

## 主成分回帰分析を用いる改良形 GMDH\*

近 藤 正\*\*

### Revised GMDH Algorithm Using Principal Component-Regression Analysis\*

Tadashi KONDO\*\*

In this paper, a revised GMDH (Group Method of Data Handling) of multiinput-singleoutput type algorithm is developed by using principal component-regression analysis.

In previous GMDH algorithms, the optimal partial polynomials, which are generated by using the second degree polynomial of the two variables in each selection layer, are accumulated in multilayerd structure to construct the complete polynomial. But, in a high order selection layer, many combinations of two variables generate multicollinearity in partial polynomials and this presents a severe problem in estimation accuracy of model parameters. The nonlinear models which contain multicollinearity in the partial polynomials lack in stability and can not be used in prediction problems.

The revised GMDH in this paper generates some optimal partial polynomials which are identified by using principal component-regression analysis in each selection layer, and the complete description of the system is constructed by combining these optimal partial polynomials in multilayerd structure. The principal components which construct the partial polynomials are perendicular each other and so the partial polynomials generate no multicollinearity.

The revised GMDH algorithm is applied to a simple illustrative example and compared with the result obtained by the previous GMDH algorithm.

### 1. まえがき

システムの中に非常に多くの変数やパラメータを含み、しかもそれらが相互に結合して複雑な現象を示すために、従来の専門科学のみの知識を用いて対象システムのモデリングを行うことが困難な非線形システムや、システム構造に関する先駆的な情報が得られていない実システムなどに対して、少ない入出力データから、システムに関する先駆的な知識を必要とすることなく、発見的自己組織化の原理に基づいてモデリングを行う手法としてGMDH(Group Method of Data Handling)<sup>1)~6)</sup>がある。GMDHは階層型ニューラルネットワークと同じく、多層構造と並列処理を特徴としたアルゴリズムをしているが、ニューラルネットワークと比較してネットワーク内部の情報処理方法は全く異なっており、比較的小規模

なネットワークを自己組織化することにより、非線形システムの高精度なモデリングが可能である<sup>7)</sup>。GMDHのアルゴリズムでは、各選択層において、システムの部分表現式あるいはシステムの中間表現式を、少数の入出力データを用いて同定し、これを多層構造に積み重ねることによりシステムの完全表現式を構成している。しかしながら、多層構造に層を積み重ねて行くと、多重共線性を起こす組み合わせが数多く発生して、モデル係数の推定精度が著しく悪化し予測精度が悪くなる。多重共線性を防ぎ予測精度の良いモデルを同定する方法として、(1) 情報量規準 AIC あるいは予測平方和 PSS などの評価規準を用いて変数の逐次選択を行い、多重共線性を起こさない最適部分表現式を自己選択するアルゴリズム<sup>3)</sup> (2) Ridge 回帰を用いて部分表現式のパラメータを推定し多重共線性を防止するアルゴリズム<sup>4)</sup>

などが提案されている。(1)の方法では、多重共線性を起こす変数を同時に部分表現式の中に取り込むことを避けて、最適部分表現式を構成する際に多重共線性を起こさないモデルを構成している。(2)の方法では、多重共線性が存在しても安定な Ridge 回帰係数を持つ部分表現式

\* 原稿受付 1991. 12. 27.

\*\* 徳島大学 医療技術短期大学部 School of Medical Sciences, The University of Tokushima : 3 chome 18-15, Kuramoto-cho, Tokushima city, Tokushima 770, JAPAN

**Key Words :** GMDH, nonlinear system identification, modeling, principal component-regression analysis, network.

を構成出来るが、Ridge回帰係数で用いるRidgeパラメータの最適値は相違なる2変数の組み合わせごとに異なり、さらにこの値は各選択層において変化するため、各々の相違なる2変数の組み合わせごとにRidge曲線などを用いて最適値を見つけだすことは實際上困難である。このため適當なRidgeパラメータを仮定してRidge回帰係数を計算することになるが、この場合、得られた回帰係数に含まれる偏りが大きくなり良い予測精度を得ることができなくなる。また、これらの多重共線性の発生を防ぐアルゴリズムとは考え方方が異なるが、Ivakhnenkoらのグループにより、逐次変数導入型GMDHアルゴリズム<sup>5)</sup>や一般化GMDHアルゴリズム<sup>6)</sup>などの研究が行われているのでこれらの手法についてふれておく。逐次変数導入型GMDHアルゴリズムは、まず、出入力データの前処理として、直交化された入力変数( $x_1, x_2, \dots, x_s$ )を作成する。そして、第1層では、出力( $\phi$ )と最も寄与の大きい変数( $x_i$ )との部分多項式( $f_1(x_i)$ )を作成し、第2層では、残差( $\phi - f_1(x_i)$ )を出力とする部分多項式( $f_2(x_j)$ )を作成する。この操作を繰り返して多層構造を構成し、残差の関数( $f_i$ )から構成される完全記述式( $\phi = f_1(x_i) + f_2(x_j) + \dots + f_s(x_k)$ )を作成するアルゴリズムである。一般化GMDHアルゴリズムは、逐次変数導入型GMDHアルゴリズムを拡張したアルゴリズムになっている。これらのアルゴリズムは、モデル式の中に入力変数間の相互結合ができるかぎり含まないようにしておらず、工業プロセスを対象にした制御用モデル式の作成を主要な目的にしている。このため、本論文で対象にしているようなシステムの入出力間に変数の相互結合を多く含むような、本質的に複雑な構造をした非線形システムのモデリングに適用することは困難である。

本論文では、複雑な構造をした非線形システムを対象として、主成分回帰分析を用いて各選択層において直交化した変数を発生させ、これらの変数を用いて多重共線性を起こさない最適な部分表現式を作成する。そして、このようにして作成した最適部分表現式を多層に積み重ねてシステムの完全表現式を作成する改良形GMDHのアルゴリズムを提案する。つぎに、この改良形GMDHを簡単な同定問題に適用し、従来から提案されている評価規準を用いて変数の逐次選択を行い最適な部分表現式を自己選択するGMDHとの比較を行い、その予測精度が大幅に改善されることを示す。

## 2. 評価規準を用いて変数の逐次選択を行い最適な部分表現式を自己選択するGMDH<sup>3)</sup>

まず始めに、評価規準を用いて変数の逐次選択を行い

最適な部分表現式を自己選択するGMDHのアルゴリズムについて簡単に述べる。

システムの完全表現式として、Kolmogorov-Gavorの多項式

$$\Phi = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \dots \quad (1)$$

を想定する。また、システムの部分表現式は2変数の2次多項式

$$y_k = b_{0k} + b_{1k} x_i + b_{2k} x_j + b_{3k} x_i x_j + b_{4k} x_i^2 + b_{5k} x_j^2 \quad (2)$$

に対して変数選択の評価規準(情報量規準 AIC<sup>8)</sup>または予測平方和 PSS<sup>9)</sup>を用いて変数選択を行い、多重共線性を起こさない最適な部分表現式を自己選択する。変数の逐次選択法として変数増減法を用いる。

情報量規準 AIC は、

$$AIC = n \log_e S_m^2 + 2(m+1) + C \quad (3)$$

$$S_m^2 = \frac{1}{n \sum_{\alpha=1}^n (\Phi_{\alpha} - y_{\alpha})^2} \quad (4)$$

によって与えられる。ここで、 $n$  はデータの個数、 $m$  は変数の個数。 $C$  は  $m$  に無関係な定数、 $S_m^2$  は(4)式から計算される残差の自乗平均である。予測平方和 PSS は、

$$PSS = \sum_{\alpha=1}^n \left\{ \frac{\Phi_{\alpha} - y_{\alpha}}{1 - \mathbf{x}_{\alpha}^T (\mathbf{X}^T \cdot \mathbf{X})^{-1} \mathbf{x}_{\alpha}} \right\}^2 \quad (5)$$

によって与えられる。ただし、

$$\mathbf{x}_{\alpha}^T = (1, x_{i\alpha}, x_{j\alpha}, x_{i\alpha} \cdot x_{j\alpha}, x_{i\alpha}^2, x_{j\alpha}^2) \\ \alpha = 1, 2, \dots, n$$

$$\mathbf{X}^T = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$$

ここで、 $\Phi_{\alpha}$  は  $\alpha$  番目のデータの出力値、 $y_{\alpha}$  は中間変数  $y_k$  の  $\alpha$  番目のデータに対応する値である。システムの完全表現式(1)式は、最適な部分表現式を積み重ねて構成し、その次数は各選択層において変数を組み合わせることにより増加する。層を通過して計算を続けて行くと多重共線性を起こす変数の組み合わせが増加して最適な部分表現式に含まれる変数の数が減少する。最終層では、すべての2変数の組み合わせについて最適な部分表現式の形が、

$$y_k = b_{0k} + b_{1k} x_i \quad (6)$$

となり、前層と同じ中間変数を発生し、予測精度が改善されなくなるため部分表現式の積み重ねを停止する。

以上のアルゴリズムによって評価規準を用いて変数の逐次選択を行い最適な部分表現式を自己選択する

GMDH を構成できる。

本研究では、後で述べる主成分回帰分析を用いる改良形 GMDH との比較を行うために、(2) 式の部分表現式を計算する前に入出力データの前処理を行い、部分表現式を構成することにする。これは、(2) 式の代わりに、まず、

$$x'_i = \frac{x_i - \bar{x}_i}{s_{x_i}} \quad (7a)$$

$$x'_j = \frac{x_j - \bar{x}_j}{s_{x_j}} \quad (7b)$$

$$x_i \cdot x_j' = \frac{x_i \cdot x_j - \bar{x}_i \cdot \bar{x}_j}{s_{x_i} s_{x_j}} \quad (7c)$$

$$x_i^{2'} = \frac{x_i^2 - \bar{x}_i^2}{s_{x_i}^2} \quad (7d)$$

$$x_j^{2'} = \frac{x_j^2 - \bar{x}_j^2}{s_{x_j}^2} \quad (7e)$$

$$y_k' = \frac{y_k - \bar{y}_k}{s_{y_k}} \quad (7f)$$

のように、変数  $(x_i, x_j, x_i \cdot x_j, x_i^2, x_j^2, y_k)$  から変数  $(x'_i, x'_j, x_i \cdot x_j', x_i^{2'}, x_j^{2'}, y_k')$  への変換（規準化）を行う。ここで、 $(\bar{x}_i, \bar{x}_j, \bar{x}_i \cdot x_j, \bar{x}_i^2, \bar{x}_j^2, \bar{y}_k)$  はそれぞれの変数の平均値（期待値）、 $(s_{x_i}, s_{x_j}, s_{x_i \cdot x_j}, s_{x_i^2}, s_{x_j^2}, s_{y_k})$  はそれぞれの変数の標準偏差を示す。この変換により、各々の変数は、平均 0、分散 1 に規準化される。次に、規準化された変数を用いて部分表現式

$$y_k = b_{1k} x'_i + b_{2k} x'_j + b_{3k} x_i \cdot x_j' + b_{4k} x_i^{2'} + b_{5k} x_j^{2'} \quad (8)$$

を作成し、この式に対して変数選択の評価規準を用いて変数選択を行い、多重共線性を起こさない部分表現式を自己選択する。この GMDH を、以下において従来形 GMDH と呼ぶことにする。

### 3. 主成分回帰分析を用いる GMDH

変数の逐次選択を行い最適な部分表現式を自己選択する従来形 GMDH においては AIC または PSS などの評価規準を用いて予測誤差を評価し、多重共線性を起こす変数などを部分表現式の中に同時に取り込むことを避けて多重共線性を起こさない部分表現式を作成した。ここでは、(2) に含まれている 5 個の変数に主成分分析を通して直交化した新しい変数  $z_1 \sim z_5$  を作成し、これらの変数を用いて直交回帰分析により部分表現式を作成する。このようにして作成した部分表現式は、入力変数  $z_1 \sim z_5$  が直交化されているため多重共線性を起こさない。

この改良形 GMDH の構造を Fig. 1 に示す。以下に、主成分回帰分析を用いる GMDH のアルゴリズムを示す。

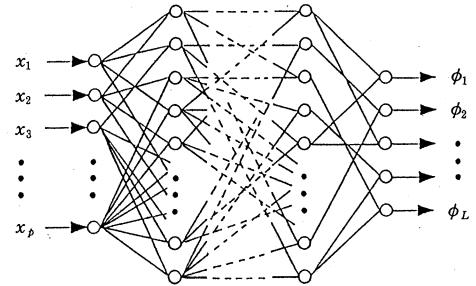


Fig. 1 Structure of the revised GMDH

### 3.1 主成分回帰分析による最適部分表現式の作成

はじめに、2 変数の 2 次多項式 (2) 式に含まれる 5 個の変数  $x_{ij} = (x_i, x_j, x_i \cdot x_j, x_i^2, x_j^2)$  を、(7) 式のように、平均 0、分散 1 に規準化する。規準化した変数  $(x'_i, x'_j, x_i \cdot x_j', x_i^{2'}, x_j^{2'})$  を  $x_{ij}'$  と表わす。つぎに、規準化した変数  $x_{ij}'$  を、

$$z = C \cdot x_{ij}' = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{15} \\ c_{21} & c_{22} & \cdots & c_{25} \\ \vdots & \vdots & & \vdots \\ c_{51} & c_{52} & \cdots & c_{55} \end{bmatrix} \begin{bmatrix} x'_i \\ x'_j \\ x_i \cdot x_j' \\ x_i^{2'} \\ x_j^{2'} \end{bmatrix} \quad (9)$$

のように直交変換し、新しい変数  $z^T = (z_1, z_2, \dots, z_5)$  を作成する。ここで、 $C$  は  $(5 \times 5)$  の正規直交行列を示す。正規直交行列  $C$  は  $x_{ij}'$  の各変数の分散が 1 になるように規準化した  $x_{ij}'$  から構成する相関行列  $R$  の固有値問題

$$R \cdot C = C \cdot A \quad (10)$$

を解くことにより求める。ここで  $R$  は  $(5 \times 5)$  の相関行列、 $A$  は固有値  $(\lambda_1, \lambda_2, \dots, \lambda_5)$  を対角要素を持つ  $(5 \times 5)$  の対角行列である。つぎに、(9) 式より作成した新しい変数  $z$  を入力変数として直交回帰分析を行い多重共線性を起こさない最適部分表現式

$$y_k = z^T \cdot d_k = [z_1, z_2, \dots, z_5] \begin{bmatrix} d_{k1} \\ d_{k2} \\ \vdots \\ d_{k5} \end{bmatrix} \quad (11)$$

を作成する。ここで、 $y_k$  は  $k$  番目の中間変数を示し、 $d_k$  は  $y_k$  に対応する係数ベクトルを示す。係数ベクトル  $d_k$

は、入力変数  $z$  が直交化されているため以下の簡単な計算により求めることができる。

(11) 式に対応する正規方程式は、

$$Z^T \cdot y_k = (Z^T \cdot Z) \cdot d_k \quad (12)$$

である。ここで、 $n$  をデータの個数とすると  $y_k^T = (y_{k1}, y_{k2}, \dots, y_{kn})$ 、 $Z$  は  $n$  個の  $z^T$  から構成されるデータ行列である。(12) 式を各要素を用いて書き直すと、

$$\begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_n^T \end{bmatrix}^T \begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kn} \end{bmatrix} = \begin{bmatrix} r_1 & 0 \\ r_2 & \ddots \\ 0 & r_5 \end{bmatrix} \begin{bmatrix} d_{k1} \\ d_{k2} \\ \vdots \\ d_{k5} \end{bmatrix} \quad (13)$$

となる。入力変数  $z$  が直交化されているため(12)式の行列  $(Z^T \cdot Z)$  は対角行列となり、係数ベクトル  $d_k$  の  $i$  番目の成分  $d_{ki}$  ( $i = 1, 2, \dots, 5$ ) は、

$$d_{ki} = \frac{(Z^T \cdot y_k)_i}{r_i}, \quad i = 1, 2, \dots, 5 \quad (14)$$

により求まる。ここで  $r_i$  は対角行列  $(Z^T \cdot Z)$  の第  $(i, i)$  成分である。(11)式の最適部分表現式を作成する時に、有用な変数のみを自己選択するために情報量規準 AIC を用いて変数の逐次選択を行う。変数  $z$  が直交化されて無相関であるために変数の逐次選択方法(変数増減法、変数減増法など)としては、どの方法を採用しても作成されるモデルは同じになる。AIC の計算に用いる残差の自乗平均は、

$$S_m^2 = S_{m-1}^2 - \frac{d_{ki}}{n}, \quad i = 1, 2, \dots, 5 \quad (15)$$

$$S_0^2 = \frac{y_k^T \cdot y_k}{n} \quad (16)$$

により求める。ここで(15)式の第2項は、変数  $z_i$  を部分表現式の中に取り込んだ場合の自乗平均の減少量を示す。(15)式と(16)式により  $z_1 \sim z_5$  を部分表現式の中に取り込んだ場合の残差自乗和を計算して、残差自乗和を最も小さくする変数から部分表現式の中に順次取り込む。そして、(15)式と(3)式から計算される AIC の値が増加する時に変数の取り込みを停止する。このようにして、選択した有用な変数のみを用いて(11)式の最適部分表現式を作成する。

### 3.2 中間変数の自己選択

最適な部分表現式によって発生される中間変数に対して AIC の値の小さいものを  $L$  個自己選択する。選択個数はなるべく大きくとることが望ましい。

### 3.3 多層構造の計算停止方法

GMDH では層を通過して計算を続けていくと多重共線性を起こす変数の組み合わせが数多く発生するため、主成分分析を行う時に変数の次元が縮小する。変数の次元は 5 次から 2 次に縮小して、最後に固有値の絶対値の大きいものが二つ残り、ほかの三つの固有値はすべてゼロに近づく。多重共線性を起こす中間変数を組み合わせても予測精度の改善が行われないため、多重共線性の発生により変数の次元が縮小する層で多層構造の層の積み重ねを打ち切り、計算を停止する。

変数の次元が縮小する層は固有値を用いて以下のようにして判定する。各選択層のすべての中間変数に対して、

$$\frac{\lambda_{k, \max_1} + \lambda_{k, \max_2}}{\sum_{i=1}^5 \lambda_{k, i}} > E \quad (17)$$

が満足された時に変数の次元が縮小したと見なし多層構造の積み重ねを打ち切る。ここで  $\lambda_{k, \max_1}, \lambda_{k, \max_2}$  は第  $k$  番目の中間変数に対する固有ベクトル  $\lambda_k = [\lambda_{k,1}, \lambda_{k,2}, \lambda_{k,3}, \lambda_{k,4}, \lambda_{k,5}]^T$ , ( $k = 1 \sim L$ ) の 5 個の要素の中で 1 番目と 2 番目に大きな値をとる固有値を示す。 $L$  は中間変数の選択個数を示す。 $E$  は多層構造の打ち切り判定規準値を示す。そして、同定されたモデルとして、前層で作成された中間変数の中で最小の AIC 値を示す変数を採用する。これは、(17) 式に含まれる固有値は、前層で作成された中間変数の分散を示しており、このため(17)式では、前層で作成された中間変数の次元が縮小しているかどうかを判定しているためである。以下では、(17)式を多層構造の計算停止方法(1)と呼ぶことにする。また、別の多層構造の停止方法として AIC の値を用いることも出来る。AIC の値は、次元が縮小する層より後の層では予測精度が改善されないために、徐々に増加する。このため、AIC の値が最小値をとる層で計算を停止する方法が考えられる。これを多層構造の計算停止方法(2)と呼ぶこととする。

以上のアルゴリズムによって、主成分回帰分析を用いる改良形 GMDH を構成できる。

### 4. 簡単な同定問題への適用

本論文で提案した改良形 GMDH と従来の GMDH の比較を行うために、簡単な同定問題への適用を行う。対象システムの入出力関係式は次式で表わされるものとする。

$$\Phi = (1.0 + 1.1x_1 + 1.2x_2 + 1.3x_3)^4 + \varepsilon \quad (18)$$

ここで、 $\varepsilon$  は  $N(0, 25^2)$  の正規性白色雑音を示す。

GMDH の入力変数として,  $x_1 \sim x_4$  の 4 变数を用いる。

(18) 式は、入力変数  $x_1 \sim x_3$  により構成されており、 $x_4$  は含んでいない。入力変数に不必要的変数  $x_4$  が含まれている場合においても、システムの同定が可能であること示すために入力変数に  $x_4$  をつけ加える。同定に用いる入出力データとして 20 個のデータを用い、同定されたモデルの評価を行うために別の 20 個の入出力データを用いる。本研究において、数値実験例として (18) 式の入出力関係式を用いたのは、GMDH は一般的に入力変数が相互に結合して構造が複雑で強い非線形性を有するシステムの同定問題に適用される場合が多いために、このような特徴を持つシステムの一例として (18) 式を用いた。(18) 式は、入力変数  $x_1 \sim x_3$  に関して 1 次から 4 次までの項をすべて含み、しかも、入力変数が相互に結合して比較的複雑な入出力関係を構成している。まず始めに改良形 GMDH による同定結果を示す。

#### 4.1 改良形 GMDH による同定結果

##### (1) 中間変数

AIC 値の小さいものを 4 個選択する。

##### (2) 多層構造の計算停止方法

ここでは、多層構造の計算停止方法として計算停止方法 (1) を採用する。多層構造の計算停止規準値は、 $E = 0.99$  とする。

##### (3) 同定されたモデル

第 5 層において計算が終了して次式が求まった。

(第 4 層)

$$\hat{\Phi} = \mathbf{z}_1^{(4)T} \cdot \mathbf{d}_1^{(4)} \quad (19a)$$

$$\mathbf{z}_1^{(4)} = C_1^{(4)} \cdot \mathbf{y}_{12}^{(3)} \quad (19b)$$

(第 3 層)

$$\mathbf{y}_1^{(3)} = \mathbf{z}_1^{(3)T} \cdot \mathbf{d}_1^{(3)} \quad (19c)$$

$$\mathbf{z}_1^{(3)} = C_1^{(3)} \cdot \mathbf{y}_{12}^{(2)} \quad (19d)$$

$$\mathbf{y}_2^{(3)} = \mathbf{z}_2^{(3)T} \cdot \mathbf{d}_2^{(3)} \quad (19e)$$

$$\mathbf{z}_2^{(3)} = C_2^{(3)} \cdot \mathbf{y}_{14}^{(2)} \quad (19f)$$

(第 2 層)

$$\mathbf{y}_1^{(2)} = \mathbf{z}_1^{(2)T} \cdot \mathbf{d}_1^{(2)} \quad (19g)$$

$$\mathbf{z}_1^{(2)} = C_1^{(2)} \cdot \mathbf{y}_{24}^{(1)} \quad (19h)$$

$$\mathbf{y}_2^{(2)} = \mathbf{z}_2^{(2)T} \cdot \mathbf{d}_2^{(2)} \quad (19i)$$

$$\mathbf{z}_2^{(2)} = C_2^{(2)} \cdot \mathbf{y}_{34}^{(1)} \quad (19j)$$

$$\mathbf{y}_4^{(2)} = \mathbf{z}_4^{(2)T} \cdot \mathbf{d}_4^{(2)} \quad (19k)$$

$$\mathbf{z}_4^{(2)} = C_4^{(2)} \cdot \mathbf{y}_{14}^{(1)} \quad (19l)$$

(第 1 層)

$$\mathbf{y}_1^{(1)} = \mathbf{z}_1^{(1)T} \cdot \mathbf{d}_1^{(1)} \quad (19m)$$

$$\mathbf{z}_1^{(1)} = C_1^{(1)} \cdot \mathbf{x}_{12}' \quad (19n)$$

$$\mathbf{y}_2^{(1)} = \mathbf{z}_2^{(1)T} \cdot \mathbf{d}_2^{(1)} \quad (19o)$$

$$\mathbf{z}_2^{(1)} = C_2^{(1)} \cdot \mathbf{x}_{13}' \quad (19p)$$

$$\mathbf{y}_3^{(1)} = \mathbf{z}_3^{(1)T} \cdot \mathbf{d}_3^{(1)} \quad (19q)$$

$$\mathbf{z}_3^{(1)} = C_3^{(1)} \cdot \mathbf{x}_{14}' \quad (19r)$$

$$\mathbf{y}_4^{(1)} = \mathbf{z}_4^{(1)T} \cdot \mathbf{d}_4^{(1)} \quad (19s)$$

$$\mathbf{z}_4^{(1)} = C_4^{(1)} \cdot \mathbf{x}_{23}' \quad (19t)$$

ここで、 $\mathbf{y}_i^{(M)}$  は第  $M$  層の  $i$  番目の中間変数、 $\mathbf{z}_i^{(M)}$  は第  $M$  層の  $i$  番目の非線形主成分、 $\mathbf{d}_i^{(M)}$  は第  $M$  層の  $i$  番目の最適部分表現式の係数ベクトル。 $\mathbf{y}_{ij}^{(M)}$  は第  $M$  層の中間変数  $y_j^{(M)}$  と  $y_i^{(M)}$  を規準化して作成したベクトル。 $(y_i^{(M)}, y_j^{(M)}, y_i^{(M)} \cdot y_j^{(M)}, y_i^{(M)2}, y_j^{(M)2})^T$ 、 $\mathbf{x}_{ij}'$  は入力変数  $x_i'$  と  $x_j'$  を規準化して作成した入力ベクトル。 $(x_i', x_j', x_i \cdot x_j', x_i^2, x_j^2)^T$ 、 $C_i^{(M)}$  は第  $M$  層の非線形主成分の係数行列を示す。同定した  $\mathbf{d}_i^{(M)}$  および  $C_i^{(M)}$  の値を Table 1 に示す。

同定されたモデル (19) 式では、入力変数を規準化して作成した入力ベクトルとして、 $\mathbf{x}_{12}', \mathbf{x}_{13}', \mathbf{x}_{14}', \mathbf{x}_{23}'$  の 4 個の入力ベクトルが自己選択されている。その他の入力ベクトル  $\mathbf{x}_{24}', \mathbf{x}_{34}'$  は自己選択されなかった。自己選択された入力ベクトルの中には  $\mathbf{x}_{14}'$  が含まれており、このベクトルの成分としてシステムの同定に不必要的入力変数  $x_4$  がある。しかしながら、Table 1 のモデル係数を参照すると、 $\mathbf{d}_3^{(1)}$  および  $C_3^{(1)}$  の係数値から、入力変数  $x_4$  の影響が中間変数  $y_3^{(1)}$  に対して非常に小さくなっていることがわかる。このように、本来、システムの同定に不必要的入力変数を含む入力ベクトルは、自己選択されないように計算が行われるが、もし、一部の不必要的変数を含む入力ベクトルが自己選択された場合においても、その影響は小さくなるものと思われる。

##### (4) 各選択層における固有値の変動と AIC 値の変化

Table 2 に、第 5 層まで計算をした時の各選択層における非線形主成分の固有値を示す。大きな値を持つ 2 個の固有値に下線をつけ、その累積寄与率を右側の列に示す。累積寄与率は第 2 層で一度小さくなり、第 3 層で再び大きくなつた。そして、第 5 層ではすべての中間変数に対して 0.99 以上の値になつた。この累積寄与率の通過層に対する変化を Fig. 2 に示す。第 5 層では 4 個の中間変数とも累積寄与率はほぼ同じ値になり、GMDH の計算が第 4 層の計算においてほぼ収束していることがわ

Table 1 Parameters of the identified model by the revised GMDH

The 4-th layer	$d_1^{(4)}$	0.582	0.000	0.000	0.000	0.402
	$C_1^{(4)}$	0.582	0.582	-0.328	-0.326	-0.329
		-0.690	0.689	-0.004	0.158	-0.153
		-0.003	0.003	0.816	-0.403	-0.413
		0.154	-0.157	-0.004	0.693	-0.687
The 3-rd layer	$d_1^{(3)}$	0.402	0.401	0.475	0.475	0.475
	$C_1^{(3)}$	0.564	-0.167	0.413	0.049	0.000
		0.548	0.579	-0.345	-0.300	-0.339
		-0.689	0.679	-0.035	0.200	-0.164
		0.401	0.403	0.487	0.473	0.466
The 2-nd layer	$d_2^{(3)}$	0.164	-0.201	0.004	0.691	-0.675
	$C_2^{(3)}$	0.023	-0.035	-0.802	0.410	0.432
	$d_1^{(2)}$	0.570	-0.245	0.416	0.087	0.000
	$C_1^{(2)}$	0.590	0.565	-0.358	-0.296	-0.341
		-0.676	0.680	0.013	0.184	-0.217
The 1-st layer	$d_2^{(2)}$	0.402	0.413	0.467	0.465	0.463
	$C_2^{(2)}$	0.178	-0.220	-0.041	0.717	-0.636
	$d_4^{(2)}$	-0.040	0.020	-0.795	0.384	0.467
	$C_4^{(2)}$	0.377	0.363	0.455	0.506	-0.427
	$d_4^{(1)}$	0.770	-0.226	-0.011	-0.519	0.294
The 1-st layer	$C_1^{(1)}$	0.014	0.517	-0.356	0.225	0.745
	$d_2^{(1)}$	-0.104	0.595	0.685	-0.406	0.039
	$C_2^{(1)}$	0.518	0.049	0.464	0.716	-0.038
	$d_3^{(1)}$	-0.357	-0.570	0.435	0.047	0.596
	$C_3^{(1)}$	0.422	0.174	0.428	-0.070	-0.569
The 1-st layer	$d_1^{(1)}$	0.709	0.064	-0.011	-0.681	0.171
	$C_1^{(1)}$	-0.172	0.694	-0.116	0.061	0.687
	$d_2^{(1)}$	0.482	-0.059	0.639	0.547	0.239
	$C_2^{(1)}$	0.367	-0.286	-0.747	0.432	0.217
	$d_3^{(1)}$	-0.316	-0.655	0.145	-0.235	0.628
The 1-st layer	$d_4^{(1)}$	0.609	-0.125	-0.257	0.000	0.534
	$C_4^{(1)}$	0.631	0.386	0.323	0.590	-0.031
	$d_1^{(1)}$	-0.260	0.748	0.025	-0.195	0.578
	$C_1^{(1)}$	-0.297	-0.454	0.525	0.356	0.551
	$d_2^{(1)}$	-0.022	-0.090	-0.782	0.528	0.318
The 1-st layer	$C_2^{(1)}$	0.668	-0.278	-0.088	-0.457	0.510
	$d_3^{(1)}$	-0.675	-0.093	0.000	0.492	-1.941
	$C_3^{(1)}$	0.703	0.014	0.011	-0.711	-0.017
	$d_4^{(1)}$	-0.490	0.512	-0.017	-0.487	0.511
	$C_4^{(1)}$	-0.270	-0.428	0.820	-0.262	-0.053

かる。つぎに、Table 2 で示した各中間変数の固有値の大きさを Fig. 3 に示す。図より第 2 層では、5 個の固有値の分布が各々の中間変数でかなり異なっているが、層

Table 2 Eigenvalues in each selection layer

Layer	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$\lambda_5$	Proportion
1	0.019	1.395	0.039	3.527	0.020	0.9844
	0.017	1.983	0.058	2.927	0.014	0.9820
	0.018	2.903	0.091	1.987	0.000	0.9780
	1.959	2.991	0.035	0.000	0.015	0.9900
2	0.553	0.324	0.504	1.455	2.163	0.7236
	0.800	0.331	1.316	0.755	1.798	0.6228
	1.452	0.134	2.833	0.564	0.017	0.8570
	1.674	0.270	2.854	0.059	0.144	0.9056
3	1.192	0.014	3.624	0.170	0.000	0.9632
	1.157	0.020	3.587	0.235	0.001	0.9488
	1.129	0.032	3.721	0.118	0.001	0.9700
	1.219	0.005	3.759	0.017	0.000	0.9956
4	1.092	0.001	0.000	0.002	3.906	0.9996
	1.097	0.001	0.000	0.003	3.900	0.9994
	1.100	0.001	3.898	0.000	0.001	0.9996
	0.010	1.160	0.000	3.752	0.078	0.9824
5	1.100	0.000	3.900	0.000	0.000	1.0000
	1.092	0.000	3.908	0.000	0.000	1.0000
	1.094	0.000	3.905	0.001	0.000	0.9998
	1.099	0.000	0.000	3.899	0.001	0.9996

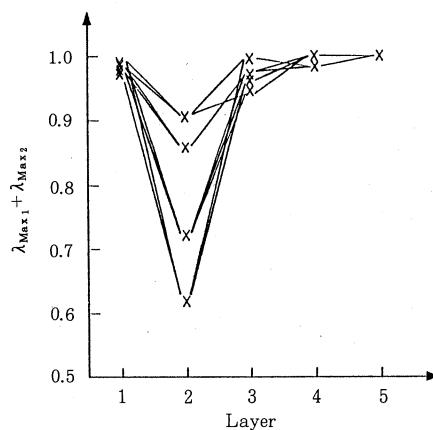


Fig. 2 Change of eigenvalues in the revised GMDH

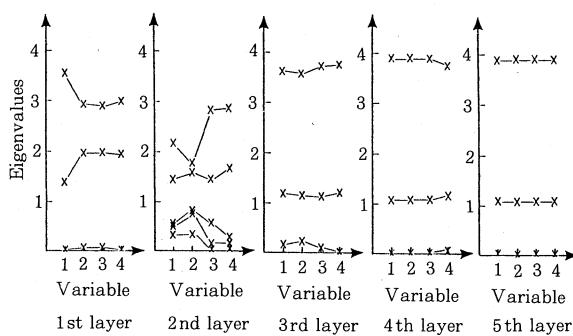


Fig. 3 Eigenvalues in each selection layer

を通過することにより各々の中間変数の固有値の分布が類似してきており、第5層では各々の中間変数ともほぼ同じ分布になった。これらのことにより、第4層以後の中間変数は同じ特徴を持った変数と見なすことができ、これ以上層を通過させて計算を続けても予測精度の向上は望めない。つぎに、各選択層における AIC の変化を Fig. 4 に示す。図より、第4層において AIC の値が最小になっていることがわかる。このため、多層構造の計算停止方法(2)を採用した場合においても、第4層のモデルが採用されることがわかる。

#### (5) 同定精度

同定に用いた 20 組の入出力データに対する同定精度を示す。同定精度は次式によって評価する。

$$J_1 = \frac{\sum_{\alpha=1}^{20} |\Phi_\alpha - \hat{\Phi}_\alpha|}{\sum_{\alpha=1}^{20} \Phi_\alpha} \times 100 (\%) \quad (20)$$

ここで、 $\hat{\Phi}_\alpha$  は出力変数  $\Phi$  の  $\alpha$  番目の出力データに対する推定値を表わす。

$$J_1 = 0.85\%$$

となった。

#### (6) 予測精度

予測に用いた残り 20 組の入出力データに対する予測精度を示す。予測精度は次式によって評価する。

$$J_2 = \frac{\sum_{\alpha=21}^{40} |\Phi_\alpha - \hat{\Phi}_\alpha|}{\sum_{\alpha=21}^{40} \Phi_\alpha} \times 100 (\%) \quad (21)$$

ここで、 $\hat{\Phi}_\alpha$  は出力変数  $\Phi$  の  $\alpha$  番目の出力データに対する推定値を表わす。

$$J_2 = 0.79\%$$

となった。Fig. 5 に同定と予測に用いた 40 組の入出力データに対する推定結果を示す。Fig. 5 より、同定精度、予測精度の両方とも非常に良いことがわかる。Fig. 5 では、真値と推定値との区別ができるないために、Fig. 6 に真値と推定値との差（残差）を示す。Fig. 6 より、残差は約 ±300 の範囲に分布しており、Fig. 5 から真値の変動が 0~40,000 の範囲であることを考慮すると、残差の変動が非常に小さいことがわかる。なお、残差の変動には対象システムの出力に加えたノイズ  $\varepsilon$  の影響が含まれている。参考までに、Table 3 に、対象システムの出力に加えたノイズ  $\varepsilon$  の標準偏差 ( $\sigma$ ) を変化させた場合の同定結果の一部を示す。3 ケースとも、同定されたモデルの構造は同じになった。

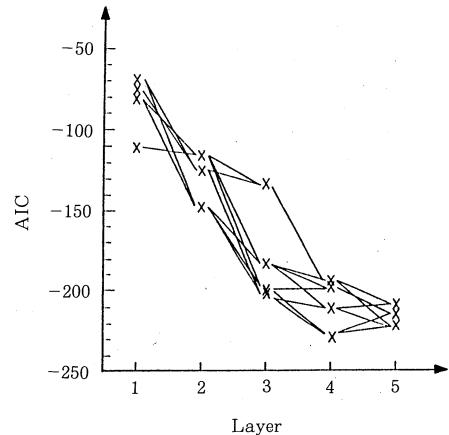


Fig. 4 Change of AIC in the revised GMDH

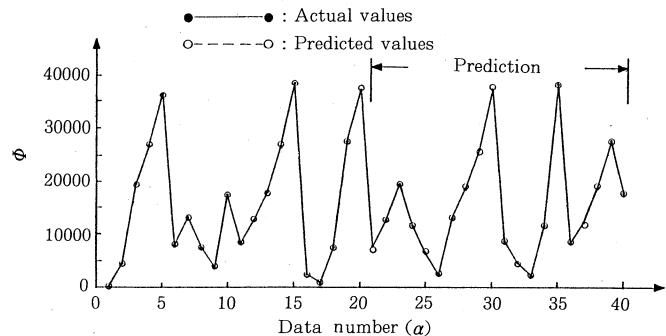


Fig. 5 The prediction results by the revised GMDH

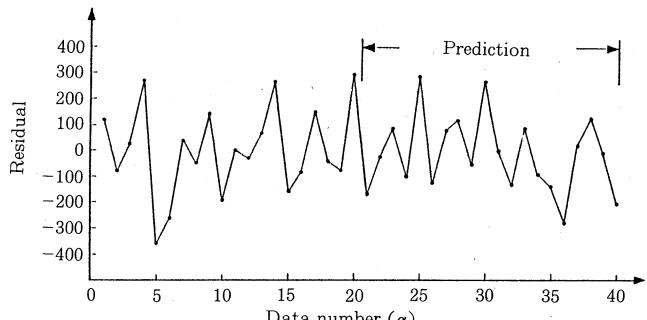


Fig. 6 Residuals of predicted values by the revised GMDH

Table 3 Accuracy for various standard deviations of noise ( $\varepsilon$ )

S.D. of noise ( $\varepsilon$ ) ( $\sigma$ )	$J_1$ (%)	$J_2$ (%)	AIC	$\frac{\lambda_1 + \lambda_2}{\Sigma \lambda}$
25	0.85	0.79	-226.8	0.9996
50	0.88	0.86	-225.0	0.9996
75	0.93	0.95	-222.8	0.9995

$\sigma$ を大きくしていくと、同定精度 ( $J_1$ )、予測精度 ( $J_2$ )、AIC 値とも徐々に増加している。これらのことから、対象システムに非常に大きな変動を伴うノイズが含まれているような同定問題では、多層構造の計算停止規準値 ( $E$ ) は 0.99 よりも少し小さな値にする必要があるものと思われる。

#### 4.2 従来形 GMDH による同定結果

##### (1) 中間変数

AIC 値の小さい変数を 4 個選択する。

##### (2) 同定されたモデル

第 5 層で計算を終了して次式が求まった。

(第 5 層)

$$\hat{\Phi} = w_1' \quad (22a)$$

(第 4 層)

$$w_1' = v_1' \quad (22b)$$

(第 3 層)

$$\begin{aligned} v_1' &= 1.163z_1' - 0.1595z_4' \\ &+ 0.4798z_1 \cdot z_4' - 0.5309z_4^2 \end{aligned} \quad (22c)$$

(第 2 層)

$$\begin{aligned} z_1' &= 0.6652y_2' + 0.5325y_4' \\ &+ 0.2116y_2 \cdot y_4' + 0.0838y_2^2 \\ &- 0.0592y_4^2 \end{aligned} \quad (22d)$$

$$z_4' = 0.6977y_1' + 0.2938y_2' \quad (22e)$$

(第 1 層)

$$y_1' = 0.7722x_1 \cdot x_2' + 0.1971x_1^2 \quad (22f)$$

$$y_2' = 0.8163x_1^2 + 0.2824x_3^2 \quad (22g)$$

$$y_4' = 0.6793x_2' \quad (22h)$$

##### (3) 多層構造の打ち切り

各選択層における AIC の変化を Fig. 7 に示す。第 5 層では部分表現式がすべて (6) 式の形となり、第 5 層の 4 個の中間変数はすべて同じ構造となった。

##### (4) 同定精度

同定に用いた 20 組の入出力データに対する同定精度を示す。同定精度は (20) 式によって評価する。

$$J_1 = 5.63\%$$

となった。

##### (5) 予測精度

予測に用いた残りの 20 組の入出力データに対する予測精度を示す。予測精度は (21) 式によって評価する。

$$J_2 = 5.57\%$$

となった。Fig. 8 に同定と予測に用いた 40 組の入出力

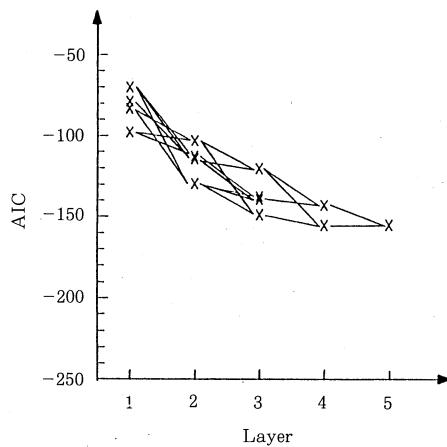


Fig. 7 Change of AIC in the previous GMDH

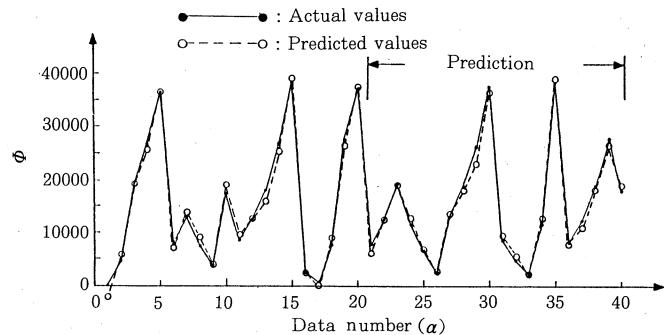


Fig. 8 The prediction results by the previous GMDH

Table 4 Accuracy of the revised GMDH and the previous GMDH

Type of GMDH	$J_1$ (%)	$J_2$ (%)	AIC
Previous GMDH	5.63	5.57	-156.0
Revised GMDH	0.85	0.79	-226.8

データに対する推定結果を示す。

#### 4.3 同定結果のまとめ

4.1 と 4.2 の同定結果を以下にまとめる。

Table 4 に、各々の GMDH の同定精度 ( $J_1$ ) と予測精度 ( $J_2$ ) および AIC の値を示す。

(1) 改良形 GMDH では同定精度 ( $J_1$ ) と予測精度 ( $J_2$ ) の両方とも従来形 GMDH と比較して精度が非常に良くなっている。多重共線性を起こさないモデル式が得られていることがわかる。

(2) AIC の値も改良形 GMDH の方が大幅に小さくなっている。良いモデル式が得られていることがわかる。

(3) 改良形 GMDH では、各選択層における中間変数の収束状態を、中間変数を構成する非線形主成分の固有値の分布と累積寄与率を調べることにより、正確に判断できることがわかった。このため、中間変数の固有値と累積寄与率を用いて多層構造の計算停止方法を構成するこ

とが可能になった。さらに、従来形 GMDH と同様に AIC 値を用いた多層構造の計算停止規準も構成することができた。

## 5. むすび

主成分回帰分析を用いる改良形 GMDH のアルゴリズムを提案し、簡単な同定問題に適用した。そして、従来から提案されている GMDH によって得られた結果と比較することにより次のことが明らかになった。

改良形 GMDH は、同定精度と予測精度の両方の精度、更に AIC の値とも非常に良い精度が得られており、多重共線性を起こさない精度の良いモデルを同定することができた。さらに、改良形 GMDH では、各選択層における中間変数の収束状態を、中間変数を構成する非線形主成分の固有値の分布と累積寄与率を調べることにより判断できることがわかった。このため、従来形 GMDH と比較して、より正確な多層構造の計算停止規準を構成することが可能になった。さらに、従来形 GMDH と同様に AIC の値を用いた多層構造の計算停止規準も構成することができた。

以上により、本研究で提案した改良形 GMDH の有效性が確認できた。

最後に、GMDH の研究に関して、長年にわたりご指導ご鞭撻を賜りました、大阪大学工学部田村坦之教授に深甚の謝意を表す。

## 参考文献

- 1) Stanly J. Farlow (Ed.) : Self-Organizing Methods In Modeling-GMDH Type Algorithms, Mercel Dekker, Inc. (1984)
- 2) 近藤：非線形指数形モデルを同定する改良形 GMDH；計測自動制御学会論文集, Vol. 22, No. 12, pp. 1283~1289 (1986)
- 3) 田村、近藤：モデル選択の評価規準に予測平方和 (PSS) を用いる改良形 GMDH；計測自動制御学会論文集, Vol. 14, No. 5, pp. 519~524 (1978)
- 4) 池田ほか：GMDH における部分モデルの正規化と Ridge 回帰；第 6 回システムシンポジウム (SICE), pp. 119~124 (1980)
- 5) A. G. Ivakhnenko, M. M. Touda and Yu. V. Chukin : GMDH Algorithm with Successive Determination of Polynomial Trends Using the Most Essential Variables ; Soviet Auto. Control, Vol. 5, No. 2, pp. 44~54 (1972)
- 6) A. G. Ivakhnenko, M. M. Touda and A. A. Fomichev : Structural Synthesis of Filters for Indirect Measurement of Variables on the GMDH Algorithm ; Soviet Auto. Control, Vol. 6, No. 1, pp. 37~47 (1973)
- 7) 近藤：複雑な構造をした非線形システムの同定を目的としたニューラルネットワーク；システム制御情報学会論文誌, Vol. 4, No. 7, pp. 259~266 (1991)
- 8) 赤池：情報量規準 AIC とは何か；数理科学, No. 153, pp. 5~11 (1976)
- 9) 奥野ほか：統多変量解析法, 日科技連 (1976)