福山平成大学経営学部紀要
第１３号（2016），*－**頁

社会システム分析のための統合化プログラム３０
－異常検知－

福井正康[*1]、大山知之[*2]、織田望[*2]
[*1] 福山平成大学経営学部経営学科
[*2] 日本ニューマチック工業株式会社

要旨：我々は教育分野での利用を目的に社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成してきた。今回は品質管理で実践される異常検知についてのプログラムを紹介する。ここでは複数変数の正規性を持つデータと正規性を持たないデータ、時系列データ、入力出力がある場合のデータの異常検知についてその手法を示し、プログラムの利用法について紹介する。

キーワード：College Analysis、経営科学、品質管理、異常検知、機械学習、EM アルゴリズム
URL：http://www.heisei-u.ac.jp/ba/fukui/

1. はじめに

製造現場における検査過程では多くのデータが測定されるが、正常なデータと異常なデータの迅速な選別は品質管理の上で非常に重要である。ここではその主要な、複数変数の異常検知、時系列データの異常検知、入力と出力がある場合の異常検知について、その方法を説明し、分析を実行するプログラムを紹介する。このプログラムは参考文献 [1] の手法に基づいている。

複数変数の異常検知では、データが多変量正規分布に従うと仮定される場合とそうでない場合を扱う。データが多変量正規分布に従う場合、マハラノビス距離の 2 乗を元にしたホテリングの t^2 統計量に基づく判定法を用いる。また、多変量正規分布に従わない場合は、混合正規分布モデルを仮定する方法を用いている。ここでは EM アルゴリズムを用いて混合正規分布のパラメータを推定している。また、１次元データについては、ガンマ分布による異常検知の方法も加えている。

時系列データの異常検知では、周波数変化や波形変化に対応した検知法に特異スペクトル変換法を用いている。

入力と出力がある場合のデータでは、重回帰分析を用いている。但し、問題は説明変数の多重共線性である。これに対してプログラムではリッジ回帰分析と PLS 回帰分析を加えてある。これらを利用することにより、多重共線性の問題は回避される。

2. 異常検知の理論

2.1 複数変数の異常検知

1) 多変量正規分布に基づく異常検知

一般に p 変量正規分布の密度関数は以下で与えられる。

$$f(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}\,|}}\exp\left[-\frac{1}{2}\,{}^t(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

データ $D=\left\{\mathbf{x}_1,\cdots,\mathbf{x}_N\right\}$ が与えられた場合の対数尤度関数は以下で与えられる。

$$L(\mu,\Sigma\,|\,D) = -\frac{pN}{2}\log(2\pi) - \frac{N}{2}\log|\,\boldsymbol{\Sigma}\,| - \frac{1}{2}\sum_{\lambda=1}^{N}{}^t(\mathbf{x}_\lambda-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_\lambda-\boldsymbol{\mu}) \quad (1)$$

我々は最尤法を用いて(1) 式を最大化するが、その解は以下となる。

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{\lambda=1}^{N}\mathbf{x}_\lambda\,,\quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\lambda=1}^{N}(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}})^t(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}})$$

ここで、同じ正規分布の確率変数 $\mathbf{x}'$ に対する異常度 $a(\mathbf{x}')$ を $-2\log f(\mathbf{x}'\,|\,\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ を元に以下のように定義する。

$$a(\mathbf{x}') = {}^t(\mathbf{x}'-\hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}'-\hat{\boldsymbol{\mu}})$$

ここで上式と $-2\log f(\mathbf{x}'\,|\,\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ の差は定数であるので、評価関数として本質的な差はない。また上式は１次元変数の場合の変数の標準化の一般形である。

異常度の式については、以下のように定数を掛けると、分布が自由度 $p, N-p$ の F 分布に従うことが知られている。

$$T^2 \equiv \frac{N-p}{(N+1)p}\,{}^t(\mathbf{x}'-\hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}'-\hat{\boldsymbol{\mu}}) \sim F_{p,N-p}$$

この T^2 をホテリング統計量という。

異常検知には、この統計量を使って確率の値を指定するか、直接 T^2 値を指定して閾値とする。

2) 混合多変量正規分布に基づく異常検知

p 変数、 n 種混合多変量正規分布の密度関数は、種類 α の確率密度関数

$$f_\alpha(\mathbf{x}\,|\,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha) = \frac{1}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}_\alpha\,|}}\exp\left[-\frac{1}{2}\,{}^t(\mathbf{x}-\boldsymbol{\mu}_\alpha)\boldsymbol{\Sigma}_\alpha^{-1}(\mathbf{x}-\boldsymbol{\mu}_\alpha)\right]$$

（$\alpha=1,\cdots,n$）

を利用して以下で与えられる。

$$f(\mathbf{x}\,|\,\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_1,\cdots,\boldsymbol{\Sigma}_\alpha)=\sum_{\alpha=1}^{n}\pi_\alpha f_\alpha(\mathbf{x}\,|\,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha)$$
$$=\sum_{\alpha=1}^{n}\frac{\pi_\alpha}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}_\alpha\,|}}\exp\left[-\frac{1}{2}{}^t(\mathbf{x}-\boldsymbol{\mu}_\alpha)\boldsymbol{\Sigma}_\alpha^{-1}(\mathbf{x}-\boldsymbol{\mu}_\alpha)\right]$$

ここに、π_α は群α の生起確率である。

この密度関数に従うデータ $D=\{\mathbf{x}_1,\cdots,\mathbf{x}_N\}$ による対数尤度は以下である。

$$L(\pi_1,\cdots,\pi_n,\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_1,\cdots,\boldsymbol{\Sigma}_n\,|\,D)=\sum_{\lambda=1}^{N}\log\left[\sum_{\alpha=1}^{n}\pi_\alpha f_\alpha(\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha\,|\,\mathbf{x}_\lambda)\right]$$
$$=\sum_{\lambda=1}^{N}\log\left[\sum_{\alpha=1}^{n}\frac{\pi_\alpha}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}_\alpha\,|}}\exp\left\{-\frac{1}{2}{}^t(\mathbf{x}_\lambda-\boldsymbol{\mu}_\alpha)\boldsymbol{\Sigma}_\alpha^{-1}(\mathbf{x}_\lambda-\boldsymbol{\mu}_\alpha)\right\}\right] \quad (2)$$

最尤法を用いてこの対数尤度の最大値を求めるが、その際以下のアルゴリズムを利用する。

①パラメータ $\pi_\alpha,\boldsymbol{\mu}_\alpha,\boldsymbol{\Sigma}_\alpha$ に初期値 $\hat{\pi}_\alpha,\hat{\boldsymbol{\mu}}_\alpha,\hat{\boldsymbol{\Sigma}}_\alpha$ を与える。

② $\hat{\pi}_\alpha,\hat{\boldsymbol{\mu}}_\alpha,\hat{\boldsymbol{\Sigma}}_\alpha$ の値を用いて、各データの群α への帰属度 $q_\alpha(\mathbf{x}_\lambda)$ を以下で求める。

$$q_\alpha(\mathbf{x}_\lambda)=\frac{\hat{\pi}_\alpha f_\alpha(\mathbf{x}_\lambda\,|\,\hat{\boldsymbol{\mu}}_\alpha,\hat{\boldsymbol{\Sigma}}_\alpha)}{\sum_{\beta=1}^{n}\hat{\pi}_\beta f_\beta(\mathbf{x}_\lambda\,|\,\hat{\boldsymbol{\mu}}_\beta,\hat{\boldsymbol{\Sigma}}_\beta)}$$

③この帰属度を使い、新しいパラメータを以下のように決定する。

$$\hat{\pi}_\alpha=\frac{1}{N}\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda),\quad \hat{\boldsymbol{\mu}}_\alpha=\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda)\mathbf{x}_\lambda\Big/\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda),$$
$$\hat{\boldsymbol{\Sigma}}_\alpha=\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda)(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}}_\alpha)^t(\mathbf{x}_\lambda-\hat{\boldsymbol{\mu}}_\alpha)\Big/\sum_{\lambda=1}^{N}q_\alpha(\mathbf{x}_\lambda)$$

④新しいパラメータと元のパラメータを比較し、十分近ければ（プログラムではすべての成分が 0.001 未満）終了し、そうでなければ②へ戻る。

この方法によって求めたパラメータを使って、異常判定には以下の指標を用いる。

$$a(\mathbf{x}')=-\log\sum_{\alpha=1}^{n}\hat{\pi}_\alpha f_\alpha(\mathbf{x}'\,|\,\hat{\boldsymbol{\mu}}_\alpha,\hat{\boldsymbol{\Sigma}}_\alpha)$$

判定基準は各個体についてこの指標を小さい方から順番に並べ、分位点を閾値として決めるか、直接指標の閾値を指定する。

このモデルの適合度は赤池情報量基準 AIC、ベイズ情報量基準 BIC などを使って求める。今このモデルのパラメータ数をM_nとすると、AIC と BIC はそれぞれ以下のように表現される。

$$AIC = -2L(\Theta \mid D) + 2M_n$$

$$BIC = -2L(\Theta \mid D) + M_n \log N, \quad M_n = \frac{p}{2}(n+1)(n+2)$$

ここに、$L(\Theta \mid D)$ はパラメータをΘ で代表させて書いた対数尤度である。具体的には(2)式に、求めたパラメータの値を代入したものである。適合度を求めるために、交差検証を使う方法も考えられるが、プログラムでは使用していない。

2.2 時系列データの異常検知

時系列データの異常検知で、周波数変化や波形変化に対応した検知法に特異スペクトル変換法がある。これは、現時刻tに対して、時系列データx_{t-w}からx_{t-1}までのw個のデータを抽出し、それを１つのベクトル${}^t\mathbf{x}_{t-w} = (x_{t-w}\ x_{t-w+1} \cdots x_{t-1})$とする。そのベクトルの開始時点を１つずつずらして$k$ （$k < w$）本並べ、以下のような行列を作る。

$$\mathbf{X}_1^{(t)} = (\mathbf{x}_{t-k-w+1}\ \mathbf{x}_{t-k-w+2} \cdots \mathbf{x}_{t-w})$$

これに対して時間Lだけ経過した時点から作ったデータを以下のように$\mathbf{X}_2^{(t)}$とする。

$$\mathbf{X}_2^{(t)} = (\mathbf{x}_{t-k-w+1+L}\ \mathbf{x}_{t-k-w+2+L} \cdots \mathbf{x}_{t-w+L})$$

この二つのデータの間で違いを見ることになる。これらのデータ数やずれ等は、状況によって適当な値に設定する。

今、${}^t\mathbf{X}_1^{(t)}$, ${}^t\mathbf{X}_2^{(t)}$の各列ベクトルに係数をかけて足し合わせ、特徴的な量を求めるために、その大きさを以下のように最大化する。

$$\left\| {}^t\mathbf{X}_1^{(t)}\mathbf{u}^{(t)} \right\|^2 \rightarrow \text{最大化}\quad \text{但し、} {}^t\mathbf{u}^{(t)}\mathbf{u}^{(t)} = 1\ ,\ \mathbf{u}^{(t)} \text{は} (w \times 1) \text{ベクトル}$$

$$\left\| {}^t\mathbf{X}_2^{(t)}\mathbf{v}^{(t)} \right\|^2 \rightarrow \text{最大化}\quad \text{但し、} {}^t\mathbf{v}^{(t)}\mathbf{v}^{(t)} = 1\ ,\ \mathbf{v}^{(t)} \text{は} (w \times 1) \text{ベクトル}$$

このベクトルは、以下の固有方程式から求められ、

$$\mathbf{X}_1^{(t)}\,{}^t\mathbf{X}_1^{(t)}\mathbf{u}^{(t)} = \lambda \mathbf{u}^{(t)}$$

$$\mathbf{X}_2^{(t)}\,{}^t\mathbf{X}_2^{(t)}\mathbf{v}^{(t)} = \mu \mathbf{v}^{(t)}$$

固有値の大きい順に固有ベクトルをm個並べて表した$(w \times m)$行列を以下のように定義する。

$$\mathbf{U}_m^{(t)}(w \times m) = \left(\mathbf{u}_1^{(t)}\ \mathbf{u}_2^{(t)} \cdots \mathbf{u}_m^{(t)}\right)$$

$$\mathbf{V}_m^{(t)}(w \times m) = \left(\mathbf{v}_1^{(t)}\ \mathbf{v}_2^{(t)} \cdots \mathbf{v}_m^{(t)}\right)$$

これら２つの行列を用いて、変化度$a(t)$は行列${}^t\mathbf{U}_m^{(t)}\mathbf{V}_m^{(t)}$ $(m \times m)$の２ノルムを用いて以下のように定義される。

$$a(t)=1-\left\| {}^t\mathbf{U}_m^{(t)}\mathbf{V}_m^{(t)} \right\|_2^{\ 2}=1-\left({}^t\mathbf{U}_m^{(t)}\mathbf{V}_m^{(t)} \text{の最大特異値} \right)^2$$

ここに、実行列 $\mathbf{A}$ の 2 ノルム $\left\|\mathbf{A}\right\|_2$ は以下のように定義される。

$$\left\|\mathbf{A}\right\|_2=\max_{\mathbf{u}}\sqrt{\frac{{}^t(\mathbf{Au})\mathbf{Au}}{{}^t\mathbf{uu}}}=\sqrt{\lambda_{\max}}$$

注）行列 $\mathbf{A}$ の特異値とは、行列 $\mathbf{A}$ と $\mathbf{A}$ の随伴行列 $\mathbf{A}^*$ の積 $\mathbf{AA}^*$（$\mathbf{A}^*\mathbf{A}$）の非負の固有値の平方根である。

2.3　入力と出力がある異常検知

入力と出力の関係で生じる異常の検知方法については、重回帰分析を用いる手法が考えられる。しかし、重回帰分析は、入力変数が多くその値が似通っている場合に、多重共線性の問題が発生する可能性があり、予測が不安定となる。これに対して改善方法と考えられている代表的な手法がリッジ回帰分析と PLS 回帰分析である。リッジ回帰分析は、多重共線性の元となる分散共分散行列に手を加える手法であり、PLS 回帰分析は多重共線性を与える変数間の自由度を制約する手法である。我々のプログラムは 3 者を比較するように作成しており、その違いを理解し易くなっている。

1）重回帰分析

重回帰分析の目的変数を y_λ（$\lambda=1,2,\cdots,N$）、説明変数を $x_{i\lambda}$（$i=1,2,\cdots,p$）とし、それらの関係を ε_λ を誤差項として以下とする。

$$y_\lambda=\sum_{i=1}^{p}b_ix_{i\lambda}+b_0+\varepsilon_\lambda$$

最小 2 乗法としての重回帰分析では、以下の値 D が最小になるように、パラメータ b_i, b_0 を決定する。

$$D=\sum_{\lambda=1}^{N}\left(y_\lambda-\sum_{i=1}^{p}b_ix_{i\lambda}-b_0 \right)^2={}^t(\mathbf{y}-\mathbf{Xb})(\mathbf{y}-\mathbf{Xb})$$

ここに、

$$(\mathbf{X})_{\lambda i}=\tilde{x}_{i\lambda}=x_{i\lambda}-\overline{x}_i\ ,\ (\mathbf{y})_\lambda=\tilde{y}_\lambda=y_\lambda-\overline{y}\ ,\ \mathbf{b}={}^t(b_1,b_2,\cdots,b_p)$$

である。パラメータは以下で与えられる。

$$\mathbf{b}=({}^t\mathbf{XX})^{-1}{}^t\mathbf{Xy}\ ,\ b_0=\overline{y}-\sum_{i=1}^{p}b_i\overline{x}_i$$

問題となる多重共線性は、行列 ${}^t\mathbf{XX}$ の非正則性から生じる。

多重共線性の判定については、i 番目の説明変数を、他の説明変数で予測して重相関係数 r_i を求め、以下の式で定義される VIF 指標を利用している。

$$VIF_i = 1/(1-r_i^2)$$

一般に VIF 指標が 10 以上であれば多重共線性の疑いがあるとみなされる。

異常度について、通常の重回帰分析では以下で定義する。

$$a(y', \mathbf{x}') = \frac{(y' - b_0 - \mathbf{b}\mathbf{x}')^2}{\sigma^2}$$

ここで分散 σ^2 については以下で推測する。

$$\sigma^2 = \frac{D}{N} = \frac{1}{N}\,{}^t(\mathbf{y} - \mathbf{X}\mathbf{b})(\mathbf{y} - \mathbf{X}\mathbf{b})$$

2) リッジ回帰分析

リッジ回帰分析は重回帰分析の多重共線性の問題に対して、以下のように置くことによって正則性を確保しようとする手法である。

$$\mathbf{b}' = ({}^t\mathbf{X}\mathbf{X} + \eta\mathbf{I})^{-1}\,{}^t\mathbf{X}\mathbf{y}$$

これは、以下を最小化する解でもある。

$$D' = {}^t(\mathbf{y} - \mathbf{X}\mathbf{b}')(\mathbf{y} - \mathbf{X}\mathbf{b}') + \eta\,{}^t\mathbf{b}'\mathbf{b}'$$

ここでパラメータ η の値は以下のようにして求められる。λ 番目の個体を抜いた 1 個抜き交差検証のリッジ回帰パラメータを $\mathbf{b}'^{(-\lambda)}$ とすると、そのときの平均 2 乗誤差 $e(\eta)$ は以下で与えられる。

$$e(\eta) = \frac{1}{N}\sum_{\lambda=1}^{N}(\tilde{y}_\lambda - \sum_{i=1}^{p}\tilde{x}_{i\lambda}b_i'^{(-\lambda)})^2, \qquad \tilde{y}_\lambda = y_\lambda - \overline{y},\ \tilde{x}_{i\lambda} = x_{i\lambda} - \overline{x}$$

これは、以下のように書くこともできる[1]。

$$e(\eta) = \frac{1}{N}\,{}^t\mathbf{A}\mathbf{A}$$

ここに、

$$\mathbf{A} = diag(\mathbf{I} - \mathbf{H})^{-1}(\mathbf{I} - \mathbf{H})\mathbf{y},\quad \mathbf{H}(N \times N) = \mathbf{X}({}^t\mathbf{X}\mathbf{X} + \eta\mathbf{I})^{-1}\,{}^t\mathbf{X}$$

また、$diag(\mathbf{I} - \mathbf{H})^{-1}$ は対角要素が $(\mathbf{I} - (\mathbf{H})_{ii})^{-1}$ となる対角行列である。運用上はパラメータ η の値を変化させて、この $e(\eta)$ が最小になるようなパラメータ η を選ぶ。

もう少し安全性を考えて、以下の一般化交差確認検証法と呼ばれる方法から与えられる誤差 $e_{GCV}(\eta)$ を最小化する場合もある。

$$e_{GCV}(\eta) = \frac{1}{N}\,{}^t\mathbf{A}'\mathbf{A}'$$

ここに、$\mathbf{A}' = (\mathbf{I} - \mathbf{H})\mathbf{y} / [1 - t\mathbf{H}/N]$ である。我々のプログラムでは前者の判定法を利用している。

多重共線性がある場合、重回帰分析の予測は、そのデータに対してだけは良い精度を与えるが、他の新しいデータを用いた場合、予測の精度が著しく低下する。そのため１個抜き交差検証は必須である。

異常度について、リッジ回帰分析では以下で定義する。

$$a(y', \mathbf{x}') = \frac{(y' - b_0' - \mathbf{b}'\mathbf{x}')^2}{\sigma'^2}$$

ここで分散 σ'^2 については以下で推測する。

$$\sigma'^2 = \frac{D'}{N} = \frac{1}{N}[{}^t(\mathbf{y} - \mathbf{X}\mathbf{b}')(\mathbf{y} - \mathbf{X}\mathbf{b}') + \eta\,{}^t\mathbf{b}'\mathbf{b}']$$

我々のプログラムでは、重相関係数と寄与率について、残差分散を通常の重回帰分析通り以下で計算している。

$$\sigma^2 = \frac{1}{N}\,{}^t(\mathbf{y} - \mathbf{X}\mathbf{b}')(\mathbf{y} - \mathbf{X}\mathbf{b}')$$

そのため、出力される分散の値が他の分析（例えば以下の PLS 回帰）より大きいのに、重相関係数の値が大きく表示されるということもある。この定義が妥当かどうか今後考えることとする。

3）ＰＬＳ回帰分析

PLS 回帰分析ではまず、変数の線形結合を考える。

$$r_{i\lambda} = \sum_{j=1}^{p} u_{ij}\tilde{x}_{j\lambda} \qquad (i = 1, 2, \cdots, r\,;\, r < p)$$

この式を、行列記号を用いて書くと以下となる。

$$\mathbf{R} = \mathbf{X}\mathbf{U} \qquad \mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2\ \cdots \mathbf{u}_r)$$

ここで、行列 $\mathbf{U}$ の各列ベクトルは直交し、順番に $\mathbf{X}\mathbf{u}_i$ と $\mathbf{y}$ との内積が最大化されるように選ばれる。詳細は後に示す。

この新しい変数を用いて、目的変数を以下のように予測する。

$$\tilde{y}_\lambda = \sum_{j=1}^{r} \beta_j r_{j\lambda} + \varepsilon_\lambda$$

即ち、

$$\mathbf{y} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}\mathbf{U}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

最小２乗法を使い、以下の量を最小化するようにパラメータを決定する。

$$D'' = {}^t(\mathbf{y} - \mathbf{R}\boldsymbol{\beta})(\mathbf{y} - \mathbf{R}\boldsymbol{\beta})$$

その解は次のように与えられる。

$$\boldsymbol{\beta} = ({}^t\mathbf{RR})^{-1}{}^t\mathbf{Ry} = ({}^t\mathbf{U}{}^t\mathbf{XXU})^{-1}{}^t\mathbf{U}{}^t\mathbf{Xy}$$

これから、標準化偏回帰係数$\tilde{\mathbf{b}}$ は以下となる。

$$\tilde{\mathbf{b}} = \mathbf{U}\boldsymbol{\beta}$$

また、回帰係数は、以下で与えられる。

$$b_i'' = \tilde{b}_i s_y / s_i \ , \quad b_0'' = \bar{y} - \sum_{i=1}^{p} b_i'' \bar{x}_i$$

多重共線性の改善の程度については、変数を$\mathbf{U}$ 行列で変換した後のi 番目の説明変数を、他の説明変数で予測して重相関係数r_i を求め、以下の式で定義される VIF 指標を利用している。

$$VIF_i = 1/(1-r_i^2)$$

異常度については以下で定義する。

$$a(y', \mathbf{x}') = \frac{(y' - b_0'' - \mathbf{b}''\mathbf{x}')^2}{\sigma^2}$$

ここで分散σ^2 については以下で推測する。

$$\sigma^2 = \frac{D}{N} = \frac{1}{N}{}^t(\mathbf{y} - \mathbf{Xb}'')(\mathbf{y} - \mathbf{Xb}'')$$

最後に行列$\mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2 \cdots \mathbf{u}_r)$の決定法について述べる。この行列の１列目$\mathbf{u}_1$ は$\mathbf{Xu}_1$ が最も$\mathbf{y}$ の方向に向くように、以下のように求める。

$$L_1 = {}^t\mathbf{yXu}_1 - \mu_1({}^t\mathbf{u}_1\mathbf{u}_1 - 1) \quad \rightarrow \quad \text{最大化}$$

この解は以下で与えられる。

$$\mathbf{u}_1 = {}^t\mathbf{Xy} / \left\| {}^t\mathbf{Xy} \right\|$$

次の$\mathbf{u}_2$ については、$\mathbf{X}$ から$\mathbf{u}_1$ 方向の成分を取り除き、以下のように求める。

$$L_2 = {}^t\mathbf{y}(\mathbf{X} - \mathbf{d}_1{}^t\mathbf{d}_1\mathbf{X})\mathbf{u}_2 - \mu_2({}^t\mathbf{u}_2\mathbf{u}_2 - 1) \quad \rightarrow \quad \text{最大化}$$

ここに、$\mathbf{d}_1 = \mathbf{Xu}_1 / \left\| \mathbf{Xu}_1 \right\|$ である。確かに$\mathbf{X} - \mathbf{d}_1{}^t\mathbf{d}_1\mathbf{X}$ は $\mathbf{u}_1$ 方向の成分を取り除いている。

$$(\mathbf{X} - \mathbf{d}_1{}^t\mathbf{d}_1\mathbf{X})\mathbf{u}_1 = \mathbf{Xu}_1 - \mathbf{Xu}_1{}^t(\mathbf{Xu}_1)\mathbf{Xu}_1 / \left\| \mathbf{Xu}_1 \right\|^2 = \mathbf{0}$$

この解は以下で与えられる。

$$\mathbf{u}_2 = {}^t(\mathbf{X} - \mathbf{d}_1{}^t\mathbf{d}_1\mathbf{X})\mathbf{y} / \left\| {}^t(\mathbf{X} - \mathbf{d}_1{}^t\mathbf{d}_1\mathbf{X})\mathbf{y} \right\|$$

このベクトル$\mathbf{u}_2$ は$\mathbf{u}_1$ と直交する。

$${}^t\mathbf{u}_1\mathbf{u}_2 \propto {}^t\mathbf{u}_1{}^t(\mathbf{X} - \mathbf{d}_1{}^t\mathbf{d}_1\mathbf{X})\mathbf{y} = {}^t(\mathbf{Xu}_1 - \mathbf{Xu}_1{}^t(\mathbf{Xu}_1)\mathbf{Xu}_1 / \left\| \mathbf{Xu}_1 \right\|^2)\mathbf{y} = 0$$

これを続けると、k 番目の係数ベクトル$\mathbf{u}_k$ は以下のように求められることが分かる。

$$L_k = {}^t\mathbf{y}\left(\mathbf{X}-\sum_{i=1}^{k-1}\mathbf{d}_i\,{}^t\mathbf{d}_i\mathbf{X}\right)\mathbf{u}_k-\mu_k({}^t\mathbf{u}_k\mathbf{u}_k-1)\quad\rightarrow\qquad \text{最大化}$$

$$\mathbf{u}_k = {}^t\left(\mathbf{X}-\sum_{i=1}^{k-1}\mathbf{d}_i\,{}^t\mathbf{d}_i\mathbf{X}\right)\mathbf{y}\Bigg/\left\|{}^t\left(\mathbf{X}-\sum_{i=1}^{k-1}\mathbf{d}_i\,{}^t\mathbf{d}_i\mathbf{X}\right)\mathbf{y}\right\|$$

どこまでの次元数を求めればよいかは、１つの方法として１個抜き交差検証法の重相関係数または同じことであるが、残差分散の大きさを元にして決めればよい。我々のプログラムではこの方法を用いている。

3.　プログラムの利用法

メニュー［分析－ＯＲ－品質管理－異常検知］を選択すると、図１のような分析実行メニューが表示される。

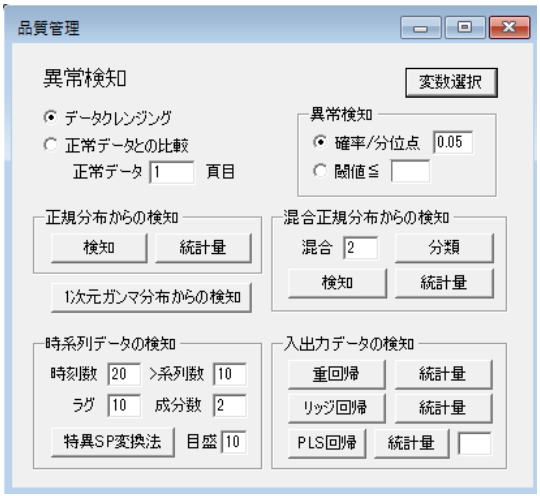


図１　分析実行メニュー

このプログラムでは、グリッドエディタに表示されているデータから、異常データを選別する「データクレンジング」機能と、別頁にある正常データを用いて、現在表示されているデータの異常データを選別する「正常データとの比較」機能がある。後者の場合は、正常データがどの頁にあるかを指定しなければならない。

分析には、(多変量)「正規分布からの検知」、(多変量)「混合正規分布からの検知」、「時系列データの検知」、「入出力データの検知」がある。最初の「正規分布からの検知」は１つの多変量正規分布からのデータのずれを検知するもので、マハラノビス距離を基にした手法である。これは分布が限定されているが、多くのデータでほぼ正規分布の仮定が成り立つと考えられるので、利用範囲は広い。「混合正規分布からの検知」は、多変量正規分布が仮定できない場合で、しかもどのような分布になっているか予想が困難な場合に適用が可能である。これは、分布を

複数の正規分布の重ね合わせとして考えるモデルで、いくつの正規分布の重ね合わせで考えると効果的かという判断も可能である。「時系列データの検知」では、最も実用的と思われる「特異 SP 変換法」（特異スペクトル変換法）が利用できる。「入出力データの検知」では、「重回帰」分析を基本として、多重共線性のある場合の手法で「リッジ回帰」分析と「PLS 回帰」分析が利用可能である。他に「１次元ガンマ分布からの検知」があるが、これは変数が１つの場合にしか適用できないので、現実には利用しにくいかも知れない。

最初に 2 変数のデータを用いて、正規分布と混合正規分布からの検知について、プログラムの説明をする。図 2 に示すファイル「異常検知 1(正規分布).txt」の 3 頁目は、2 変数で異常値を含んだデータである。

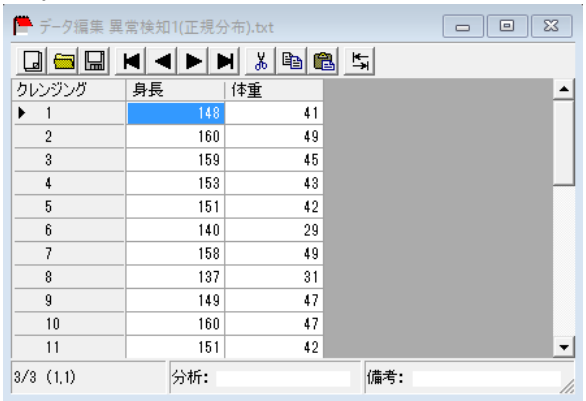


図 2　異常値を含んだデータ

この中から異常値を検出するには、「データクレンジング」ラジオボタンを選び、「確率/分位点」ラジオボタンを選んで、異常値の確率値を指定する（この場合は確率値となる）。ここでは 5%に設定している。その後、正規分布からの異常検知の「検知」ボタンをクリックすると図 3 のような出力結果を得る。

図 3　データクレンジング検知結果

出力は、このデータから求めた多変量正規分布の平均からのマハラノビス距離の 2 乗、それを元にしたホテリング統計量（F 分布の f 値）、その検定確率、異常かどうかの判定である。判定は正常と異常でそれぞれ 0 または 1 で出力される。

次に、「統計量」ボタンをクリックすると、図４のように、平均や共分散等のパラメータ推定値等と共に、異常の判定に使われるホテリング統計量の閾値が出力される。利用者はこの値を参考にして、閾値としてホテリング統計量を用いてもよい。

	身長	体重
平均	149.000	38.700
分散共分散		
身長	51.733	39.433
体重	39.433	40.343
閾値確率	0.050	
自由度	2,28	
閾値(Ht2-f値)	3.340	

図４　パラメータ推定値

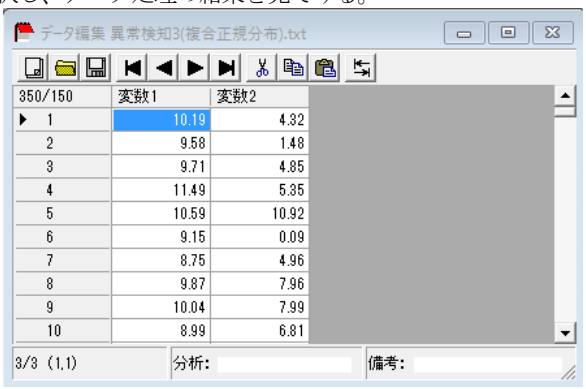
次に「正常データとの比較」ラジオボタンをクリックし、同じファイルの１頁目を開く、正常データは２頁目に入っているものとして、「正常データ」テキストボックスの中に２を入力する。「検知」ボタンをクリックすると、図５のように２頁目の正常データから求められるパラメータを元にした、異常検知結果が出力される。

	Maha2乗	異常度(f値)	確率	異常
1	2.295	1.036	0.368	0
2	0.484	0.219	0.805	0
3	10.957	4.948	0.014	1
4	0.368	0.166	0.848	0
5	2.384	1.077	0.354	0
6	0.645	0.292	0.749	0
7	2.763	1.248	0.303	0
8	2.048	0.925	0.408	0
9	2.412	1.089	0.350	0
10	6.417	2.898	0.072	0

図５　正常データとの比較検知結果

出力項目については図３と同様である。「統計量」ボタンをクリックした結果は、正常データを元にした結果であり、図４と同じ様式であるので省略する。

ファイル「異常検知 3(複合正規分布).txt[2]」を図６のように読み込み、「データクレンジング」ラジオボタンを選択し、データ処理の結果を見てみる。



350/150	変数1	変数2
1	10.19	4.32
2	9.58	1.48
3	9.71	4.85
4	11.49	5.35
5	10.59	10.92
6	9.15	0.09
7	8.75	4.96
8	9.87	7.96
9	10.04	7.99
10	8.99	6.81

図６　非正規分布のデータ

このデータは、実際には２つの正規分布を合わせたものであるが、今の段階ではそれが分からないものとする。仮に「混合」テキストボックスを２とし、処理を進める。後にこの数字を変更して最も良いモデルを選択する。

「分類」ボタンをクリックすると、図７のように、２つの群についてのデータの帰属度と分類結果が得られる。

	帰属度1	帰属度2	分類
343	0.999	0.001	1
344	1.000	0.000	1
345	0.999	0.001	1
346	0.984	0.016	1
347	0.999	0.001	1
348	0.999	0.001	1
349	0.999	0.001	1
350	1.000	0.000	1
351	0.000	1.000	2
352	0.000	1.000	2

図７　レコード毎の帰属度と分類結果

この分類結果をコピーして、元のデータに図８のように貼り付け、

350/150	変数1	変数2	
1	10.19	4.32	1
2	9.58	1.48	1
3	9.71	4.85	1
4	11.49	5.35	1
5	10.59	10.92	1
6	9.15	0.09	1
7	8.75	4.96	1
8	9.87	7.96	1
9	10.04	7.99	1
10	8.99	6.81	1

3/3 (1,3)

図８　分類項目の貼り付け

分析「相関と回帰分析」の「先頭列で群分け」ラジオボタンを選択して、図９のような散布図を描くことも可能である。

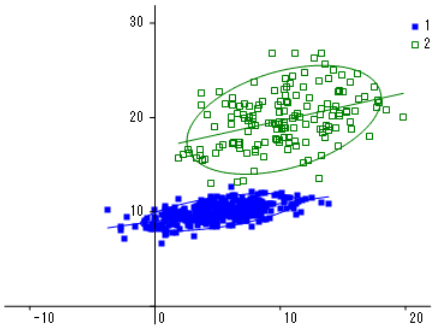


図９　データの散布図

但し、このグラフには、グラフメニュー［設定－正規楕円半径］を用いて、2σの確率楕円を加えてある。

次に「統計量」ボタンをクリックすると、図 10 のように 2 つの群の平均と共分散、群の生起確率の推測値などが表示される。この中で一番下の BIC はモデル判定によく利用されるベイズ情報量基準と呼ばれるもので、この値が小さいほど良いモデルとして評価される。

	変数1	変数2
▶ 群1		
生起確率	0.697	
平均	9.973	5.091
共分散		
変数1	0.992	1.845
変数2	1.845	9.744
群2		
生起確率	0.303	
平均	19.738	10.267
共分散		
変数1	8.540	4.666
変数2	4.666	15.133
閾値(a値)	7.305	
対数尤度	-2392.768	
AIC	4809.535	
BIC	4860.110	

図 10　2 群の場合のパラメータ推定値

現在の 2 群の場合は BIC=4860.110 であるが、3 群にすると 4888.488 となり、2 群の方が良いモデルであると判断される。これは、2 群を故意に作ったモデルであるので、当然の結果である。もちろん AIC についても小さな値が良いモデルと判定される。

また、「検知」ボタンをクリックすると、図 11 のように異常度の値と判別結果が表示される。

	異常度(a値)	異常
97	5.259	0
98	4.859	0
99	4.379	0
100	3.285	0
101	3.165	0
102	3.323	0
103	7.567	1
104	3.786	0
105	3.606	0
106	3.473	0

図 11　混合正規分布からの検知

判定には「確率/分位点」ラジオボックスの中の分位点を利用している。正規性を考えない場合は分位点で判定を与える。

1 次元ガンマ分布でも同様の結果の表示となるので、ここでは省略するが、ガンマ分布の場合は、2 つのパラメータが推測される。またこの場合、1 つのボタンで検知結果と統計量が両方表示される。

時系列データからの異常検知では、正常データを指定しないので、図 1 の分析メニューでは「データクレンジング」ラジオボタンを選んで実行する。時系列データは、図 12 のように与えられる。

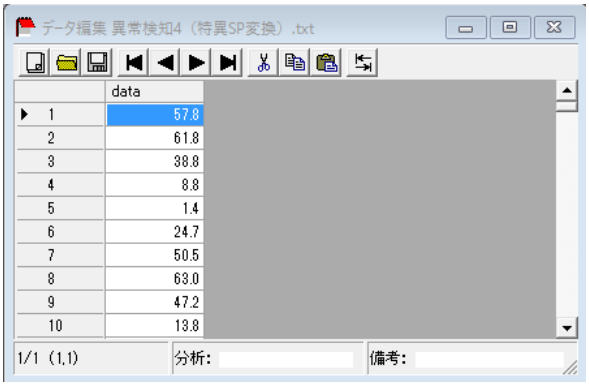


図 12　時系列データ（異常検知 4(特異 SP 変換).txt [2]）

異常検知の方法は簡単で、変数選択をして、「特異 SP 変換」ボタンをクリックする。その際、必要ならば「時刻数」、「系列数」、「ラグ」、「成分数」の値を変更する。これらの値はそれぞれ理論の説明の中の、w, k, L, m に相当する。「目盛」の値は、グラフの目盛間隔の値を与える。図 13 に実行結果を示す。

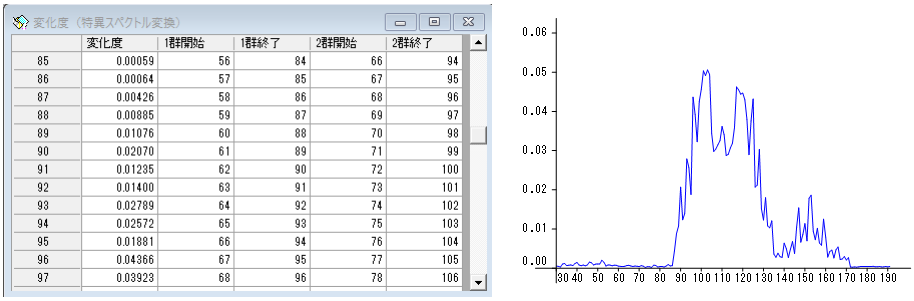


図 13　特異スペクトル変換による異常検知結果

図 13 の表の中で、左端が時刻であり、右 4 列は 2 つの群が利用したデータの範囲を示している。1 群は時刻の 1 つ前までのデータを利用し、2 群はそれから「ラグ」だけ遅れた時刻のデータを使っている。

入力と出力があるデータでは、重回帰分析が基本となる。通常の重回帰分析と同様に、目的変数、説明変数の順に変数を選んで、図 1 の分析メニューの「重回帰分析」ボタンをクリックすると図 14 のような実行結果が表示される。

異常度（重回帰分析）

	実測値	予測値	残差	異常度	判定
4	39	40.859	-1.859	0.160	0
5	81	84.237	-3.237	0.486	0
6	47	53.870	-6.870	2.192	0
7	92	91.659	0.341	0.005	0
8	75	76.076	-1.076	0.054	0
9	77	79.557	-2.557	0.304	0
10	64	61.139	2.861	0.380	0
11	76	82.910	-6.910	2.217	0
12	97	85.622	11.378	6.011	1
13	65	65.551	-0.551	0.014	0

図 14　重回帰分析実行結果（異常検知 6(入出力).txt 1 頁目）

また、隣の「統計量」ボタンをクリックすると、図 15 のような結果が表示される。

統計量（重回帰分析）

	偏回帰係数	標準化係数	VIF	残差分散	閾値	重相関R	寄与率R^2
x1	0.1490	0.0749	1.1796	21.5362	4.0705	0.943	0.889
x2	2.2233	0.0527	1.0554				
x3	2.7614	0.3522	3.9544				
x4	0.4314	0.5888	3.9670				
切片	6.8654	0.0000					

図 15　重回帰分析統計量

これは正常に重回帰分析が行われた結果である。閾値は異常度の閾値である。

データに多重共線性がある場合、例えば、図 16 は変数 x2 と x4 が殆ど同じで、1 番目のデータだけが、0.01 違っている例である。

データ編集 異常検知6(入出力).txt

完全多重共線性	y	x1	x2	x3	x4
1	83	67	3.0	7.4	3.01
2	90	71	3.7	8.0	3.7
3	80	57	3.9	6.5	3.9
4	39	43	2.8	1.8	2.8
5	81	63	3.6	6.1	3.6
6	47	51	3.7	2.7	3.7
7	92	72	4.1	7.9	4.1
8	75	62	3.8	4.6	3.8
9	77	69	3.6	5.8	3.6
10	64	59	3.6	4.2	3.6

4/5 (1,1)　分析:　備考:

図 16　強い多重共線性があるデータ（同 4 頁目）

このデータに対して重回帰分析の「統計量」ボタンをクリックすると、図 17 のような結果が得られる。

統計量（重回帰分析）

	偏回帰係数	標準化係数	VIF	残差分散	閾値	重相関R	寄与率R^2
x1	0.2251	0.1132	1.1862	38.0310	3.3607	0.8967	0.8041
x2	521.1502	12.3456	61886.5468				
x3	6.7407	0.8596	1.1435				
x4	-519.5957	-12.2978	61906.7203				
切片	18.5670	0.0000					

図 17　強い共線性があるデータの重回帰分析統計量

この結果ではほぼ同じ値である変数 x2 と x4 の偏回帰係数が極端に大きくなっている。これは予測において、x2 と x4 の値が少しずれると大きな差となって表れることを示しており、予測の頑健性において問題となる。

これを解決するための代表的な手法に、リッジ回帰分析と PLS 回帰分析がある。理論のところで示したように、リッジ回帰分析は多重共線性の元となる説明変数の共分散行列の対角成分に、ある定数（ここでは $\eta \div N$）を加える方法である。この定数 η は、1 個抜き交差検証の残差分散が最小となるように決める。

図 16 のデータに対して、「リッジ回帰分析」とその「統計量」を与えた結果を図 18 と図 19 に示す。

異常度（リッジ回帰分析）

	実測値	予測値	残差	異常度	判定
4	39	45.709	-6.709	1.005	0
5	81	78.913	2.087	0.097	0
6	47	54.497	-7.497	1.255	0
7	92	93.169	-1.169	0.030	0
8	75	69.391	5.609	0.702	0
9	77	78.488	-1.488	0.049	0
10	64	65.867	-1.867	0.078	0
11	76	78.259	-2.259	0.114	0
12	97	79.665	17.335	6.708	1
13	65	76.217	-11.217	2.809	0

図 18　リッジ回帰分析による異常検知

統計量（リッジ回帰分析）

	偏回帰係数	標準化係数	残差分散	閾値	重相関R	寄与率R^2
x1	0.2466	0.1240	44.7968	2.8086	0.8943	0.7998
x2	0.6154	0.0146			最良η	7.2000
x3	6.3469	0.8094				
x4	0.6093	0.0144				
切片	20.2521	0.0000				

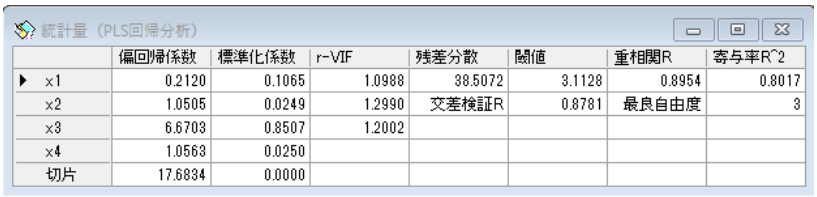
図 19　リッジ回帰分析統計量

また、PLS 回帰分析は多重共線性に対して、独立成分として与える変数の数を減らす方法で対応する。どれだけの変数を減らすかは、リッジ回帰分析と同様に 1 個抜き交差検証の残差分散が最小となるように決める。図 20 と図 21 にメニューの指定で独立成分を 3 にした PLS 回帰分析を実行した結果を示す。

異常度（PLS回帰分析）

	実測値	予測値	残差	異常度	判定
4	39	44.704	-5.704	0.845	0
5	81	79.311	1.689	0.074	0
6	47	54.299	-7.299	1.383	0
7	92	94.278	-2.278	0.135	0
8	75	69.515	5.485	0.781	0
9	77	78.581	-1.581	0.065	0
10	64	65.789	-1.789	0.083	0
11	76	78.611	-2.611	0.177	0
12	97	80.817	16.183	6.801	1
13	65	75.948	-10.948	3.113	0

図 20　PLS 回帰分析による異常検知


統計量（PLS回帰分析）

	偏回帰係数	標準化係数	r-VIF	残差分散	閾値	重相関R	寄与率R^2
x1	0.2120	0.1065	1.0988	38.5072	3.1128	0.8954	0.8017
x2	1.0505	0.0249	1.2990	交差検証R		0.8781	最良自由度	3
x3	6.6703	0.8507	1.2002				
x4	1.0563	0.0250					
切片	17.6834	0.0000					

図 21　PLS 回帰分析統計量

「最良自由度」は、目的変数の最適な自由度である。リッジ回帰分析、PLS 回帰分析とも重回帰分析で問題となった変数 x2 と x4 の偏回帰係数が小さくなっている。これにより、これらの変数のずれの影響は小さく抑えられる。

4.　おわりに

我々は、多変量正規分布に従うデータ、多変量正規分布しないデータ、時系列データ、入力と出力があるデータに関する異常検知のプログラムを作成した。多変量正規分布するデータについてはマハラノビスの距離を元にしたホテリング t^2 統計量を利用して異常検知を行っている。この方法では F 分布の確率値を用いて閾値を指定することができる。多変量正規分布しないデータについては多変量正規分布を確率的に重ね合わせてデータ分布を近似する。分布の重ね合わせには EM アルゴリズムという手法を用いている。

時系列データの異常検知については、時間をずらせて連続データを 2 本取り、その間の違いを定量化することによって異常を調べる特異スペクトル変換法という手法を用いている。入出力のあるデータについては、通常は重回帰分析を用いて予測値のずれの異常を調べるが、データの中に多重共線性が認められる場合も考えられる。その際には、リッジ回帰分析や、PLS 回帰分析を用いて多重共線性の影響を押さえるようにする。リッジ回帰分析は、多重共線性により正則性が問題となる行列の対角成分を変更し、正則性を保証するようにする手法であり、PLS 回帰分析は説明変数の自由度を減らして多重共線性を防ぐ手法である。

現在、これらの異常検知のプログラムは、データと独立に単独で与えられているが、本来はラインの中に組み込み、自動的に異常を検知できるようにして、力を発揮するものである。このプログラムの中から必要な部分が取り出され、実際にライン上で動作するように作り変えられれば大変興味深い。

参考文献

[1] 井出剛, 入門機械学習による異常検知, コロナ社, 2015.

[2] ホームページ http://www.heisei-u.ac.jp/ba/fukui/analysis.html 内のサンプルデータ Samples.zip 内のファイル

社会システム分析のための統合化プログラム３０
－異常検知－

Multi-purpose Program for Social System Analysis 30
- Anomaly Detection -

Masayasu FUKUI[*1], Tomoyuki OYAMA[*2] and Nozomu ODA[*2]

[*1] *Department of Business Administration, Faculty of Business Administration, Fukuyama Heisei University*
[*2] *Nippon Pneumatic Manufacturing Co., Ltd.*

Abstract: We have been constructing a unified program on the social system analysis for the purpose of education. This time, we make programs on anomaly detection in the field of quality control. We explain methods of anomaly detection for multivariate normally distributed data, non-normally distributed data, time series data and input output data, and introduce operation of our programs.

Keywords: College Analysis, management science, quality control, anomaly detection, machine learning, EM algorithm
URL: http://www.heisei-u.ac.jp/ba/fukui/