT(PPCA)は 0 個，MT(GG)は平均 216 個，MT(NULL)は 496 個のパラメータが削減されている．

図 4.1 より，サンプル数が大きい場合と小さい場合において最も安定的に高い性能を示しているのは MT(GG)であるとわかる．一方，MT(PPCA)は MT(FULL)と同じ性能を示しており，提案プロセスの導入効果は確認できない．実際，MT(PPCA)ではパラメータの削減は行われていない．そのため，MT(PPCA)と MT(FULL)における母共分散行列の推定値は同値となり，異常検知性能も理論的に一致してしまう．

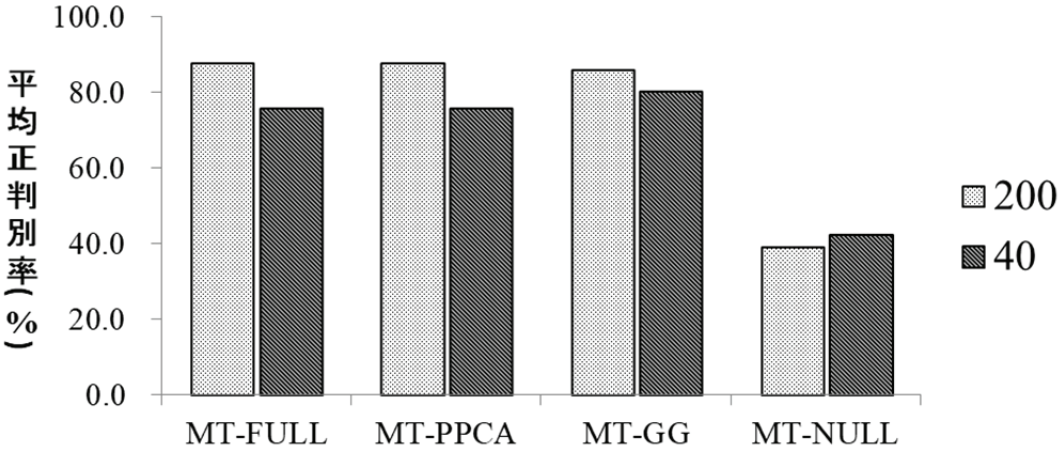
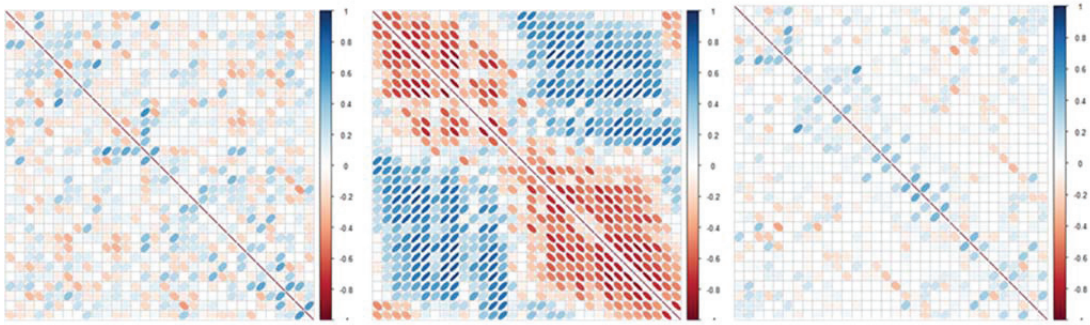


図 4.1　各手法の判定性能（Ionosphere データセット）



(a) 全データ　　(b) フルモデル　　(c) GG モデル

図 4.2　Ionosphere データセットの偏相関係数行列（$N_0 = 40$ の場合）

ここで，MT(GG)の特徴を示すため，図 4.2 にフルモデルと GG モデルに基づく母偏相関係数行列の推定値を比較した．図 4.2a は全正常データから求めた場合，図 4.2b はフルモデルの場合，図 4.2c は GG モデルの場合に対応している．ただし，図 4.2b および 4.2c は，40 個の単位空間サンプルから母偏相関係数行列を推定した結果である．

図 4.2 より，フルモデルよりも GG モデルの方が全正常データから求めた偏相関係数行列に近い相関構造を推定できているとわかる．また，図 4.2c からは厳密にゼロの値をとる非対角要素が数多く存在することが観察できる．このことから GG モデリングと同様の知見を獲得できる．すなわち，偏相関係数行列の非対角要素がゼロとなる変数間には本質的な相関がないと判断できる．この性質を用いれば，異常を検知後，その原因追究を行う際，異常が発生した変数を的確に絞り込めるといえる．

4.5.　モンテカルロ・シミュレーション

前節では，実データ解析を通じて統計モデルの選択が異常検知性能に大きな影響を与

えることを示した．一方，異常原因の追究まで考慮する場合，GG モデリングとの整合性が高い GG モデルを設定することは実用上，有益であるといえる．また実際，統計的モデリングの枠組みでは，必ずしも真の分布を推定することが目的ではないため，GG モデルを設定することは異常原因の追究の観点から妥当であるといえる．しかしながら，この選択は予測の観点から言えば，適切なモデル設定とは限らない．

そこで本研究では，次のようなモンテカルロ・シミュレーションを実施する．まず，母集団分布が GG モデルでない場合の一例として，PPCA モデルを仮定したデータを発生させる．次に，PPCA モデルおよび GG モデルを設定した提案プロセスによる異常検知の性能を予測の観点から評価する．

4.5.1. 生成モデル

本実験では単位空間およびテストデータに対して次の生成モデルを仮定する．

単位空間．データ生成モデルは PPCA モデルとする．観測変数の次元は 30，潜在変数の次元は m (m = 5, 10, 15, 20, 25)とする．ここで，$\boldsymbol{\mu}$ はゼロベクトル，$\boldsymbol{\Gamma}$ は共分散行列の対角要素がすべて 1 となるように各行の 2 乗和が $1-\sigma^2$ となるような乱数で定める．また，σ の値は σ^2 = 0.2, 0.4, 0.6, 0.8 を満たすように設定する．サンプル数は 60 個である．

テストデータ．正常データの生成モデルは単位空間と同様である．一方，異常データの生成モデルは正常データと同じ母平均ベクトルと，その母共分散行列を定数倍（1.5^2倍）したパラメータをもつ多変量正規分布とする．正常データおよび異常データのサンプル数はともに 100,000 個である．

4.5.2. 評価方法

本実験では次節で示す評価対象の判別性能を比較する．ここで，判定性能の評価方法は次の通りである．

判定性能．判定性能の評価指標はテスト用異常データの正判別率を用いる．ここで，判別のための閾値は，テスト用正常データの誤判別率が 1%となるように設定する．

4.5.3. 評価対象

本実験では，母共分散行列がフルモデル，PPCA モデル，GG モデル，ナルモデルの場合に対応した解析プロセスを MT 法に導入した場合の判定性能を比較する．ここで，各 MT 法の略称は前節と同様である．

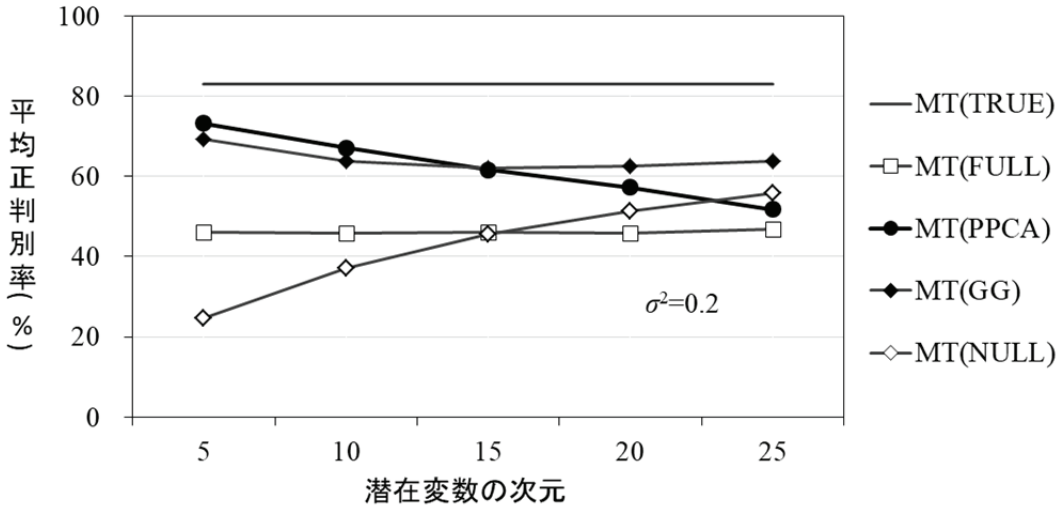


図 4.3　各手法の性能比較（$\sigma^2 = 0.2$ の場合）

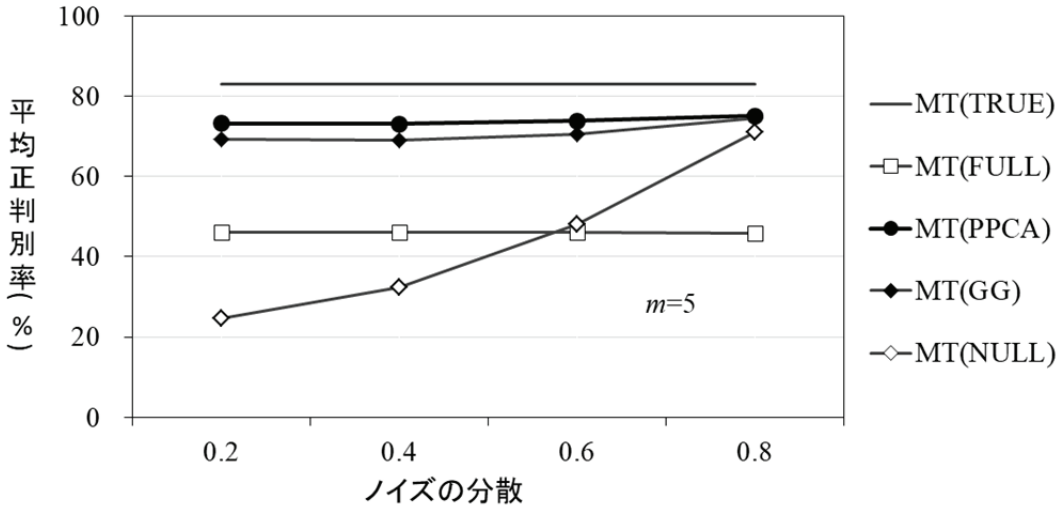
MT(PPCA)および MT(GG)の設定について説明しておく．まず，MT(PPCA)は真のモデルが既知であるため, 真の潜在変数の次元 m を所与として母共分散行列を推定する．一方，MT(GG)はモデル評価以外のプロセスは 4.3 節で示した通りとする．ただし，モデル評価には，真の分布と推定した統計モデルとの Kullback-Leibler（KL）ダイバージェンス（Kullback and Leibler (1951)）を用いる．そして，KL ダイバージェンスが最小となるモデルを選択するものとする．

4.5.4.　実験結果

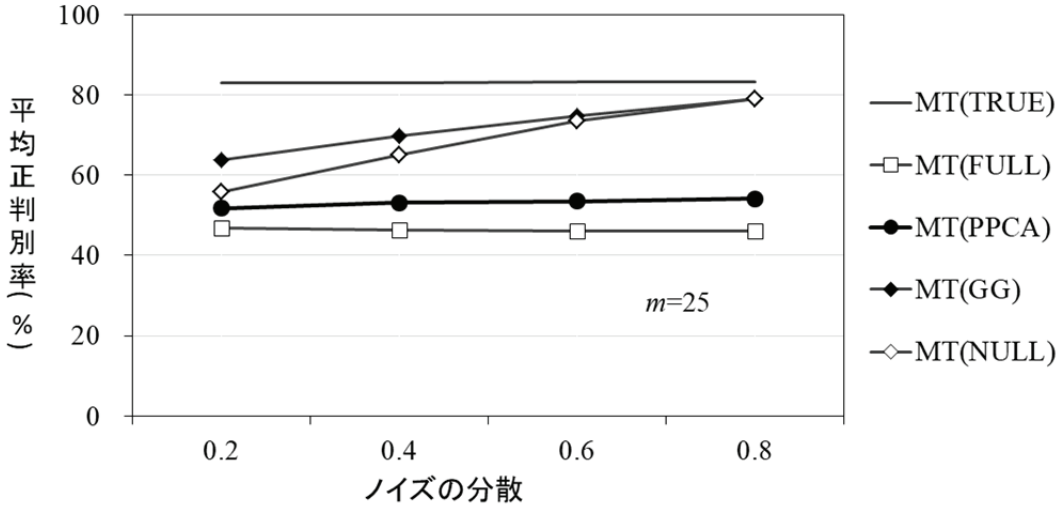
実験結果を図 4.3 および図 4.4 に示す．

図 4.3 は $\sigma^2 = 0.2$ の場合における各手法の判定性能と潜在変数の次元の関係を示したグラフである．図中には前述の評価対象に加えて，母マハラノビス距離を用いた MT 法（MT(TRUE)）の解析結果も比較のため示している．ここで，グラフの縦軸は正判別率であり，横軸は潜在変数の次元 m の大きさである．なお，各手法の判定性能は単位空間およびテストデータを 100 組用意して解析を行った結果の平均である．

図 4.3 より，MT(PPCA)は潜在変数の次元が判定性能に大きな影響を与えているとわかる．$m = 5$ の場合，評価対象の中で最も高い判定性能を示しているものの，$m = 25$ の場合には MT(FULL)とほぼ同等の正判別率にまで低下している．一方，MT(GG)は潜在変数の次元によらず，安定的に正判別率を向上させていることが観察できる．



(a) $m = 5$ の場合



(b) $m = 25$ の場合

図 4.4　ノイズによる各手法の判定性能の変化

また図 4.4 は各手法の判定性能とノイズの分散 σ^2 の関係を示したグラフである．図 4.4a は潜在変数の次元 $m = 5$ の場合，図 4.4b は $m = 25$ の場合に対応している．各図中には前述の評価対象に加えて，母マハラノビス距離を用いた MT 法（MT(TRUE)）の解析結果も比較のため示している．ここで，グラフの縦軸は正判別率であり，横軸はノイズの分散 σ^2 の大きさである．なお，各手法の判定性能は単位空間およびテストデータを 100 組用意して解析を行った結果の平均である．

図 4.4 より，MT(PPCA)はノイズの分散 σ^2 の大きさにかかわらず，ほぼ一定の正判別率を示しているとわかる．一方，MT(GG)は潜在変数の次元によらず，ノイズの分散 σ^2 の大きさが増加するほど，正判別率が向上することが観察できる．

ここで，PPCA モデルのもとでは，ノイズ $\boldsymbol{\varepsilon}$ に関して次の 2 つの性質があることに注意する．第 1 にノイズ $\boldsymbol{\varepsilon}$ が相関のない多変量正規分布に従うことである．第 2 に，ノイズの分散 σ^2 が大きいほどノイズ $\boldsymbol{\varepsilon}$ の共分散構造が最終的な共分散構造に与える影響が大きくなることである．この 2 つの性質からノイズの分散 σ^2 が大きくなるほど，PPCA モデルの母共分散行列，延いては母相関係数行列の非対角要素はゼロの値に近づくといえる．このとき，対応する母精度行列の非対角要素もゼロの値に近づくため，σ^2 が大きくなるほど GG モデルの仮定に近づくといえる．

したがって，MT(GG)は PPCA モデルのもとでもノイズの影響が大きくなるほど，精度よく異常検知できる可能性がある．一方，MT(PPCA)は本実験のように潜在変数の次元を真の次元に固定する場合，このノイズの影響を考慮したモデル選択ができない．言い換えれば，ノイズの影響を考慮するためには，予測の観点に基づいてナルモデルを含む他のモデルを選択した方がよいといえる．

このように真のモデルを選択することが最もよいモデル選択の方法であるとは限らない．そのため，解析の目的，推定すべきパラメータ数やサンプル数を考慮しながら，適切なモデルを選択する必要がある．提案プロセスを導入すれば，統計的モデリングの枠組みに基づいて自然にモデル選択が実行可能であり，小標本データ解析に関わる実務的な需要にも適切かつ柔軟に対応できるといえる．

4.6. 本章のまとめ

本章では，MT 法を小標本データに適用する場合を想定したうえで，統計的モデリングの具体的な実施方法について考察した．一般に小標本データを解析する場合，学習データへの過剰適合が発生してしまう．そのため，推定すべきパラメータ数の削減が有用となるものの，これまでの対策では未知パターンの異常が検知できない危険性があった．そこで，本研究では未知パターンの異常検知を考慮した小標本データの解析プロセスを提案した．提案プロセスでは，統計的モデリングの枠組みに基づいて自然に推定すべきパラメータ数を削減できる．提案プロセスの導入効果を数値実験で検証した結果，予測と診断の両面で有用との結論を得た．特にガウシアン・グラフィカル・モデルの設定は異常原因追及の観点から実用性が高いといえる．

第 5 章　高次元データ解析プロセスの提案

本章では，MT システムを高次元データに適用する場合の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する．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 が遥に多いデータ（$p >> n$）を取り扱う必要がある．しかしながら，そのような場合には，そもそも現行プロセスでは標本マハラノビス距離が計算不能となるため，異常検知を実行できない．そこで本研究では，$p >> n$ でも計算可能な 2 種類の距離を取り上げ，その距離に基づく異常検知方式の特徴について考察する．そして，そのパラメータを統計的モデリングの枠組みに基づいて推定することを提案する．また数値実験を通して，提案プロセスが予測と診断の両面で有用であることを示す．

本章の内容は大久保・永田 (2012)および Ohkubo and Nagata (2017a)に基づく．

5.1.　はじめに

近年，我が国の製造業では，大量のセンサーから取得・蓄積した保全データに基づく設備機器の異常検知に関する技術に注目が集まっている．また，医療分野においてもマイクロアレイ・データ等の極めて高次元なデータを扱う必要性が生じている．しかしながら，このような高次元データを対象とする場合，現行 MT 法の解析プロセスでは適切な分析ができるとは限らない．

第 2 章で述べたように，MT 法はマハラノビス距離に基づく異常検知方式である．この異常検知方式の問題点の一つは，標本共分散行列の逆行列計算を必要とするため，観測変数の次元 p がサンプル数 n よりも大きな（$p > n$）場合，マハラノビス距離が計算できないことである．そのため，本研究で想定するような高次元小標本（$p >> n$）データには MT 法は適用できない．

このような高次元小標本データ解析に関わる問題を解決するためのアプローチとして，次の 2 つを考えることができる．

第 1 のアプローチは，MT 法における現行プロセスの改良である．前章で提案した小標本データのための解析プロセスは，その一例となる．提案プロセスを導入すれば，$p > n$ の場合あるいは $p >> n$ の場合にも標本マハラノビス距離が計算可能となり，異常検知を実行できる．ここで，前章の提案プロセスにおいて Tipping and Bishop (1999)の PPCA モデルを設定する場合を考える．前章では，PPCA モデルを多変量正規分布における母共分散行列に対する仮定として取り扱った．一方，本章では，PPCA モデルを固有値分布への仮定として捉え直したうえで，$p >> n$ の場合の対策として再考する．

第 2 のアプローチは，高次元データを対象とした新たな MT システムの提案である．例えば，田口 (2006)の RT 法とその改良手法である大久保・永田 (2012)の RT-PC 法が提案されている．RT 法は主に高次元バイナリ･データを対象とした方法論であるため，適用対象は限定的である（永田・土居 (2009)，大久保・永田 (2012)）．一方，RT-PC 法は高次元連続量データにも適用可能なように RT 法を改良している．本章では，高次元データがもつ幾何学的な特徴，すなわち球面集中現象（例えば，Yata and Aoshima (2012) に詳しい）に着目したうえで，その特徴を活用した異常検知方式として RT-PC 法について再考する．球面集中現象を積極的に利用することで，$p >> n$ となるデータが解析可能となるだけでなく，高次元であるほど異常検知性能の向上が期待できる．

しかしながら，このようなアプローチを考える際，パラメータの推定方法についても慎重に検討する必要があることを注意しておく．前章あるいは大久保・永田 (2012)において提案されたプロセスは高次元データを対象とする場合，パラメータを精度よく推定できない可能性がある．また，現行の推定法では，データが高次元になるほど，パラメータの解釈容易性が失われてしまう．例えば，品質管理の分野では，異常の検知だけでなく，異常原因の特定に重要な知見の獲得が分析の目的となる場合がある．そのため，推定すべきパラメータ数を削減し，解釈容易性を高めることも必要である．

そこで本研究では，高次元データ，特に高次元小標本データを対象とした MT システムの新たな解析プロセスを提案する．具体的には，まず $p >> n$ となる場合でも計算可能な 2 種類の距離を取り上げ，その距離に基づく異常検知方式がもつ特徴について考察する．また，その算出に必要なパラメータが母共分散行列の固有値および固有ベクトルであることを示す．その後，そのパラメータを統計的モデリングの枠組みに基づいて推定する．提案プロセスでは，統計モデルの母数の推定に際して，Sparse Principal Component Analysis（SPCA）に基づく推定法を使用する．SPCA を用いることで，高次元データを対象とした場合でも固有値および固有ベクトルが精度よく推定されることが期待できる．また，推定された固有ベクトルの非ゼロの要素数が減少するため，主成分の解釈，延いては異常の原因が技術的に考察し易くなる．

5.2. 現行プロセスの問題点

第 2 章で示した MT 法は変数の次元 p がサンプル数 n よりも大きな（$p > n$）データを解析対象とする場合，標本マハラノビス距離が計算不能となってしまう．現行プロセスでは，標本マハラノビス距離を算出する際，母相関係数行列の推定量 $\hat{\mathbf{\Pi}}$ として標本相関係数行列を用いている．そのため，$p > n$ となる場合，標本相関係数行列の逆行列が

存在しなくなり，標本マハラノビス距離も計算できなくなる．本節では， $p > n$ となるデータを対象とする場合でも計算可能な距離として，前章および大久保・永田 (2012) において提案された 2 つの距離を取り上げる．また，その距離に基づく異常検知方式の特徴について考察する．さらに，現行プロセスにおける問題点を指摘する．

5.2.1. PPCA モデルに基づくマハラノビス距離

本項では，$p > n$ の場合にも計算可能な距離の一つとして，前章の提案プロセスにおいて PPCA モデルを設定した場合の母マハラノビス距離を取り上げる．また，その距離がもつ特徴をホテリング多変量管理図の観点から考察する．

まずホテリング多変量管理図では，母相関係数行列のスペクトル分解を，母相関係数行列 $\mathbf{\Pi}$ の固有値 $\lambda_1, \lambda_2, \cdots, \lambda_p$ および対応する長さ 1 の固有ベクトル $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_p$ を用いて

$$\mathbf{\Pi} = \sum_{j=1}^{q} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^{T} + \sum_{k=q+1}^{p} \lambda_k \boldsymbol{\xi}_k \boldsymbol{\xi}_k^{T} \tag{5.1}$$

と表現して, データ空間全体を第 1 項と第 2 項に対応した固有ベクトルが張る空間に分けて異常原因を考察する. いま第 1 項に対応した固有ベクトルが張る空間を主部分空間, 第 2 項に対応した固有ベクトルが張る空間をノイズ空間と呼ぼう．このとき，主部分空間に発生した異常は高々 q 個の主成分の変動が原因であると考える．一方，ノイズ空間に発生した異常は本質的な相関構造の崩れが原因だと考える．そして，各空間の異常を T^2 統計量と Q 統計量を独立に利用して異常検知を行う（Jackson and Mudholkar (1979)）．

次に，PPCA モデルでは，p 次元観測変数が q 次元潜在変数と p 次元ノイズから生成されると仮定する．このとき，母相関係数行列は

$$\mathbf{\Pi}_{\mathbf{ppca}} = \sum_{j=1}^{q} \left(\lambda_j - \sigma^2\right) \boldsymbol{\xi}_j \boldsymbol{\xi}_j^{T} + \sum_{k=1}^{p} \sigma^2 \boldsymbol{\xi}_k \boldsymbol{\xi}_k^{T} = \sum_{j=1}^{q} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^{T} + \sum_{k=q+1}^{p} \sigma^2 \boldsymbol{\xi}_k \boldsymbol{\xi}_k^{T} \tag{5.2}$$

と表現される．すなわち，(5.1)式で第 q+1 固有値以降の固有値がすべて σ^2 とした特別な場合に相当する．そして，このモデルのもとでの母マハラノビスに基づいて異常検知を実行することは, 主部分空間とノイズ空間に発生した異常を 1 つの統計量に基づいて検知することに対応する．ここで，母マハラノビス距離が(4.1)式で示したように，主成分得点を対応する固有値の逆数で重みづけた和と解釈できることに注意する．そして，PPCA モデルにおける母マハラノビス距離は

$$\Delta_{ppca}^2(\boldsymbol{u}) = \sum_{j=1}^{q} \frac{\left(\boldsymbol{\xi}_j^{T} \boldsymbol{u}\right)^2}{\lambda_j} + \sum_{k=q+1}^{p} \frac{\left(\boldsymbol{\xi}_k^{T} \boldsymbol{u}\right)^2}{\sigma^2} = \sum_{j=1}^{q} \frac{\left(\boldsymbol{\xi}_j^{T} \boldsymbol{u}_i\right)^2}{\lambda_j} + \frac{SE_{(q)}}{\sigma^2} \tag{5.3}$$

$$SE_{(q)} = \boldsymbol{u}^T\boldsymbol{u} - \sum_{j=1}^{q}\left(\boldsymbol{\xi}_j{}^T\boldsymbol{u}\right)^2 \tag{5.4}$$

$$\sigma^2 = \frac{p - \sum_{j=1}^{q}\lambda_j}{p-q} \tag{5.5}$$

で定義される．このとき，(5.3)式は $q = p-1$ とすれば，(4.1)式の母マハラノビス距離となることに注意する．すなわち，PPCA モデルにおける母マハラノビス距離は，母集団分布への仮定がない場合の母マハラノビス距離の自然な近似となっている．また，(5.3)式の右辺第 1 項は T^2 統計量，(5.4)式の統計量は Q 統計量に一致する．そのため，PPCA モデルにおける母マハラノビス距離は T^2 統計量と Q 統計量を統合した尺度に基づく異常検知方式の一つといえる．そして，PPCA モデルにおける母マハラノビス距離は高々 q 個の固有値とそれに対応する固有ベクトルを推定すれば，標本空間上の対応するマハラノビス距離が計算できる．そのため，$p > n$ の場合にも計算可能であり，異常検知方式を実行できるとわかる．

なお, 前章での提案プロセスは母相関係数行列の固有値および固有ベクトルの推定量として，標本相関係数行列から出発した PCA に基づく推定量を使用している．これは前章で述べたように PPCA モデルのもとで母数の最尤推定を行った結果が，PCA に基づく固有値および固有ベクトルによって表現できることを利用している．

5.2.2. RT-PC 法の距離

前項で示した距離は, 母集団分布への仮定がない場合の母マハラノビス距離の自然な近似となるため，その距離に基づく異常検知方式は同様の性質を有するといえる．例えば，正規母集団を仮定する場合，いずれも分布の裾に発生した異常を検知するための異常検知方式となる．ところで，$p >> n$ となる高次元小標本データに対して，(5.2)式のような母相関係数行列あるいは固有値の分布を考える場合，ノイズ空間に発生するデータは球面集中現象（Yata and Aoshima (2012)）と呼ばれる幾何学的な特徴を有することが知られている．すなわち，データは原点からの距離が等しい球面上に集中し，分布の中心付近にはほとんど発生しなくなる．

このような高次元データの特徴を異常検知に利用するための距離として，田口 (2006)の RT 法およびその改良手法である大久保・永田 (2012)の RT-PC 法で用いる距離の利用を考えることができる．本項では，RT-PC 法で使用される距離（以降，RD と呼ぶ）とその距離に基づく異常検知方式がもつ性質について説明する．ただし，RT-PC 法の解析プロセスは，用いる距離の違いを除いて MT 法と同様である．

まず母集団上の RD（以降，母 RD と呼ぶ）を定義するためには，p 次元変数 $\boldsymbol{x}$ を 2 次元変数 $\boldsymbol{z}$ に集約する必要がある．ここで，p 次元変数 $\boldsymbol{x}$ に対して，$\boldsymbol{z}=\left(Z_1, Z_2\right)^T$ を

$$Z_1=\boldsymbol{\xi}_1^{\ T}\boldsymbol{u}/\boldsymbol{\xi}_1^{\ T}\boldsymbol{\xi}_1 \tag{5.6}$$

$$Z_2=\sqrt{SE/(p-1)} \tag{5.7}$$

$$SE=\boldsymbol{u}^T\boldsymbol{u}-\left(\boldsymbol{\xi}_1^{\ T}\boldsymbol{u}\right)^2/\boldsymbol{\xi}_1^{\ T}\boldsymbol{\xi}_1 \tag{5.8}$$

と定義する．ただし，(5.8)式は $q=1$ とした場合の(5.4)式の統計量に相当する．

次に，2 次元変数 $\boldsymbol{z}$ に対する母マハラノビス距離

$$\Delta^2(\boldsymbol{z})=\left(\boldsymbol{z}-\boldsymbol{\mu}_z\right)^T\boldsymbol{\Sigma}_z^{-1}\left(\boldsymbol{z}-\boldsymbol{\mu}_z\right) \tag{5.9}$$

を母 RD として定義する．ただし，2 次元変数 $\boldsymbol{z}$ の期待値を $\mathrm{E}(\boldsymbol{z})=\boldsymbol{\mu}_z$，分散を $\mathrm{V}(\boldsymbol{z})=\boldsymbol{\Sigma}_z$ としている．ここで，第 1 固有値とそれに対応する固有ベクトルを推定すれば，標本空間上の RD（以降，標本 RD と呼ぶ）が計算できることに注意する．そのため，$p>n$ の場合にも計算可能であり，異常検知方式を実行できることがわかる．

なお，大久保・永田 (2012)では，固有値および固有ベクトルの推定量として，標本相関係数行列から出発した PCA に基づく推定量を使用している．また，2 次元変数 $\boldsymbol{z}$ の期待値および分散の推定量として，母集団からの大きさ n の無作為標本 $\boldsymbol{x}_i$ (i = 1,2, ..., n)に対する $\boldsymbol{z}_i=\left(Z_{i1}, Z_{i2}\right)^T$ (i = 1,2, ..., n)から求めた標本平均ベクトルおよび標本共分散行列を使用している．

最後に，標本 RD と PPCA モデルにおける標本マハラノビス距離に基づく異常検知方式は本質的に異なる異常検知方式として捉える必要があることを注意しておく．以降，PPCA モデルにおける母マハラノビス距離において q = 1 とした場合の距離を単に母 MD と呼ぶ．また，標本空間上の対応する距離を標本 MD と呼ぶことにする．

標本 MD と標本 RD は同様の統計量（第 1 主成分とその残差）に関する距離であるものの，各距離に基づく異常検知方式は検出する異常の定義が異なる．標本 MD に基づく方式では，単位空間の中心に発生したデータほど，単位空間に属する（正常である）度合いが高いと判断される．一方，標本 RD に基づく方式では，単位空間の中心付近に発生したデータも異常と判定される可能性がある（永田・土居 (2009)）．

このことは特に異常原因について考察する必要がある場合，大きな影響があるといえる．標本 RD に基づく方式は，第 2 固有ベクトル以降が張るノイズ空間の中心付近に発

生したデータも異常と捉えるため，第 2 主成分以降が技術的な意味をもつ場合，定性的な意味合いが不明になる．一方，第 2 主成分以降がノイズによる変動を意味すれば，データの本質的な構造に変化があるものと判断できる．よって，第 2 主成分以降が技術的な意味をもつか否かを事前に確認したうえで，両者を使い分ける必要がある．

5.2.3. 従来型 PCA に基づく推定法の問題

5.2.1 項と 5.2.2 項では, $p > n$ となるデータを対象とする場合でも計算可能な距離として，標本 MD および標本 RD があることを述べた．また，各距離を計算する際は，母相関係数行列の第 1 固有値および固有ベクトルを推定すればよいことも述べた．ここで，現行プロセスでは，各距離の算出時に従来型 PCA に基づく推定量を用いていることに注意する必要がある．本節では，高次元データを適用対象とする場合，従来型 PCA に基づく推定法が推定精度および解釈容易性の両面から問題があることを指摘する．

5.1 節で述べたように，MT システムに属する手法は，異常を検知するだけでなく，異常原因の特定等に有用な知見を獲得することも解析の目的となる場合がある．したがって，固有値および固有ベクトルの推定精度を向上させることは，主成分の解釈を行ううえで重要であり，異常の発生原因を追究するうえでも有益であるといえる．また，5.5 節で示すように，異常検知性能にも大きな影響を与える．

しかしながら，従来型 PCA はこれまでの多変量解析法の枠組み，すなわち大標本漸近理論のもとで，その精度保証が与えられていることに注意する．このことはサンプル数 n を固定したもとで高次元のデータを扱う問題設定においては，必ずしもよい方法論とは限らないことを意味する．実際，変数の次元 p が $p \to \infty$ となる前提のもとで，従来型 PCA に基づく母相関係数行列の固有値および固有ベクトルの推定量が一致性を満たす領域は限定的であることが報告されている（Jung and Marron (2009)，Yata and Aoshima (2009) ，Jung et al. (2012)）．すなわち，高次元小標本データを適用対象とする場合，従来型 PCA では推定量の精度保証を与えることが困難となる．

また知見獲得の観点からは，固有ベクトルの解釈容易性も重要である．高次元になるほど固有ベクトルの要素数は増加するため，高次元データを対象とする場合，固有ベクトルの解釈は困難になるといえる．すなわち，もとの変数が主成分に与える影響を，固有ベクトルの要素をもとに判断する場合，技術的な考察は高次元になるほど難しくなる．これはもとの変数が主成分に与える影響が全くない（固有ベクトルの対応する要素がゼロである）場合でも，従来型 PCA で推定すると非ゼロの値をとることに起因する．

5.3. 提案プロセス

本研究では，高次元データ，特に高次元小標本データを対象とした MT システムの新たな解析プロセスを提案する．提案プロセスでは，5.1 節で示したような統計的モデリングを通して統計モデルを推定後, そのパラメータ推定量を用いて標本 MD および標本 RD を計算する．本節では，統計的モデリングにおける各ステップの詳細を示す.

5.3.1. モデル設定と母数の推定

いま単位空間の真の分布が多変量正規分布であることが想定できるものとする. 加えて，p 次元の観測変数 $\boldsymbol{x}$ が 1 次元の変数で説明可能との事前情報があるとする．本節では，このような事前情報がある場合のモデルの設定と母数の推定について考える.

まず，統計モデルとして Single component Spike covariance Gaussian（SSG）モデル（例えば，Shen and Huang (2008)）を設定する．SSG モデルとは，高次元データの特徴を示すモデルの一つである．SSG モデルでは，母共分散行列の固有値分布に対して

$$\lambda_1 = p^{\alpha}, \lambda_2 = \lambda_3 = \cdots = \lambda_p = {\sigma_{SSG}}^2 \tag{5.10}$$

を仮定する．ここで，α は非負の定数であり，σ_{SSG}^2 には通常 1 が仮定される（例えば，Shen and Huang (2008)）. このモデルは第 1 固有値が変数の次元の増加に伴って急激に増加する一方, 第 2 固有値以降は変数の次元によらず一定と仮定するモデルであるといえる．なお，SSG モデルでは通常，固有ベクトルに関する仮定は存在しない．ただし，本研究では，第 1 固有ベクトルの大部分がゼロの値をとること，すなわちスパース性を仮定する．また，特に断りのない限り，通常の SSG モデルと区別なく，第 1 固有ベクトルに対してスパース性を仮定したモデルも SSG モデルと呼ぶことにする.

この SSG モデルのもとで，統計的モデリングを実行する場合，SPCA のアルゴリズムを積極的に活用した方がよい．前章の GG モデルにおけるパラメータ推定に GG モデリングのアルゴリズムを利用したのと同様に, SSG モデルのもとでもモデル設定とパラメータ推定を同時に実行することを考える．SPCA アルゴリズムを利用することで，最良のモデルの候補となるモデルを効率よく設定できる．また，非ゼロ要素に対する縮小推定も行うため，小標本における推定精度の向上も期待できる.

SPCA には様々なアルゴリズムが存在するものの，本研究では MD と RD の算出に必要な第 1 固有値と対応する固有ベクトルを推定することに焦点を絞り，Shen and Huang (2008)の Regularized SPCA（RSPCA）を用いる．RSPCA では, 高次元漸近理論のもとで, 従来型 PCA よりも緩い条件の下で固有ベクトルの推定量の一致性が保証されており，

従来型 PCA を改良する方法論となっている．RSPCA では，最適化問題

$$\min_{\boldsymbol{f},\boldsymbol{l}} \left\{ \left\| \mathbf{X} - \boldsymbol{f}\boldsymbol{l}^{T} \right\|_{F}^{2} + \zeta P(\boldsymbol{l}) \right\} \tag{5.11}$$

を解くことで，固有ベクトルの推定量を得る．ここで，$\mathbf{X}$ は $n \times p$ のデータ行列，$\boldsymbol{f}$ は分散 1 に基準化した第 1 主成分得点ベクトル（n 次元ベクトル），$\boldsymbol{l}$ は第 1 固有ベクトル（p 次元ベクトル）である．また，$\|\cdot\|_F$ はフロベニウス・ノルム，$P(\boldsymbol{l})$は Least Absolute Shrinkage and Selection Operator（Lasso）あるいは L_1 ノルム，adaptive lasso, Smoothly Clipped Absolute Deviation（SCAD）等の正則化項，ζは正則化項の重み（非負定数）である．このとき，ζをゼロにすれば，通常の PCA となることに注意しておく．一方，第 1 固有値は(5.11)式の最適解 $\boldsymbol{f}^*$および $\boldsymbol{l}^*$を用いて次のように推定できる．

$$\lambda_{1(RSPCA)} = \left(\boldsymbol{f}^{*T} \mathbf{X} \boldsymbol{l}^{*} / \sqrt{n-1} \right)^2 \tag{5.12}$$

なお，正則化項に SCAD を使用する場合，(5.11)式は非凸最適化問題となることに注意する必要がある．したがって，本研究では，SCAD ペナルティの局所 2 次近似を利用した近似解を母共分散行列の固有ベクトルの推定量として使用する．

5.3.2. モデル評価

モデル評価は AIC，BIC，Cross-Validation 等の各種規準および技術的な考察に基づいて総合的に判断する．ただし，前章と同様の理由により，異常原因の特定に有益な知見の獲得を重視したモデル選択を実行するため，BIC タイプの情報量規準を利用するのがよいといえる．ここで，RSPCA アルゴリズムを利用する場合，Shen and Huang (2008) によって提案されている BIC が使用できる．なお，RSPCA アルゴリズムとモデル評価規準の関係については，Sill et al. (2015)に詳しい．

5.4. 実データ解析

本節では，統計的モデリングの枠組みに基づく新たな高次元データ解析プロセスの有用性を評価するために実データ解析を行う．使用するデータセットは Gordon et al. (2002)のマイクロアレイ・データである．なお，公開されている高次元小標本データセットの多くは，評価用のサンプル数が十分ではないため，将来のデータに対する異常検知性能を適切に評価することが困難である．そのため，本実験では，主に固有ベクトルの解釈容易性の観点から，提案プロセスの有用性を評価する．

5.4.1. データセットの概要

Gordon et al. (2002)で公開されたマイクロアレイ・データは，各サンプルを悪性胸膜中皮腫（MPM; malignant pleural mesothelioma）か肺腺癌（ADCA; adenocarcinoma）かに分類したデータセットである．変数は 12,533 次元であり，MPM に属するデータは 31 個，ADCA に属するデータは 150 個である．

なお，本実験では ADCA に属するデータを正常データ，MPM に属するデータを異常データとする．また，単位空間は正常データ 150 個の中から，無作為に 75 個のサンプルを抽出して定める．テスト用の正常データは残りの正常データ 75 個であり，テスト用の異常データは 31 個すべてを用いる．

5.4.2. 評価方法

本実験では次節で示す評価対象の判別性能および第 1 固有ベクトルの解釈容易性を比較する．ここで判定性能および解釈容易性の評価方法は次の通りである．

判定性能．評価方法は，テスト用異常データの正判別率とする．ここで，判別のための閾値は，テスト用正常データの誤判別率が 5%となるように設定する．

解釈容易性．評価方法は，第 1 固有ベクトルの全要素数 p に対するゼロの値をとる要素数の割合である（以降，この評価指標を zero elements rate と呼ぶ）．zero elements rate (%) は 0 から 100 までの値をとり，100 に近いほどよいと考える．ただし，実際には非ゼロ要素とゼロ要素が正しく選択されているか技術的に考察する必要がある．

5.4.3. 評価対象

本実験では，MD および RD に基づく異常検知方式を評価対象とする．パラメータ推定法は PCA および SPCA を使用する．また，その推定法に基づく異常検知方式の略称を MTS(PCA)，MTS(SPCA)とする．このとき，MTS(SPCA)が提案プロセスに対応しており, 解析プロセスは 5.3 節で示した通りである. ただし, RSPCA の正則化項は adaptive lasso を用いる（詳細は Sill et al. (2015)を参照されたい）．

5.4.4. 実験結果

実験結果を表 5.1 に示す.

表 5.1 は PCA および SPCA を各距離に適用した場合の判別性能および第 1 固有ベクトルの解釈容易性を比較した表である．なお，正判別率と zero elements rate は単位空間

表 5.1　各手法の判定性能および固有ベクトルの解釈容易性

p	n of unit space	method	distance	positive discrimination rate (%)	zero elements ratio (%)
12,533	75	MTS(SPCA)	MD	88.4	69.1
12,533	75	MTS(PCA)	MD	88.2	0.0
12,533	75	MTS(SPCA)	RD	88.6	69.1
12,533	75	MTS(PCA)	RD	88.3	0.0

およびテストデータを 100 組用意して解析を行った結果の平均である．

表 5.1 より，正判別率は同等であるものの，用いる推定法によって zero elements rate には大きな差異があるとわかる．PCA は固有ベクトルの要素の中にゼロの値をとる要素が全く存在していない．一方，SPCA の zero elements rate は約 70%であるため，全体の約 30%のみが非ゼロの値をとっている．したがって，SPCA を用いることで，正判別率を維持したまま，解釈容易性を向上できているといえる．

5.5.　モンテカルロ・シミュレーション

前節では解釈容易性の観点から提案プロセスの有用性を示したものの, 異常検知性能については同等であった．ここで，SPCA の特別な場合として PCA が定義できることに注意すると, あるデータセットに対して異常検知性能を最大化するようにパラメータを決定すれば，SPCA を利用した MT システムは PCA を利用した場合と比べて同等かそれ以上の性能が実現できる．しかしながら，MT システムでは未知のパターンが発生した場合でも，その異常を検知することが要求される．そのため，異常の発生パターンを様々に用意して，その性能を評価する必要性が生じる．そこで本章では，モンテカルロ・シミュレーションを通して，提案プロセスを予測力の観点から評価する．

5.5.1.　生成モデル

本実験では, 正常データおよび異常データの生成モデルに多変量正規分布を仮定したうえで，単位空間およびテストデータを次のような設定で発生させる．

単位空間．データの生成モデルを p_0 (p_0 = 1,000)次元ベクトル $\boldsymbol{\mu}_0$ と $p_0 \times p_0$ の正定値対称行列 $\boldsymbol{\Sigma}_0$ をパラメータにもつ多変量正規分布 $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$として，$N_0$ (N_0 = 30) 個のサンプルを発生させる．$\boldsymbol{\mu}_0$ はゼロベクトル, $\boldsymbol{\Sigma}_0$ は固有値分布に SSG モデルを仮定する．ただし，SSG モデルのパラメータには $p = p_0$, α = 0.1, 0.5, 0.9，$\sigma_{SSG}^2 = 1$ を設定する．また，第 1

固有ベクトルは$\lfloor p_0^{\beta} \rfloor$個の要素が非ゼロの値$\lfloor p_0^{\beta} \rfloor^{-\frac{1}{2}}$をとるものとする（$\beta = 0.1, 0.3, 0.5, 0.7, 0.9$）．ただし，$\lfloor \cdot \rfloor$はガウス記号である．

テストデータ．正常データの生成モデルは，単位空間と同一のパラメータをもつ多変量正規分布 $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$とする．異常データの生成モデルは，$p_0$ 次元ベクトル $\boldsymbol{\mu}_0$ と $p_0 \times p_0$ の正定値対称行列 $\boldsymbol{\Sigma}_0$ を定数 c^2 倍$(c^2 = 1.1, 0.75)$した行列をパラメータにもつ多変量正規分布 $N(\boldsymbol{\mu}_0, c^2\boldsymbol{\Sigma}_0)$とする．なお，正常データおよび異常データともに 1,000 個のサンプルを発生させる．

5.5.2. 評価方法

本実験では次項で示す評価対象の判別性能を比較する．ここで，判定性能の評価方法は次の通りである．

判定性能．評価方法はテスト用異常データの正判別率である．ここで，判別のための閾値は，テスト用正常データの誤判別率が 5%となるように設定する．

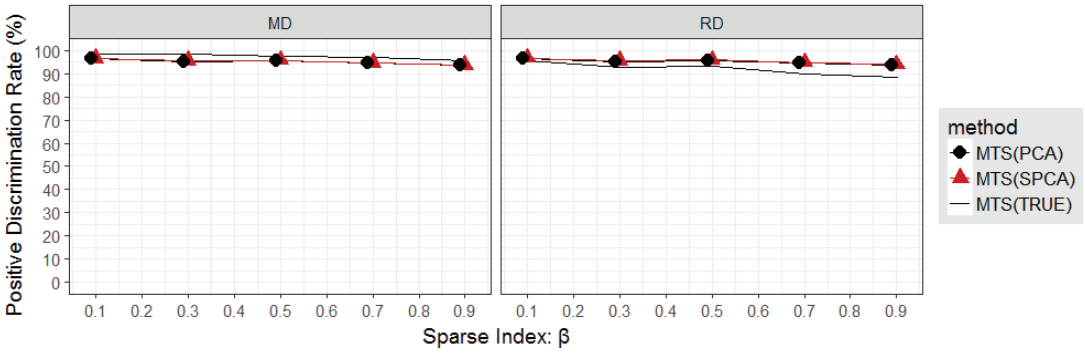
5.5.3. 評価対象

本実験では，MD および RD に基づく異常検知方式を評価対象とする．パラメータ推定法は PCA および SPCA を使用する．また，その推定法に基づく異常検知方式の略称を MTS(PCA)，MTS(SPCA)とする．このとき，MTS(SPCA)が提案プロセスに対応しており, 解析プロセスは 5.3 節で示した通りである. ただし, RSPCA の正則化項は adaptive lasso を用いる（詳細は Sill et al. (2015)を参照されたい）．

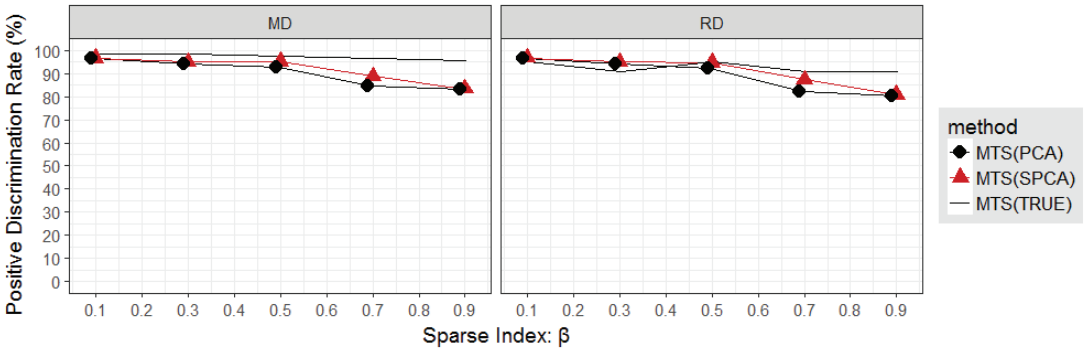
5.5.4. 実験結果

実験結果を図 5.1 から図 5.2 に示す．

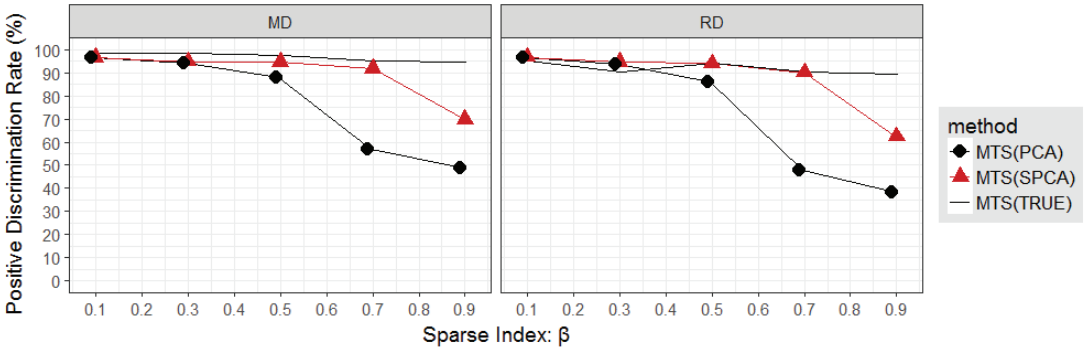
図 5.1 から図 5.2 は各手法の判定性能を比較したグラフである．図 5.1 は単位空間の外側に異常が発生した場合（MD および RD に基づく異常検知方式の両方が異常とみなすべきパターン），図 5.2 は単位空間の内側に異常が発生した場合（RD に基づく異常検知方式のみが異常とみなすべきパターン）を示す．各図とも正判別率が高い方がよい．なお，各図には PCA および SPCA に基づく推定法に加えて，母相関係数行列の固有値および固有ベクトルを始めとして，母数にすべて真値を用いた場合（以降，TRUE と呼ぶ）も併記している．なお，正判別率は単位空間およびテストデータを 100 組用意して解析を行った結果の平均である．



(a) α=0.1, c^2 = 1.1



(b) α=0.5, c^2 = 1.1

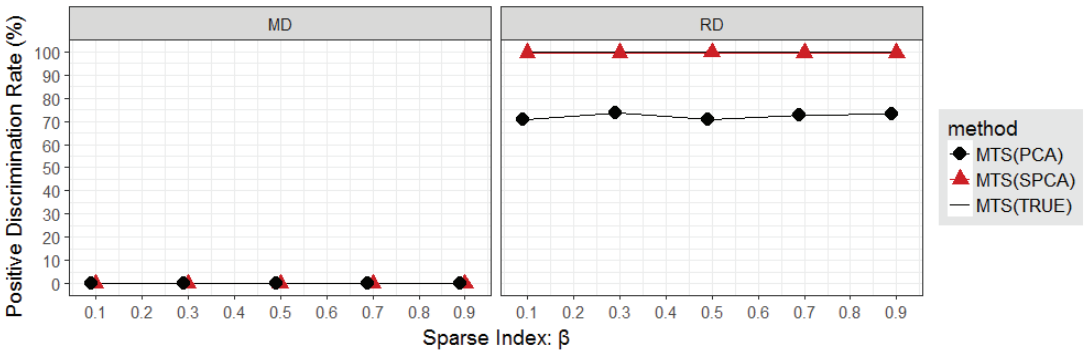


(c) α=0.9, c^2 = 1.1

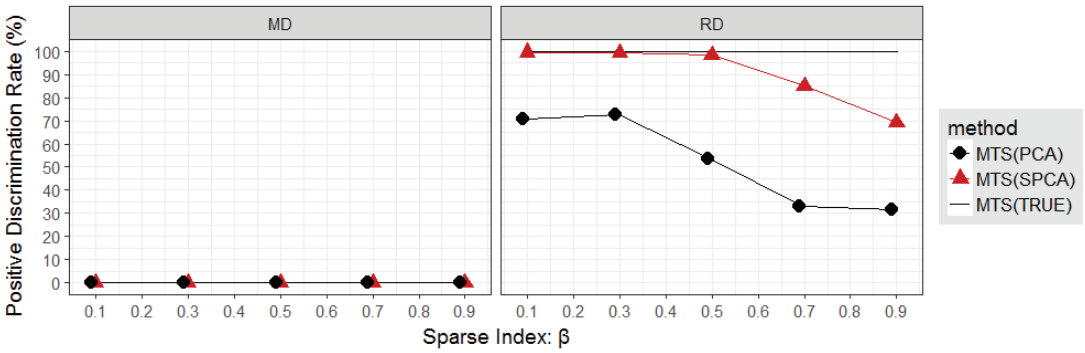
図 5.1　各手法の判定性能（ c^2 = 1.1 の場合）

図 5.1 より，単位空間の外側に異常が発生した場合，距離の違い（MD と RD）による判定性能への影響はほとんどないといえる．また，推定法に関しても PCA に基づく方法を用いても十分に判定できるパターンも多い．

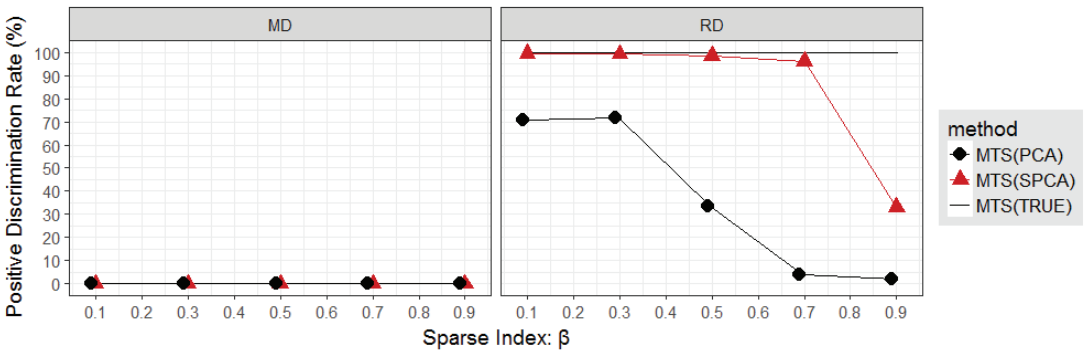
一方，図 5.2 より，単位空間の内側に異常が発生した場合，MD は正常とみなすため，手法間に差はなく，すべて正判別率（この場合に限り誤判別率と捉えてよい）が 0%となっている．それに対して RD では SPCA を導入した方が PCA に比べて性能が高くなっていることがわかる．このパターンは，高次元空間に発生するノイズがもつ球面集中



(a) α=0.1, c^2 = 0.75



(b) α=0.5, c^2 = 0.75



(c) α=0.9, c^2 = 0.75

図 5.2　各手法の判定性能（ c^2 = 0.75 の場合）

現象を応用しなければ，異常を検知できないパターンである．そのため，RD に基づく異常検知方式を用いる場合には，提案プロセスの導入効果は高いといえる．このことは次のように説明できる．

まず，第 1 固有値は常に過大に推定される（Yata and Aoshima (2012)）ことに注意する．そして，(5.8)式の統計量 *SE* および Z_2 は真の母数を用いた場合よりも小さな値をとるようになる．したがって，Z_2 の分布は，真の分布よりも内側異常が発生する方向にシフトしてしまう．それゆえ，内側異常は検知が困難になる一方で，外側異常は検知しや

すくなる傾向をもつといえる．実際，図 5.1 をみると，PCA と SPCA に基づく方法の方が真値よりも正判別率が高い場合もあることからこの傾向が確認できる．なお，本議論は第 1 固有値が回転しない場合あるいは回転しても影響が小さい場合に成立する．

5.6. 本章のまとめ

本章では，MT 法を高次元データに適用する場合を想定したうえで，統計的モデリングの具体的な実施方法について考察した．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 が遥に多いデータ（$p >> n$）を取り扱う必要がある．しかしながら，そもそも現行プロセスでは標本マハラノビス距離が計算不能となるため，異常検知を実行できない．そこで本研究では，$p >> n$ でも計算可能な 2 種類の距離を取り上げ，その距離に基づく異常検知方式の特徴について考察した．そして，そのパラメータを統計的モデリングの枠組みに基づいて推定することを提案した．提案プロセスの導入効果を数値実験で検証した結果，提案プロセスが予測と診断の両面で有用との結論を得た．特に球面集中現象を応用した異常検知方式である RT-PC 法を用いる場合には，提案プロセスの導入効果が高いことも明らかになった．

第 6 章　汚染データ解析プロセスの提案

本章では，MT システムを汚染データに適用する場合の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する．単位空間にミスラベルのデータが混入したデータを汚染データと呼ぶ．一般に汚染データを解析する場合，統計モデルのパラメータを精度よく推定することが困難となる．そのため，各種ロバスト推定法を使用する必要性が生じる．そこで本研究では，母数の推定法に焦点をあて，ロバスト推定法の導入効果について検証する．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知を実現するための解析プロセスを提案する．また数値実験を通して，提案プロセスの有用性を示す．

本章の内容は Ohkubo and Nagata (2017b)に基づく．

6.1.　はじめに

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・蓄積されるようになっている．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た設備機器の異常検知技術の確立が急務となっているといえる．しかしながら，履歴データを対象とする場合，現行 MT 法の解析プロセスでは適切な分析ができるとは限らない．蓄積されている履歴データの多くは，正常か異常かを示すラベルが付与されていない可能性があるからである．

そもそも現行 MT 法は，学習データの真のラベルがすべて正常であるという強い仮定のもとで成立している．第 2 章で示したように，MT 法は標本平均ベクトルおよび標本共分散行列を用いて計算したマハラノビス距離に基づく異常検知方式である．ところが，この標本平均ベクトルおよび標本共分散行列は，外れ値の影響を強く受けてしまうことが知られている．したがって，ホテリング多変量管理図の分野では，外れ値にロバストな推定法を導入した解析プロセスが数多く提案されている（例えば，Vargas (2003), Cetin and Aktas (2003), Alfaro and Ortega (2008, 2009), Chenouri and Variyath (2011), Haddad et al. (2013), Abu‐Shawiesh et al. (2014)等を参照されたい）．このようなロバスト推定法を MT 法に導入すれば，誤ったラベルをもつデータ（以降，ミスラベル・データと呼ぶ）が混入した場合でも，混入数が十分に小さければ，適切な分析ができるといえる．

しかしながら，全くラベルが付与されていないデータを対象とする場合，ミスラベル・データが少数である保証はないことに留意する必要がある．6.4 節および 6.5 節の数値実験で示すように，ミスラベル・データが大量に混入する場合，従来のロバスト推

定法を導入しても, MT 法の異常検知性能に深刻な影響が発生してしまう危険性がある．また，従来のロバスト推定法は，ミスラベル・データの個数だけでなく，分布に対しても暗に仮定をおいている．そのため，異常検知問題のように異常の分布が想定できない場合，ミスラベル・データの混入の状況に応じてロバスト推定法を使い分ける必要性が生じることも実用上の問題となるといえる．

そこで本研究では, 全くラベルが付与されていないデータを対象とした MT 法の新たな解析プロセスを提案する．具体的には，母数の推定法に焦点をあて，ロバスト推定法の導入効果について検証する．そして，Fujisawa and Eguchi (2008)の γ ダイバージェンスに基づくロバスト推定法を MT 法に用いる．γ ダイバージェンスに基づくロバスト推定法はミスラベル・データの混入数に対して一切の仮定をおいていない．また 6.3 節で述べるように, ミスラベル・データの分布に対する仮定も異常検知問題では自然である．したがって，γ ダイバージェンスに基づくロバスト推定法を用いれば，全くラベルが付与されていない学習データから精度よく母数を推定できる可能性がある．ゆえに，提案プロセスの異常検知性能も，ミスラベル・データが混入していない場合と同等の性能となることが期待できる．

6.2. 現行プロセスの問題点

第 3 章で述べたように, 現行 MT 法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．ここで，単位空間の分布に対して厳密な仮定をおいていないことに注意すると, 現行 MT 法は経験分布に基づくモーメント法によってパラメータ推定を行っているといえる．一方，本研究では形式的に多変量正規分布に基づく最尤推定法によってパラメータ推定を行っていると考える．本節では，この前提のもとで, 現行 MT 法におけるパラメータ推定法をダイバージェンス最小化の観点から考察する．また，汚染データを解析する場合の問題について指摘する．

いま母集団の真の分布を $g(\boldsymbol{x})$とするとき，$g(\boldsymbol{x})$を近似する統計モデル $f(\boldsymbol{x})$が与えられたとしよう. 特に$f(\boldsymbol{x})$がパラメトリックなモデルである場合, $f(\boldsymbol{x}|\boldsymbol{\theta})$あるいは$f_{\boldsymbol{\theta}}(\boldsymbol{x})$と記す. このとき，$g(\boldsymbol{x})$と $f_{\boldsymbol{\theta}}(\boldsymbol{x})$の近さが最小となるように $\boldsymbol{\theta}$ の推定量$\hat{\boldsymbol{\theta}}$を求める問題を考える．

まず，分布間の近さを定量化するための尺度として，KL ダイバージェンス（Kullback and Leibler (1951)）を次式で定義する．

$$D_{\mathrm{KL}}(g \,\|\, f) = -d_{\mathrm{KL}}(g \,\|\, g) + d_{\mathrm{KL}}(g \,\|\, f) \tag{6.1}$$

ここで，右辺の第 1 項および第 2 項は各々相互エントロピーである．すなわち，分布 $g(\boldsymbol{x})$に対する分布 $f(\boldsymbol{x})$の相互エントロピーは次式のように定義される．

$$d_{\mathrm{KL}}(g \,\|\, f) = -\int g(\boldsymbol{x}) \ln f(\boldsymbol{x}) d\boldsymbol{x} \tag{6.2}$$

次に，KL ダイバージェンスが最小化されるようにパラメトリック・モデル $f_{\theta}(\boldsymbol{x})$のパラメータ $\boldsymbol{\theta}$ の推定量$\hat{\boldsymbol{\theta}}_{KL}$を求める問題を定式化する．ここで，(6.1)式の右辺第 1 項はパラメトリック・モデル $f_{\theta}(\boldsymbol{x})$に依存しないため，第 2 項の相互エントロピーを最大化するような $\boldsymbol{\theta}$ を求めればよい．すなわち，推定量$\hat{\boldsymbol{\theta}}_{KL}$は次式で与えられる．

$$\hat{\boldsymbol{\theta}}_{KL} = \arg\min_{\theta} d_{\mathrm{KL}}(g \,\|\, f_{\theta}) \tag{6.3}$$

このとき，$f_{\theta}(\boldsymbol{x})$を尤度関数とみなすと，(6.3)式は平均対数尤度の最大化問題となる．特に母集団から抽出された大きさ n の標本に基づいて$g(\boldsymbol{x})$を経験分布の確率関数に置き換えると，(6.3)式の推定量は最尤推定法と同等であることがわかる．したがって，KL ダイバージェンスの最小化に基づくパラメータ推定法と最尤推定法は本質的に同等であるといえる．ゆえに，現行 MT 法におけるパラメータ推定法は，KL ダイバージェンスに基づくパラメータ推定法を用いていると捉えることができる．

ここで，単位空間にミスラベル・データが混入した状況を考える．このとき，KL ダイバージェンスはミスラベル・データが混入した単位空間に対する統計モデル $f_{\theta}(\boldsymbol{x})$のダイバージェンスとなる．すなわち，KL ダイバージェンスを最小化するようにパラメータを推定した場合，推定量は混入したミスラベル・データの影響を受けるため，その混入状況に依存したバイアスをもってしまう．

6.3. 提案プロセス

ラベルが全く付与されていないデータを解析対象とする場合，ミスラベル・データが学習用サンプルに含まれている危険性が高い．また，ミスラベル・データの混入数が必ずしも少数である保証がない．そこで本節では，このような前提のもとでも，MT 法による異常検知を実現するための新たな解析プロセスを提案する．すなわち，統計的モデリングを通して, 汚染データから真の単位空間を近似する統計モデルのパラメータを精度よく推定するプロセスを MT 法に導入する．

6.3.1. 統計的モデリング

本章では, 統計的モデリングのステップの中でも母数の推定法に焦点をあてる. また, そのうえで，ミスラベル・データが混入した場合でも，真の単位空間に対する統計モデル$f_{\theta}(\boldsymbol{x})$のダイバージェンスが最小化されるようにパラメータを推定する方法論を MT 法

に導入する．具体的には，Fujisawa and Eguchi (2008)の γ ダイバージェンスに基づくロバスト推定法を用いる．ただし，統計モデルには多変量正規分布を設定する．

まず，母集団の分布 $g(\boldsymbol{x})$に対して真の単位空間にミスラベル・データが混入する汚染モデルを仮定する．いま真の単位空間の分布（以降，ターゲット分布と呼ぶ）を $h(\boldsymbol{x})$, 混入したミスラベル・データの分布を $\delta(\boldsymbol{x})$とする．このとき，$g(\boldsymbol{x})$に対して

$$g(\boldsymbol{x}) = \varsigma h(\boldsymbol{x}) + (1-\varsigma)\delta(\boldsymbol{x}) \ \ (0 < \varsigma < 1) \tag{6.4}$$

を仮定する．ただし，$h(\boldsymbol{x})$と $\delta(\boldsymbol{x})$は次の条件を満たすものとする．

$$\left\{\int \delta(\boldsymbol{x}) h(\boldsymbol{x})^{\gamma_0} d\boldsymbol{x}\right\}^{\frac{1}{\gamma_0}} \approx 0 \tag{6.5}$$

ここで，γ_0 は適切に大きな正定数とする．(6.5)式の条件はターゲット分布 $h(\boldsymbol{x})$の裾にミスラベル・データが混入することを仮定している．異常検知問題では，ターゲット分布 $h(\boldsymbol{x})$の裾に発生したデータ，すなわち $h(\boldsymbol{x})$の値が十分に小さいデータを異常とみなすため，(6.5)式は自然な仮定であるといえる．

次に，$g(\boldsymbol{x})$に対する統計モデル $f(\boldsymbol{x})$の γ ダイバージェンスを次式で定義する．

$$D_\gamma(g \,\|\, f) = -d_\gamma(g \,\|\, g) + d_\gamma(g \,\|\, f) \tag{6.6}$$

ここで，右辺の第 1 項および第 2 項は各々γ 相互エントロピーと呼ばれる．すなわち，分布 $g(\boldsymbol{x})$に対する分布 $f(\boldsymbol{x})$の γ 相互エントロピーは正定数 γ を用いて次式で定義される．

$$d_\gamma(g \,\|\, f) = -\frac{1}{\gamma}\ln\int g(\boldsymbol{x}) f(\boldsymbol{x})^\gamma d\boldsymbol{x} + \frac{1}{1+\gamma}\ln\int f(\boldsymbol{x})^{1+\gamma} d\boldsymbol{x} \tag{6.7}$$

Fujisawa and Eguchi (2008)は(6.4)式の汚染モデルのもと, (6.5)式の条件が成立するならば，母集団の分布 $g(\boldsymbol{x})$に対する統計モデル $f(\boldsymbol{x})$の γ ダイバージェンスは，$h(\boldsymbol{x})$に対する $f(\boldsymbol{x})$の γ ダイバージェンスと同等であることを証明した．すなわち，統計モデル $f(\boldsymbol{x})$をパラメトリック・モデル $f_\theta(\boldsymbol{x})$に置き換えた場合，$g(\boldsymbol{x})$に対する γ ダイバージェンスを最小化するようにパラメータを推定すれば，$h(\boldsymbol{x})$に対する γ ダイバージェンスを最小化するようにパラメータを推定することと同等となる．したがって，真の単位空間の分布 $h(\boldsymbol{x})$を近似する統計モデル $f_\theta(\boldsymbol{x})$のパラメータの推定量は次式で与えられる．

$$\hat{\boldsymbol{\theta}}_\gamma = \arg\min_{\theta} d_\gamma(g \,\|\, f_\theta) \tag{6.8}$$

以上より，ミスラベル・データが混入した状況でも(6.8)式の推定量は真の単位空間の分布を近似する統計モデルのパラメータのよい推定量になると期待できる．ここで，(6.4)式および(6.5)式において，$\delta(\boldsymbol{x})$の分布および ς に対する仮定が存在しないことに留意する．混入するミスラベル・データは(6.5)式の条件さえ満たせば，特定の分布を仮定

する必要がない．そのため，異常の混入パターンに応じて解析プロセスを変更する必要がなく，実用的であるといえる．また，ミスラベル・データが大量に混入する場合でも，真の単位空間を近似する統計モデルの母数を精度よく推定できる可能性がある．

最後に，γ ダイバージェンスに基づくロバスト推定法が KL ダイバージェンスに基づく推定法の自然な拡張となることを注意しておく．ここで，$\gamma \to 0$ の極限を考えると，(6.6)式で定義した γ ダイバージェンスは(6.1)式の KL ダイバージェンスに一致する（Fujisawa and Eguchi (2008)）．したがって，γ が十分に小さければ，現行 MT 法の推定量と同等の推定量を得ることができる．ゆえに，異常検知性能を最大化するように γ を探索的に決定すれば，現行 MT 法に比べて同等かそれ以上の性能を実現できるといえる．ただし，テスト用のデータが不十分な場合には，Cross-Validation（Efron(1982)）等を用いて過適合を防ぐ必要がある．

6.3.2. 関連研究

第 2 章で述べたように MT 法は Hotelling (1947)の T^2 管理図と同様のマハラノビス距離に基づく異常検知方式である．そのため，T^2 管理図の分野において提案された解析プロセスを応用することで，汚染データに対しても適切な分析が可能となる場合がある．また，母集団の母平均ベクトルおよび母共分散行列のロバスト推定法を導入することも有用であるといえる．例えば，Rousseeuw and Driessen (1999)の Minimum Covariance Determinant（MCD）法や Olive (2004)の Median Ball Algorithm（MBA）法，Maronna and Zamar (2012)の Orthogonalized Gnanadesikan-Kettenring（OGK）法等が利用できる．

しかしながら，これまでのロバスト推定法の多くは，ミスラベル・データの混入数が十分に小さいことを仮定しなければ，精度保証が得られないことに注意する必要がある．また，混入したミスラベル・データの分布を観察し，用いる推定量を慎重に選ぶことも必須となる．例えば，Maronna and Zamar (2012)の OGK 法は，母平均ベクトルおよび母共分散行列のロバスト推定量を求める際，各変数の母平均および母分散の推定量を適切に選択する必要がある．6.5 節の数値実験で示すように，この推定量の選択は異常検知性能に大きな影響を与える．

6.4. 実データ解析

本節では実データ解析を通して，提案プロセスの導入効果を確認・検証する．

6.4.1. データセットの概要

本実験では，UCI 機械学習レポジトリの Vertebral Column データセットを使用する．Vertebral Column データセットは，各サンプルを正常な脊椎か異常な脊椎かに分類したデータセットである．変数は 6 次元であり，正常な個体は 100 個，異常な個体は 210 個である．ここで，訓練データおよびテストデータの設定は次の通りである．

訓練データ．訓練データには単位空間にミスラベル・データを混入させたデータを使用する．まず単位空間を正常な個体 100 サンプルから無作為に抽出した 80 サンプルを用いて構成する．次に，単位空間に混入させるミスラベル・データは，異常な個体 210 サンプルから無作為に抽出した $80\times p$ (p = 0, 0.1, 0.2, …, 1.0)サンプルを用いて構成する（以降，p を混入比率と呼ぶ）．

テストデータ．テストデータには，正常な個体 100 サンプルおよび異常な個体 210 サンプルをすべて使用するものとする．

6.4.2. 評価方法

本実験では次節で示す評価対象の判別性能を比較する．ここで，判定性能の評価方法は次の通りである．

判定性能．判定性能の評価指標は，テスト用異常データの正判別率を用いる．ここで，判別のための閾値は，テスト用正常データの誤判別率が 10%となるように設定する．

6.4.3. 評価対象

本実験の評価対象は，6.2 節の KL ダイバージェンスに基づく推定法，6.3 節の γ ダイバージェンスに基づくロバスト推定法および OGK 法を用いた MT 法である．ここで，3 種類の推定法の略称を KL-DIV，Gamma-DIV，OGK とする．また，その推定法に基づく MT 法の略称を MT(KL-DIV)，MT(Gamma-DIV)，MT(OGK)とする．

Gamma-DIV および OGK の設定について補足しておく．まず，Gamma-DIV の γ はテスト用の正常データに対する多変量正規分布の平均対数尤度が最大となるように決定する．γ の探索は 0.01 から 1.00 までの値を 0.01 刻みで行うものとする．次に，OGK における各変数の母平均および母分散の推定量には, Maronna and Zamar (2012)と同様の設定で Yohai and Zamar (1988)の τ 推定量を使用する．

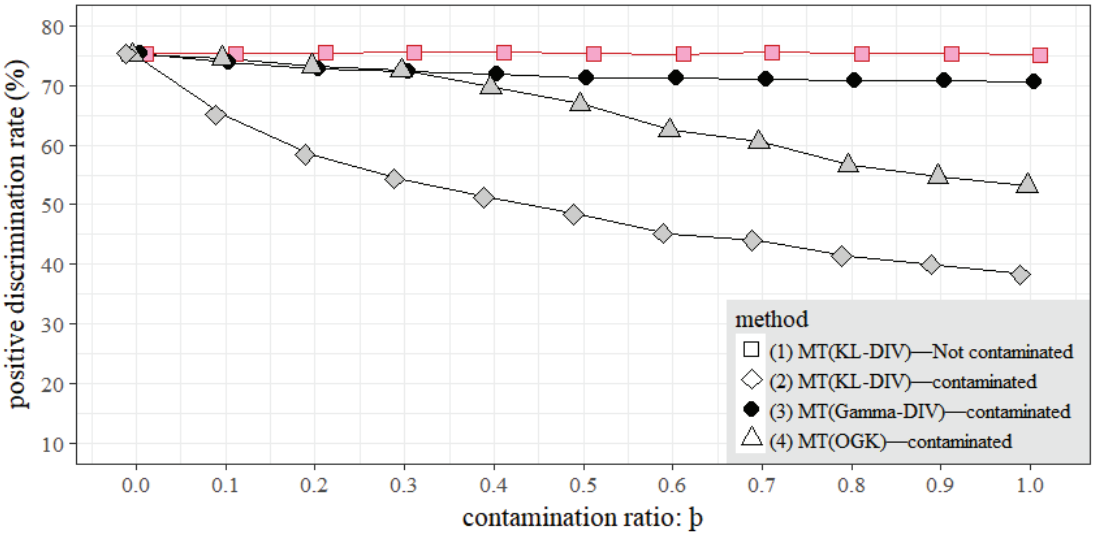


図 6.1　各手法の判定性能（Vertebral Column データセット）

6.4.4.　実験結果

実験結果を図 6.1 に示す.

図 6.1 は各手法の判定性能を示したグラフである．図中には 6.4.3 項で示した評価対象に加えて，ミスラベル・データが混入していない場合（「Not contaminated」）に KL-DIV に基づく推定法を用いた MT 法の判定性能も比較のため示している．ここで，各図の縦軸と横軸は判定性能と混入比率 p である．なお，各手法の判定性能は訓練データおよびテストデータを 100 組用意して解析を行った結果の平均である．

図 6.1 より，MT(Gamma-DIV)−contaminated は混入比率 p が増加しても MT(KL-DIV)−Not contaminated とほぼ同等の判定性能を実現できているとわかる．一方，MT(KL-DIV)−contaminated あるいは MT(OGK)−contaminated は混入比率 p が増加するほど，正判別率が大きく低下してしまうことが確認できる．特に混入比率 p が 0.5 を超える場合，MT(Gamma-DIV)−contaminated と MT(OGK)−contaminated の判定性能の差が増大している．したがって，γ ダイバージェンスに基づくロバスト推定法は，ミスラベル・データが大量混入する状況において MT 法への導入効果が高いといえる．

6.5.　モンテカルロ・シミュレーション

本節では，前節で確認した提案プロセスの導入効果の再現性を確認するため，モンテカルロ・シミュレーションを行う．

6.5.1. 生成モデル

本実験では単位空間およびミスラベル・データ，テストデータに次の生成モデルを仮定する. なお, ミスラベル・データの混入状況と判定性能の関係を的確に分析するため, ミスラベル・データとテスト用の異常データの生成モデルは異なるモデルを用いている.

単位空間. データ生成モデルは 2 変量正規分布 $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$とする．ここで，$\boldsymbol{\mu}_0$ は 2 次元ゼロベクトル，$\boldsymbol{\Sigma}_0$ は 2×2 の単位行列とする．なお，サンプル数は N_0 ($N_0 = 50$)個である.

ミスラベル・データ．単位空間の中心からの動径が長さ一定の分布に従うデータを生成する. まず, 2 次元ユークリッド空間上の点(X_1, X_2)を極座標$(\rho\cos\Phi, \rho\sin\Phi)$で表す. 次に, ρ は定数($\rho = 3, 6$)，Φ は von Mises 分布に従う確率変数としてデータを発生させる．ここで，von Mises 分布は平均方向パラメータ φ，集中パラメータ κ をもつ円周上の分布である．そして，$\kappa = 0$ のとき，円周上の一様分布となり，κ が大きな値をとるとき，平均 φ の正規分布に近づく．なお，本実験では $\varphi = 3/4\pi$，$\kappa = 0, 5, 10$ とする．また，サンプル数は 6.4.1 項の混入比率 p ($p = 0, 0.1, 0.2, \ldots, 1.0$)を用いて $N_1 = N_0 \times p$ 個とする.

テストデータ．正常データの生成モデルは単位空間と同様に $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$とする．一方，異常データの生成モデルは $N(\boldsymbol{\mu}_0, c^2\boldsymbol{\Sigma}_0)$ ($c = 7$)とする．正常データおよび異常データのサンプル数はともに 100,000 個である.

図 6.2 に訓練データの発生例を示した．図 6.2 はグラフの左から順に $\kappa = 0, 5, 10$ の各場合の単位空間およびミスラベル・データをプロットしたグラフである．なお，ρ は 3, p は 0.5 に設定している．図 6.2 より，κ が増加すると，異常のラベルをもつデータ（ミスラベル・データ）が，単位空間の中心から $3/4\pi$ の方向に集中していくことが観察できる．言い換えれば，$\kappa = 0$ のとき，ミスラベル・データは単位空間の分布の裾に均等に発生する一方，$\kappa = 5, 10$ では特定の方向に偏って発生するといえる．特に $\kappa = 10$ のとき，強い相関をもつ 2 変量正規分布に近い分布となっていることを注意しておく.

6.5.2. 評価方法

本実験では次項で示す評価対象の判別性能を比較する．ここで，判定性能の評価方法は次の通りである.

判定性能. 判定性能の評価指標はテスト用異常データの正判別率を用いる．ここで，判別のための閾値は，テスト用正常データの誤判別率が 1%となるように設定する.

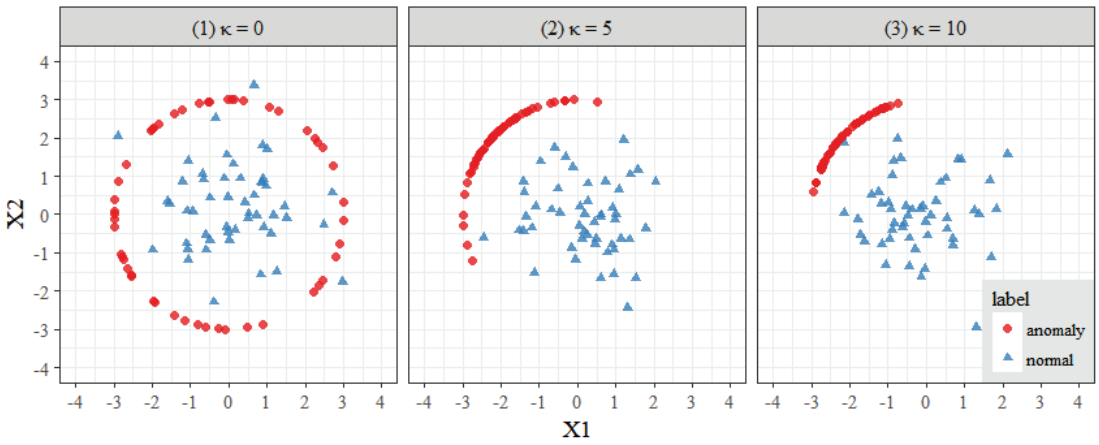


図 6.2　学習データの発生例（$\rho = 3, p = 0.5$ の場合）

6.5.3.　評価対象

本実験の評価対象は，6.2 節の KL ダイバージェンスに基づく推定法，6.3 節の γ ダイバージェンスに基づくロバスト推定法および OGK 法を用いた MT 法である．各推定法およびその推定法に基づく MT 法の略称は，6.4.3 項と同様とする．

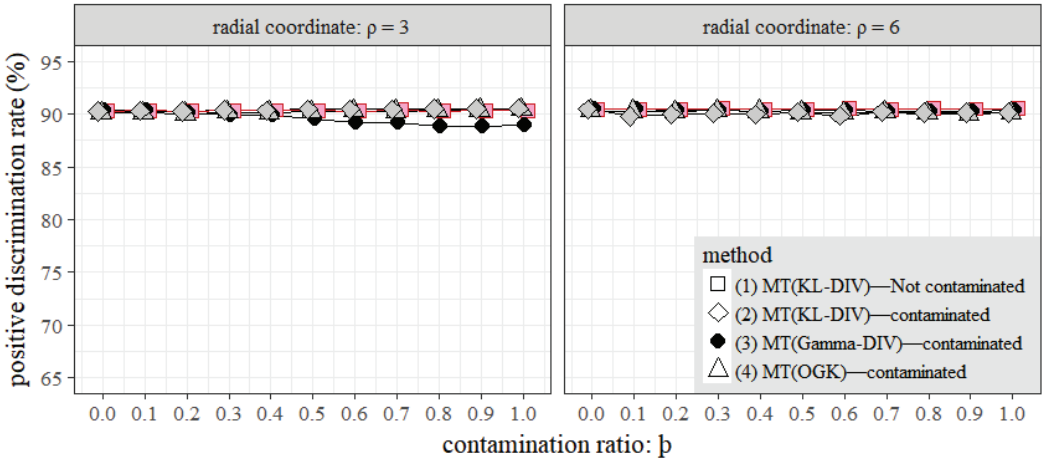
Gamma-DIV および OGK の設定は 6.4.3 項と同様である．ただし，Gamma-DIV の γ は単位空間の分布 $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$に対する統計モデル $N(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$の KL ダイバージェンスが最小となるように決定する．すなわち，次式で定義される KL ダイバージェンスが最小となるように γ を決定する．なお，γ の探索は 0.01 から 1.00 までの値を 0.01 刻みで行う．

$$-\ln\left|\boldsymbol{\Sigma}_0\hat{\boldsymbol{\Sigma}}_0^{-1}\right| + \mathrm{tr}\left(\boldsymbol{\Sigma}_0\hat{\boldsymbol{\Sigma}}_0^{-1}\right) - p + \left(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0\right)^T \hat{\boldsymbol{\Sigma}}_0^{-1}\left(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0\right) \tag{6.9}$$

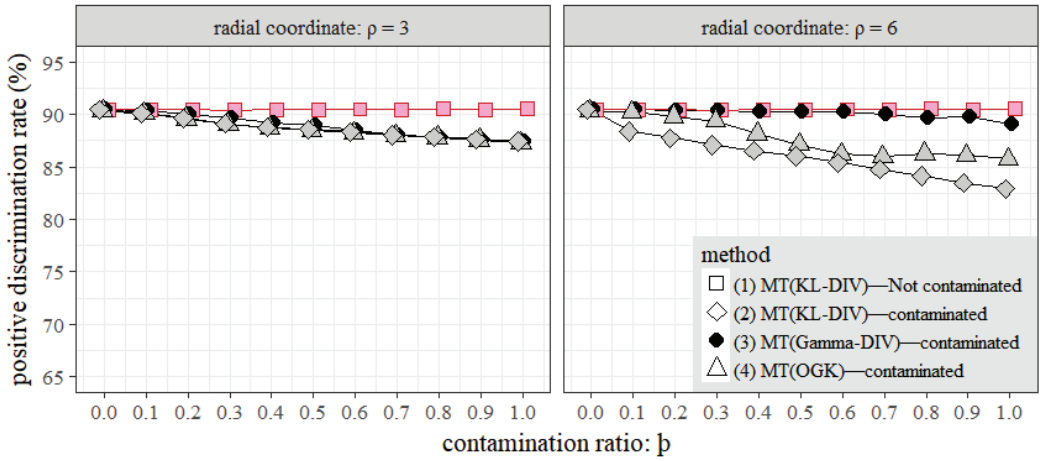
6.5.4.　実験結果

実験結果を図 6.3 に示す．

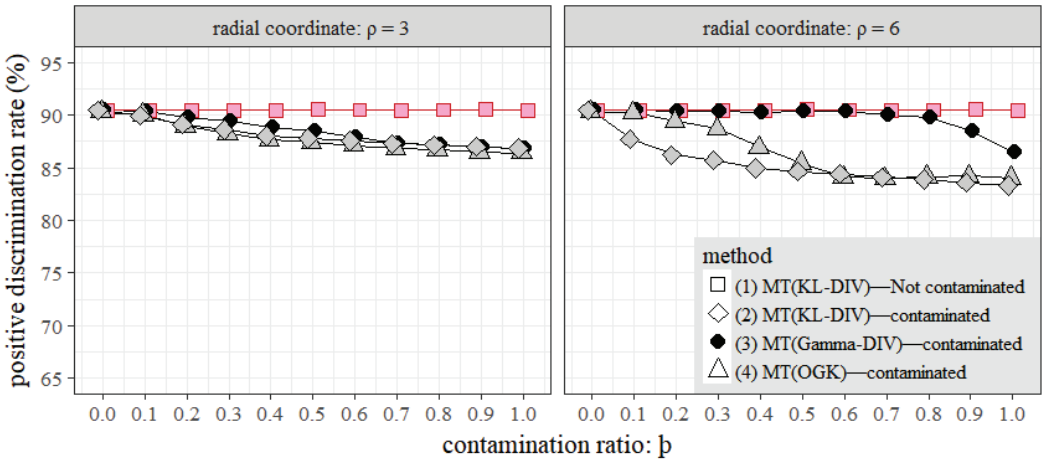
図 6.3 は各手法の判定性能を示したグラフである．図 6.3a は $\kappa = 0$，図 6.3b は $\kappa = 5$，図 6.3c は $\kappa = 10$ の場合に対応している．また，各図とも単位空間の中心からの動径の長さ ρ が 3 の場合を左側，6 の場合を右側に示している．図中の評価対象には，6.5.3 項で述べた評価対象に加えて，ミスラベル・データの混入がない場合（「Not contaminated」）に KL-DIV に基づく推定法を用いた MT 法も比較のため示している．ここで，各図の縦軸と横軸は判定性能と混入比率 p である．なお，各手法の判定性能は単位空間，ミスラベル・データ，テストデータを 100 組用意して解析を行った結果の平均である．



(a) $\kappa = 0$ の場合



(b) $\kappa = 5$ の場合



(c) $\kappa = 10$ の場合

図 6.3　各手法の判定性能（モンテカルロ・シミュレーション）

まず，図 6.3 を概観する．$\rho = 3$ の場合，手法間の優劣はほぼ存在しないといえる．一方，$\rho = 6$ の場合，MT(Gamma-DIV)−contaminated が MT(KL-DIV)−contaminated あるいは MT(OGK)−contaminated よりも高い判定性能を示すパターンが存在することが確認できる. したがって, 単位空間の中心から離れた位置にミスラベル･データが混入するほど，提案プロセスの導入効果が高いといえる．これは γ ダイバージェンスに基づくロバスト推定法が汚染モデルに対して(6.5)式を仮定しているからである．

次に，図 6.3a，6.3b，6.3c の各図における $\rho = 6$ の場合について詳細に観察する．

第 1 に図 6.3a における $\rho = 6$（右側）のパターンでは，MT(Gamma-DIV)−contaminated に加えて，他の 2 手法も混入比率 p によらず，MT(KL-DIV)−Not contaminated とほぼ同等の判定性能を実現できているとわかる．これは本パターンでは，単位空間の分布の裾に均等にミスラベル・データが混入していることに起因する．本パターンの場合，たとえ KL-DIV および OGK に基づく推定を実行しても，母平均ベクトルおよび母相関係数行列の推定には大きな影響を与えないといえる．言い換えれば，母共分散行列の推定量への影響は発生する（母共分散行列の推定量が定数倍される）ものの，相関構造は維持されるといえる．したがって，単位空間の分布の裾に均等にミスラベル・データが混入する場合，異常検知性能にはほとんど影響しない可能性がある．なお，母平均ベクトルあるいは母共分散行列の推定量を定数倍しても標本マハラノビス距離の大小関係は変化しないため，異常検知性能には影響しないことを注意しておく．

第 2 に図 6.3b における $\rho = 6$（右側）のパターンでは，MT(Gamma-DIV)−contaminated のみが混入比率 p によらず，MT(KL-DIV)−Not contaminated とほぼ同等の判定性能を実現できているとわかる．これは本パターンでは，ミスラベル・データの混入状況に偏りが発生していることに起因する．混入状況に偏りが発生した場合，KL-DIV に基づく推定にはミスラベル・データに由来するバイアスが発生する可能性がある．また OGK に基づく推定を実行する場合でも，解析プロセスを慎重に検討する必要がある．本実験では，OGK の各変数の母平均および母分散の推定量に，各変数の中央値に基づく τ 推定量を利用している．このことは単位空間の分布の裾に均等にミスラベル・データが混入することを暗に仮定している．そのため，混入状況に偏りが発生する場合には適切な解析プロセスとはならない．したがって，OGK 法を用いる場合，混入状況に応じて，中央値やトリム平均等を使い分けてロバスト推定量を計算する必要がある．

第 3 に図 6.3c における $\rho = 6$（右側）のパターンでも，MT(Gamma-DIV)−contaminated は他の 2 手法に比べて高い判定性能を実現できているとわかる．特に混入比率 p が 0.8 を超えるまでは MT(KL-DIV)−Not contaminated と同等の判定性能だといえる．ただし，

単位空間とミスラベル・データのサンプル数が同数（混入比率 p が 1.0）に近づくと，他の 2 手法とほぼ同じ水準まで判定性能が低下している．これは本パターンでは，ミスラベル･データが群を形成しているとみなせることに起因する. 図 6.2 で示したように，本パターンではミスラベル･データが強い相関をもつ 2 変量正規分布に近い分布を形成するといえる．したがって，混入比率が高くなると，Gamma-DIV がミスラベル・データの分布を近似する統計モデルの母数を推定してしまう状況が発生し得る.

最後に本実験結果に関する議論をまとめておく．まず，現行 MT 法がミスラベル・データの混入比率によらない安定的な異常検知性能を実現できる状況は非常に限定的であるといえる. また, OGK 法のような既存のロバスト推定法を MT 法に導入する場合，ミスラベル・データの混入状況に応じて解析プロセスを慎重に検討する必要がある．一方, γ ダイバージェンスに基づくロバスト推定法を MT 法に導入する場合，ミスラベル・データが群を形成していなければ, 混入比率によらない安定的な異常検知性能を実現できる可能性がある．ここで，異常検知問題では，正常な個体が均一な母集団を形成するとみなすのに対して，異常な個体は群を形成しないと仮定することに注意する．したがって，異常検知問題では提案プロセスの導入効果の再現性はあるといえる.

6.6. 本章のまとめ

本章では，MT システムを汚染データに適用する場合を想定したうえで，統計的モデリングの具体的な実施方法について考察した．一般に汚染データを解析する場合，統計モデルのパラメータを精度よく推定することが困難となる．そのため，各種ロバスト推定法を使用する必要性が生じる．そこで本研究では，母数の推定法に焦点をあて，ロバスト推定法の導入効果について検証した．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知を実現するための解析プロセスを提案した．具体的には，母数の推定法として，γ ダイバージェンスに基づくロバスト推定法を使用した．提案プロセスの導入効果を数値実験で検証した結果，提案プロセスがミスラベル・データの大量混入への対策として有用であるとの結論を得た.

第 7 章　結論

本研究では，タグチメソッドの代表的な方法論である MT システムを取り上げ，実問題を適切に分析するための新たな解析プロセスについて考察した．現行プロセスは統計的モデリングの観点から言えば，解析目的によらず予測精度を唯一の評価規準として統計モデルを評価していた．また，統計モデルのパラメータ推定法もデータの特徴を考慮せずに常に同じ推定法を使用していた．そこで本研究では，統計的モデリングの枠組みを MT システムに導入することを提案した．提案プロセスでは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．したがって，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といえる．

実際に本研究では，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の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，そ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確認した．第 1 に，小標本データの場合，提案プロセスによって異常検知性能が向上できるだけでなく，異常原因の特定に有益な知見を獲得できることを示した．第 2 に，高次元データの場合，提案プロセスによって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た．最後に，汚染データの場合，提案プロセスによってミスラベルのデータが大量に混入する状況でも安定的に異常を検知できることを示した．

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，設備機器の異常検知を実現しようという動きがある．しかしながら，このような実問題に対して現行の MT システムを適用する場合，適切な分析が実行できる対象が非常に限定的されてしまう．一方，提案プロセスでは様々な実問題に対して適切な分析ができるため，適用対象は格段に広がり，実用的な異常検知システムの実現につながるといえる．

今後の課題として，リアルタイム異常検知に向けた解析プロセスの改良を挙げる．MT システムをリアルタイム異常検知に使用する場合，単位空間データの更新等に対する適切な対応が必要となる．その際，ベイズ更新等が活用できるため，ベイズ的なアプローチが有用であると予想される．しかしながら，ベイズ的なアプローチを前提とする場合，本研究で用いたモデル選択よりも階層ベイズモデリングの実行を考えた方がよいといえる．したがって，新たなモデリング法の導入について検討する必要がある．

謝辞

本論文は著者が早稲田大学　創造理工学部および大学院　創造理工学研究科に在学中に行った研究成果をまとめたものです.

早稲田大学　創造理工学部　経営システム工学科　永田靖教授には学部時代より長期間にわたりご指導を頂きました. 統計科学に関する的確なコメントもさることながら, 研究の進め方や考え方等についても丁寧にご教示頂きました. また, 先生のご書籍からは専門知識だけでなく, 簡明な文章表現, 読者に配慮した構成等の多くの学びを得ることができました. ここに心より感謝申し上げます.

早稲田大学　創造理工学部　経営システム工学科　椎名孝之教授ならびに後藤正幸教授, 蓮池隆准教授には副査として大変有意義なご助言を頂くとともに本論文の細部まで丁寧にご指導を頂きました. ここに深謝の意を表します.

早稲田大学　創造理工学部　経営システム工学科　永田研究室の各位には本研究の遂行に際して数多くのディスカッションの時間を頂きました. 特に同期からは率直な意見をもらうことができ, 研究の質の向上につながりました. 厚く御礼申し上げます.

最後に, 学部時代より多大なるご支援を頂いている早稲田大学　創造理工学部　経営システム工学科の教員ならびにスタッフの皆様, そして日々著者の心身の支えとなってくれている家族に深く感謝いたします.

2018 年 2 月

大久保　豪人

参考文献

Abu‐Shawiesh, M. O. A., Kibria, G. & George, F. (2014), A robust bivariate control chart alternative to the Hotelling's T^2 Control Chart. *Quality and Reliability Engineering International*, **30**(1), 25-35.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. *In Proc. of 2nd International Symposium on Information Theory*, 267-281.

Alfaro, J. L. & Ortega, J. F. (2008), A robust alternative to Hotelling's T^2 control chart using trimmed estimators. *Quality and Reliability Engineering International*, **24**(5), 601-611.

Alfaro, J. L. & Ortega, J. F. (2009), A comparison of robust alternatives to Hotelling's T^2 control chart. *Journal of Applied Statistics*, **36**(12), 1385-1396.

安道知寛. (2014), 高次元データ分析の方法 ―Rによる統計的モデリングとモデル統合―, 朝倉書店.

Banerjee, O., Ghaoui, L. E. & d'Aspremont, A. (2008), Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, **9**(Mar), 485-516.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000), LOF: identifying density-based local outliers. *In ACM sigmod record*, **29**(2), 93-104.

Cetin, M. C. & Aktas, S. (2003), Hotelling's T^2 statistic based on minimum-volume-ellipsoid estimator. *GAZI University Journal of Science*, **16**(4), 691-695.

Chen, J. & Chen, Z. (2008), Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**(3), 759-771.

Chenouri, S. E. & Variyath, A. M. (2011), A comparative study of phase II robust multivariate control charts for individual observations. *Quality and Reliability Engineering International*, **27**(7), 857-865.

Dempster, A. P. (1972), Covariance selection. *Biometrics*, **28**(1), 157-175.

Efron, B. (1982), *The jackknife, the bootstrap and other resampling plans*. SIAM.

Foygel, R. & Drton, M. (2010), Extended Bayesian Information Criteria for Gaussian Graphical Models. *Advances in Neural Information Processing Systems 23 (NIPS 2010)*.

Friedman, J., Hastie, T. & Tibshirani, R. (2008), Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9** (3), 432-441.

Fujisawa, H. & Eguchi, S. (2008), Robust parameter estimation with a small bias against heavy

contamination. *Journal of Multivariate Analysis*, **99**(9), 2053-2081.

福島祥夫, 斉藤克彦, 千葉隆一, 久米原宏之. (2009), マハラノビス距離によるニアネットシェイプ鋳造モニタリングシステムに関する基礎的研究. 品質工学, **17**(2), 92-98.

Gordon, G. J., Jensen, R. V., Hsiao, L. L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S. & Bueno, R. (2002), Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, **62**(17), 4963-4967.

Haddad, F. S., Syed‐Yahaya, S. S. & Alfaro, J. L. (2013), Alternative Hotelling's T2 Charts using Winsorized Modified One‐Step M‐estimator. *Quality and Reliability Engineering International*, **29**(4), 583-593.

Hotelling, H. (1947), Multivariate quality control - illustrated by the air testing of sample bombsights. C. *Techniques of Statistical Analysis*, Eisenhart, C., Hastay, M.W. & Wallis, W.A. (eds), McGraw-Hill, 113-184.

井手剛. (2015), 入門機械学習による異常検知: R による実践ガイド. コロナ社.

Ide, T., Lozano, A. C., Abe, N. & Liu, Y. (2009), Proximity-based anomaly detection using sparse structure learning. *In Proc. of 2009 SIAM International Conference on Data Mining*, 97-108.

井手剛, 杉山将. (2015), 異常検知と変化検知. 講談社.

稲生淳紀, 永田靖, 堀田慶介, 森有紗. (2012), タグチの T 法およびその改良手法と重回帰分析の性能比較. 品質, **42**(2), 265-277.

Jackson, J. E. & Mudholkar, G. S. (1979), Control procedures for residuals associated with principal component analysis. *Technometrics*, **21**(3), 341-349.

Jugulum, R., Taguchi, G., Taguchi, S. & Wilkins, O. J. (2003), Discussion. *Technometrics*, **45**(1), 16-21.

Jung, S. & Marron, J. S. (2009), PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, **37**(6B), 4104-4130.

Jung, S., Sen, A. & Marron, J. S. (2012), Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis*, **109**, 190-203.

兼高達貳. (1987), マハラノビスの汎距離の応用例-特殊健康診断の事例-. 標準化と品質管理 **40**(11), 46-54.

小西貞則, 北川源四郎. (2004), 情報量規準. 朝倉書店.

Kullback, S. & Leibler, R. A. (1951), On information and sufficiency. *The Annals of*

Mathematical Statistics, **22**(1), 79-86.

Lauritzen, S. L. (1996), *Graphical models*. Oxford University Press.

間ヶ部明, 高田圭, 矢野宏. (1998), はんだ自動外観検査へのマハラノビスの距離の適用. 品質工学, **6**(6), 66-73.

Maronna, R. A. & Zamar, R. H. (2012), Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44**(4), 307-317.

松田里香, 池田佳起, 鴨下隆志, 東原和行. (2002), MTS法の将来宇宙機用ソフトウェアへの適用. 品質工学, **10**(1), 37-41.

Mazumder, R. & Hastie, T. (2012), The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, **6**, 2125-2149.

McCabe, G. P. (1984), Principal variables. *Technometrics*, **26**(2), 137-144.

宮川雅巳. (2000), 品質を獲得する技術: タグチメソッドがもたらしたもの. 日科技連出版社.

宮川雅巳, 永田靖. (2003), マハラノビス・タグチ・システムにおける多重共線性対策について. 品質 **33**(4), 77-85.

宮川雅巳, 田中研太郎, 岩澤智之, 中西寛子. (2007), マハラノビス・タグチ・システムにおける実際の誤判別率. 品質 **37**(1), 101-106.

永田靖. (2013), MTシステムの諸性質と改良手法. 応用統計学, **42**(3), 93-119.

永田靖. (2017), MTシステムの研究. 国際学研究 6(2), 29-36, 2017.

永田靖, 土居大地. (2009), タグチの RT 法で用いる距離の性質とその改良. 品質, **39**(3), 364-375.

中島尚登, 矢野耕也, 高田圭, 高木一郎, 小宮佐和子, 大畑充, 戸田剛太郎. (2004), 各種肝疾患に対するマハラノビスの距離による病態評価. 品質工学, **12**(3), 51-58.

大久保豪人, 永田靖. (2012), タグチの RT 法における同一次元でない連続量データへの適用方法. 品質, **42**(2), 248-264.

大久保豪人, 永田靖. (2015), MT システムにおける小標本データの解析方法. 日本経営工学会論文誌, **66**(1), 30-38.

大久保豪人, 永田靖. (2017), グラフィカル・モデリングに基づくマハラノビス・タグチ法, 応用統計学, **46**(1), 13-26.

Ohkubo, M. & Nagata, Y. (2017a), Anomaly detection in high-dimensional data with the Mahalanobis-Taguchi system. *20th QMOD conference on Quality and Service Sciences ICQSS*.

Ohkubo, M. & Nagata, Y. (2017b), Anomaly detection in contaminated unit space using the Mahalanobis-Taguchi system. *Asian Network for Quality (ANQ) Congress 2017*, Kathmandu.

Olive, D. J. (2004), A resistant estimator of multivariate location and dispersion. *Computational statistics & data analysis*, **46**(1), 93-102.

Rousseeuw, P. J. & Driessen, K. V. (1999), A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**(3), 212-223.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. (2001), Estimating the support of a high-dimensional distribution. *Neural computation*, **13**(7), 1443-1471.

Schwarz, G. (1978), Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461-464.

Shen, H. & Huang, J. Z. (2008), Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99**(6), 1015-1034.

芝野広志, 安永英明. (1999), マハラノビスの距離を活用した企業経営状態の把握. 品質工学, **7**(4), 41-48.

Sill, M., Saadati, M. & Benner, A. (2015), Applying stability selection to consistently estimate sparse principal components in high-dimensional molecular data. *Bioinformatics*, **31**(16), 2683-2690.

田口玄一. (1995), パターン認識のための品質工学(3). 品質工学, **3**(4), 2-5.

田口玄一. (2002), 機能と機能性(8) 20世紀のMTS法と21世紀のMT法. 標準化と品質管理, **55**(2), 61-70.

田口玄一. (2005), 目的機能と基本機能(6) T法による総合予測. 品質工学, **13**(3), 309-314.

田口玄一. (2006), 目的機能と基本機能(11) 認識のためのT法. 品質工学, **14**(2), 171-175.

Taguchi, G. & Jugulum, R. (2002), *The Mahalanobis-Taguchi Strategy: A Pattern Technology System*. John Wiley and Sons.

高濱正幸, 三上尚高. (2012), ガスタービンプラントの異常予兆検知. 品質工学 **20**(4), 437-443.

高橋和仁, 大和俊明, 池田和子, 鴨下隆志, 矢野宏. (2003), シュミットの直交展開を用いた測定者の能力評価. 品質工学, 11(3), 62-70.

立林和夫. (2013), 実験計画法・タグチメソッドの活用. 応用統計学, **42**(3), 161-171.

立林和夫, 手島昌一, 長谷川良子. (2008), 入門MTシステム. 日科技連出版社.

Tipping, M. E. & Bishop, C. M. (1999), Mixtures of probabilistic principal component analyzers. *Neural Computation*, **11**, 443-482.

Tracy, N. D., Young, J. C. & Mason, R. L. (1992), Multivariate Control Charts for Individual Observations. *Journal of Quality Technology*, **24**(2), 88-95.

Vargas, N, J. A. (2003), Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, **35**(4), 367-376.

山西健司. (2009), データマイニングによる異常検知. 共立出版.

Yata, K. & Aoshima, M. (2009), PCA consistency for non-Gaussian data in high dimension, low sample size context. *Communications in Statistics—Theory and Methods*, **38**(16-17), 2634-2652.

Yata, K. & Aoshima, M. (2012), Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, **105**(1), 193-215.

Yohai, V. J. & Zamar, R. H. (1988), High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical Association*, **83**(402), 406-413.

研究業績

種類別	題名、発表・発行掲載誌名、発表・発行年月、連名者（申請者含む）
論文	（学術誌原著論文）
○	[1] グラフィカル・モデリングに基づくマハラノビス・タグチ法. 応用統計学, Vol.46, No.1, pp.13-26, 2017.7. 大久保豪人, 永田靖.
○	[2] MT システムにおける小標本データの解析方法. 日本経営工学会論文誌, Vol.61, No.1, pp.30-38, 2015.4. 大久保豪人, 永田靖.
○	[3] タグチの RT 法における同一次元でない連続量データへの適用方法. 品質, Vol.42, No.2, pp.86-102, 2012.4. 大久保豪人, 永田靖.
講演	（国際会議）
○	[1] Anomaly detection in contaminated unit space using the Mahalanobis-Taguchi system. ANQ Congress 2017, Kathmandu, 2017.9. Masato OHKUBO and Yasushi NAGATA.
○	[2] Anomaly detection in high-dimensional data with the Mahalanobis-Taguchi system. 20th QMOD conference on Quality and Service Sciences ICQSS, Copenhagen / Elsinore, Denmark and Helsingborg, Sweden, 2017.8. Masato OHKUBO and Yasushi NAGATA.
	[3] Applying Graphical Modeling to the Mahalanobis-Taguchi Method. ANQ Congress 2016, Vladivostok, 2016.9. Masato OHKUBO and Yasushi NAGATA.

種類別	題名、発表・発行掲載誌名、発表・発行年月、連名者（申請者含む）
	（国内会議） [1] MT システムによる高次元データ解析. 日本品質管理学会研究発表会（関西支部）発表要旨集 Vol.115, pp.13-16, 大阪, 2017.9. 大久保豪人, 永田靖. [2] 単位空間の汚染にロバストなマハラノビス・タグチ法. 統計関連学会連合大会講演報告集, pp.355, 愛知, 2017.9. 大久保豪人, 永田靖. [3] γダイバージェンスに基づく MT 法. 日本品質管理学会研究発表会研究発表要旨集 Vol.113, pp.69-72, 東京, 2017.5. 大久保豪人, 永田靖. [4] スパース・モデリングを応用したマハラノビス・タグチ法による異常検知. 情報処理学会第 79 回全国大会講演論文集(2), pp.51-52, 愛知, 2017.3. 大久保豪人, 永田靖. [5] グラフィカル・モデリングに基づく MT 法. 日本品質管理学会年次大会講演・研究発表要旨集 Vol.46, pp.165-168, 東京, 2016.11. 大久保豪人, 永田靖. [6] MT システムにおける小標本データの解析方法. 日本品質管理学会研究発表会研究発表要旨集 Vol.101, pp.43-46, 東京, 2013.5. 大久保豪人, 永田靖.

種類別	題名、発表・発行掲載誌名、発表・発行年月、連名者（申請者含む）
	[7] 次元圧縮を用いた MT システムにおける判定方法. 日本品質管理学会年次大会講演・研究発表要旨集 Vol.42, pp.53-56, 石川, 2012.10. 大久保豪人, 永田靖. [8] タグチの RT 法におけるアンサンブル学習の導入. 日本品質管理学会研究発表会研究発表要旨集 Vol.98, pp.217-220, 東京, 2012.5. 大久保豪人, 永田靖. [9] 同一次元でない連続量データのためのタグチの RT 法の改良手法. 日本品質管理学会研究発表会研究発表要旨集 Vol.95, pp.153-156, 東京, 2011.5. 大久保豪人, 永田靖.
その他	（講演） [1] スパース・モデリングに基づくマハラノビス・タグチ・システム. 第 11 回日本統計学会春季集会, 企画セッション 4: テクノメトリックスー品質改善を実践するための数理統計ー, 講演 3, 東京, 2017.3. 大久保豪人.