類似度データへのツリーモデルの当てはめについて

前川眞一
東京工業大学　大学院社会理工学研究科 970410, 071023 *

1　ツリーモデル

本稿では，N 個の個体 $(o_i, i = 1, 2, \ldots, N)$ 間の類似度もしくは非類似度の指標 $(o_{ij}, i > j)$ が与えられた場合に，最小２乗法を用いてツリーモデルを当てはめることを考える．ツリーモデルは大きく分けて加算的ツリーモデル (additive tree model) と超距離ツリーモデル (ultrametric tree model) に分けることができるが，ここでは超距離モデルのみを取り扱う．一般に，超距離と呼ばれるものは，距離の３つの性質に加えて以下の超距離不等式と呼ばれる不等式を満たすようなものであるが，

$$\begin{array}{lcll} d_{ij} & \geq & 0 \quad \text{and} \quad d_{ii} = 0 & \\ d_{ij} & = & d_{ji} & \text{対称性} \\ d_{ij} & + & d_{jk} \geq d_{ik} & \text{三角不等式} \\ d_{ij} & \leq & \max\left(d_{ik}, d_{jk}\right) & \text{超距離不等式} \end{array} \tag{1}$$

このようなものは樹状図上の距離（樹状距離）と解釈することができる．たとえば，図 1 は 7 個の個体の樹状図を示したものであるが，各個体をつなぐ最もっとも低い水平線の高さで個体間の距離を定義するとそれが超距離の不等式を満たすことになる．この図の場合には個体間距離行列はそれを各水平線の右肩につけたクラスターを示す番号で表せば 表 1 の上段部分のようになる．

この表からわかるように，距離行列の行と列を樹状図の横軸の順にとった場合，対角要素から離れるにしたがって，個体間の距離は減少することはない．このような特性を持った距離行列を Robinson property (Kendall, 1969) と呼ぶこともある．この点に注目すれば，各個体は樹状図の横軸の順に１次元の弱い順序尺度を構成しているとも考えられるが，（ Critchley & Heiser,1988 を参照） 同一のツリー構造を指し示すツリーは数多く描くことが出来る．例えば，図 1 において，(個体 1 個体 2 ）の対と（個体 3 個体 4 ）の対の順序を入れ替えてもツリー構造は変わらない．また，クラスターの名称（番号）の付け方も自由であるが，本稿では一応，番号の大きなクラスターが番号の小さなクラスターを含むものとしている．

2　線形モデルによる樹状距離の当てはめ

このような樹状距離は，以下のようにして線形モデルの形に書くことができる．いま， $N \times N$ の非類似度データを $\boldsymbol{O}$, それをベクトル化したものを $\boldsymbol{o}$ （ $N(N-1)/2 \times 1$ ）とする．また，$\boldsymbol{D}$ を $N \times N$ の超距離行列，$\boldsymbol{D}$ をベクトル化したものを $\boldsymbol{d}$ とする．これらの記号ならびにツリー構造を表す $N(N-1)/2 \times q$ のデザイン行列 $\boldsymbol{X}$ と各クラスターの高さを表すベクトル $\boldsymbol{\beta}$ を用いれば，

$$\boldsymbol{o} \quad = \quad \boldsymbol{d} + \boldsymbol{\varepsilon} \tag{2}$$

* 〒 152-8552 目黒区大岡山 2-12-1
Telefax: 81-3-5734-3242,　e-mail: mayekawa@hum.titech.ac.jp
本稿は前川 (1997) の第 5.3 節に加筆修正したものである.

$$\boldsymbol{d} \quad = \quad \boldsymbol{X}\boldsymbol{\beta} \tag{3}$$

である．すなわち，x_{ij} は第 i 番目の個体の対間の距離が第 j 番目の高さで表される場合に 1 を取り，それ以外は 0 となるような行列である．ただし，$\boldsymbol{\beta}$ の要素間にはツリー構造が意味する高さ間の順序の制約を表す $(q-1)\times q$ の行列 $\boldsymbol{A}$ により

$$\boldsymbol{A}\boldsymbol{\beta} \geq \boldsymbol{0} \tag{4}$$

という制約がかかる．この $\boldsymbol{A}$ 行列は，各行に示すクラスターの親クラスターを 1 という要素で示したものとなっている，なお，超距離ツリーを線形モデルとして記述する方法に関しては Corter(1996, p56-60) を参照されたい．また，Mayekawa & Peng(1981) には上記の方法と少し異なる記述の方法が記してある．図 1 の例では，これらの行列は以下に示すようなものになる．

$$\boldsymbol{X} = \begin{bmatrix}
 & c1 & c2 & c3 & c4 & c5 & c6 \\
o2o1 & 1 & 0 & 0 & 0 & 0 & 0 \\
o3o1 & 0 & 0 & 1 & 0 & 0 & 0 \\
o3o2 & 0 & 0 & 1 & 0 & 0 & 0 \\
o4o1 & 0 & 0 & 1 & 0 & 0 & 0 \\
o4o2 & 0 & 0 & 1 & 0 & 0 & 0 \\
o4o3 & 0 & 1 & 0 & 0 & 0 & 0 \\
o5o1 & 0 & 0 & 0 & 0 & 0 & 1 \\
o5o2 & 0 & 0 & 0 & 0 & 0 & 1 \\
o5o3 & 0 & 0 & 0 & 0 & 0 & 1 \\
o5o4 & 0 & 0 & 0 & 0 & 0 & 1 \\
o6o1 & 0 & 0 & 0 & 0 & 0 & 1 \\
o6o2 & 0 & 0 & 0 & 0 & 0 & 1 \\
o6o3 & 0 & 0 & 0 & 0 & 0 & 1 \\
o6o4 & 0 & 0 & 0 & 0 & 0 & 1 \\
o6o5 & 0 & 0 & 0 & 0 & 1 & 0 \\
o7o1 & 0 & 0 & 0 & 0 & 0 & 1 \\
o7o2 & 0 & 0 & 0 & 0 & 0 & 1 \\
o7o3 & 0 & 0 & 0 & 0 & 0 & 1 \\
o7o4 & 0 & 0 & 0 & 0 & 0 & 1 \\
o7o5 & 0 & 0 & 0 & 0 & 1 & 0 \\
o7o6 & 0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix}
 & c1 & c2 & c3 & c4 & c5 & c6 \\
c1 & -1 & 0 & 1 & 0 & 0 & 0 \\
c2 & 0 & -1 & 1 & 0 & 0 & 0 \\
c3 & 0 & 0 & -1 & 0 & 0 & 1 \\
c4 & 0 & 0 & 0 & -1 & 1 & 0 \\
c5 & 0 & 0 & 0 & 0 & -1 & 1
\end{bmatrix}$$

したがって，データとして $N\times N$ の非類似度行列 $\boldsymbol{O}$ が与えられたときに，それに超距離ツリーモデルを当てはめるということは，制約条件 (4) 式 の下で

$$\mathrm{RSS} = \|\boldsymbol{O}-\boldsymbol{D}\|^2 = \|\boldsymbol{o}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{5}$$

を最小にするようにツリー構造 ($\boldsymbol{X}$ と $\boldsymbol{A}$) と高さ $\boldsymbol{\beta}$ を求めることに他ならない．この問題は Carroll & Pruzansky(1975), De Soete(1984a,b) 等によって解が与えられているが，$\boldsymbol{X}$ 行列の中途半端な単純性のために，解を求める手続きが通常の非線形の最小２乗法と比べて極端に複雑になる．したがって，ここでは，ツリー構造 ($\boldsymbol{X}$ と $\boldsymbol{A}$) は既知であるとして，非類似度データに最もよくフィットする超距離行列を求めることを考える．

この問題は， RSS を最小とするような $\boldsymbol{\beta}$ を上記の制約条件の下で求めることに他ならないが，そのためには

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}^{(*)^{-1}} \qquad \text{and} \qquad \boldsymbol{\gamma} = \boldsymbol{A}^{(*)}\boldsymbol{\beta} \tag{6}$$

という変数変換，ただし， $\boldsymbol{A}^{(*)}$ は $\boldsymbol{A}$ 行列に1 行加えてそれを正則にしたもの，(例えば，上記の例では，第 6 行目に第 6 番目の要素が 1 それ以外は 0 という行を加える.）を用いて

$$\boldsymbol{d} = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}{\boldsymbol{A}^{(*)}}^{-1}\boldsymbol{A}^{(*)}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\gamma} \tag{7}$$

とし，

$$\boldsymbol{\gamma} \geq \boldsymbol{0} \tag{8}$$

を満たすような $\boldsymbol{\gamma}$ を求めれば簡単になる．このような制約付きの最小２乗法の解は前川 (1997) の第 4.4.2 節や第 4.6.1 節に示した方法を用いて求めることができる．また，$\boldsymbol{\gamma}$ が求まれば，樹状図の高さは

$$\boldsymbol{\beta} = {\boldsymbol{A}^{(*)}}^{-1}\boldsymbol{\gamma} \tag{9}$$

として求まる．なお，樹状図の構造 ($\boldsymbol{X}$,$\boldsymbol{A}$) が元々データを群平均法 (goup average method) を用いて分析したものとして与えられている場合，このようにして求めた樹状図の高さは群平均法で求められた高さと同一のものとなる.

この方法の例として，表 1 の中段に示された非類似度データ行列に 図 1 に示されたツリー構造を当てはめた場合，同表の下部に示したような最小２乗解が得られた．また，図 1 の左端にはそれがツリーの高さとして表示してある.

3　クラスターの細部を無視した樹状距離の当てはめ

前節に示した方法は様々な形で応用が可能であるが，その一つは Mayekawa & Peng(1981) が示したように複数の非類似度データ行列が与えられた場合に，グループ間の差異を既知の共通のツリー構造の高さの違いとして表現する個人差ツリーモデルである．すなわち，

$$\mathrm{RSS} = \sum_{s=1}^{S} \left\| \boldsymbol{O}^{(s)} - \boldsymbol{D}^{(s)} \right\|^2 = \sum_{s=1}^{S} \left\| \boldsymbol{o}^{(s)} - \boldsymbol{X}\boldsymbol{\beta}_s \right\|^2 \tag{10}$$

を最小にするような $\boldsymbol{\beta}_s, s = 1, 2, \ldots, S$ を，制約条件

$$\boldsymbol{A}\boldsymbol{\beta}_s \geq \boldsymbol{0}, s = 1, 2, \ldots, S \tag{11}$$

のもとに求めればよいのであるが，各 $\boldsymbol{\beta}_s$ が分離可能であるため，これは上記の最小２乗解を s 回求めることに帰着する.

また，Kruskal & Wish(1978) 等が推奨しているように，多次元尺度法を用いて得られた個体の布置とクラスター分析により得られた個体間類似度の階層構造とのー致度を数値的に見るためにも利用することができる．ただ，その際に，細部のツリー構造は無視して，大きな（上位の）クラスターのみに着目したい場合，上記の方法をそのままの形では利用することができない．例えば，図 3 に示した多次元尺度法の布置と 図 2 に示した同じ類似度データのクラスター分析によるツリー構造ののー致度を見ようとする場合に，図 3 の楕円で囲んだ部分のみに着目する様な場合である.（クラスター 3, 6 の内部は無視する.）

このように，ある水準以下のクラスターを無視して超距離をフィットする場合には、それらのクラスターに含まれる個体間距離をパラメタとして上記の制約付き最小２乗解を求めればよい。例えば， 図 2 の 3 つのクラスターの場合は，$d_{(A)(B)}$ で集合 A の要素と集合 B の要素の間のすべての対間の距離を示せば，

クラスター 3

$$d_{(1234)(1234)} \quad \leq \quad d_{(1234)(567891011)} \tag{12}$$

クラスター 6

$$d_{(5678)(5678)} \quad \leq \quad d_{(5678)(123491011)} \tag{13}$$

という距離間の大小関係が示されているが，これらの関係はc_9 をクラスター 3 の親クラスター 9 の高さ，また， c_{10} をクラスター 6 の親クラスター 10 の高さとすれば，

クラスター 3

$$d_{(1234)(1234)} \quad \leq \quad c_9 \tag{14}$$

クラスター 6

$$d_{(5678)(5678)} \quad \leq \quad c_{10} \tag{15}$$

としても同じである．従って，$c_1, c_2, c_3, c_4, c_5, c_6$ をパラメタからはずし，代わりに $d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34}$ および $d_{56}, d_{57}, d_{58}, d_{67}, d_{68}, d_{78}$ を新たにパラメタとして導入し，それらの間に

$$d_{12}, d_{13}, d_{14}, d_{23}, d_{24}, d_{34} \leq c_9 \tag{16}$$

$$d_{56}, d_{57}, d_{58}, d_{67}, d_{68}, d_{78} \leq c_{10} \tag{17}$$

という合計 12 個の制約条件課すことと同値である．この例の場合の $\boldsymbol{X}$ 行列と $\boldsymbol{A}$ 行列を 表 2 に示した．

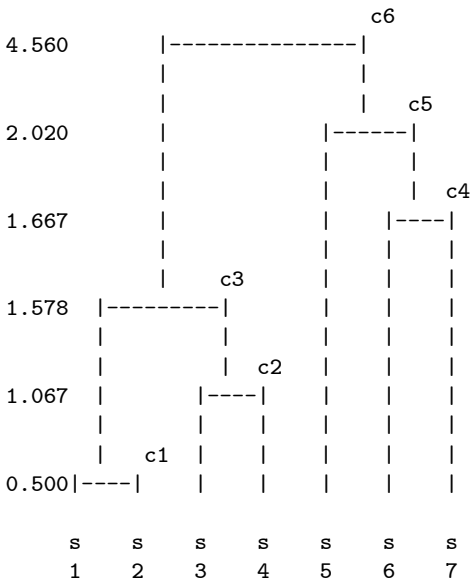

図 1: 樹状図の例

参考文献

Carroll, D.J. & Pruzansky, S.(1975) Fitting of hiearchical tree structures(HTS) models, mixtures of HTS models, and hybrid models, via mathematical programing and alternating least squares. Paper presented at the U.S.-Japan Seminar in Theory, Method, and Applications of Multidimensional Scaling and Related Techniques, San Diego, CA.

Corter, J.E. (1996) Tree models of similarity and association. Thousand Oaks, CA: SAGE Publications.

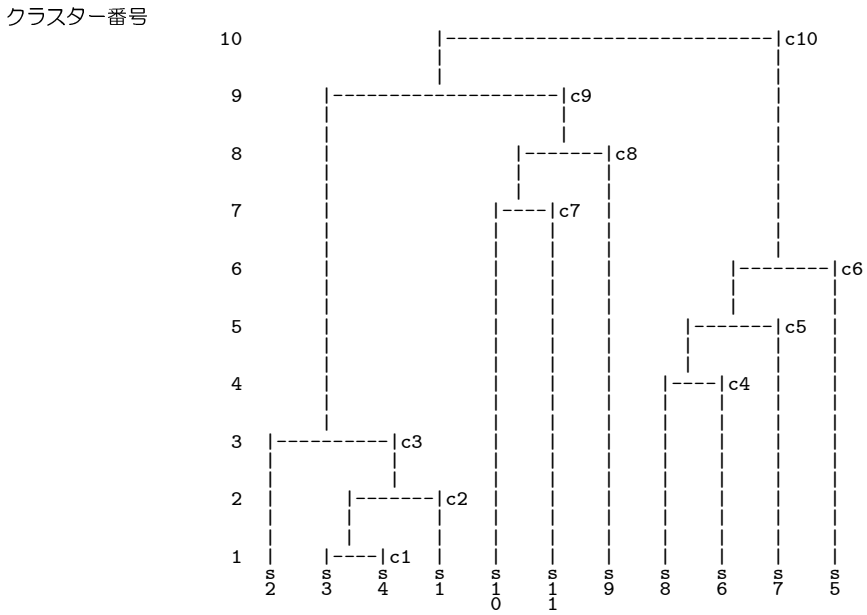

図 2: クラスター分析によるツリー構造

Critshley, F. & Heiser, W. (1988) Hierarchical Tress Can be Perfectly Scaled in One Dimension. Journal of Classification, 5, 5-20.

De Soete, G. (1984b) Ultrametric tree representations of incomplete dissimilarity data. *Journal of Classification*, 1, 235-242.

Kendall, D.G.(1969) Some problems and methods in statistical archaeology. World Archaeology, 1, 68-76.

Kruskal, J.B. & Wish, M. (1978) Multidimensional scaling. Biverly Hills: SAGE Publications. （高根芳雄（訳）(1980) 多次元尺度法. 朝倉書店

Mayekawa, S. & Peng, J. (1981) Individial differences in hierarchical cluster analysis with multiple subject. Paper presented at the annual meeting of the Psychometric Society, Chapell Hill, NC.

前川眞一 (1997) SAS による多変量データ解析 東京大学出版会

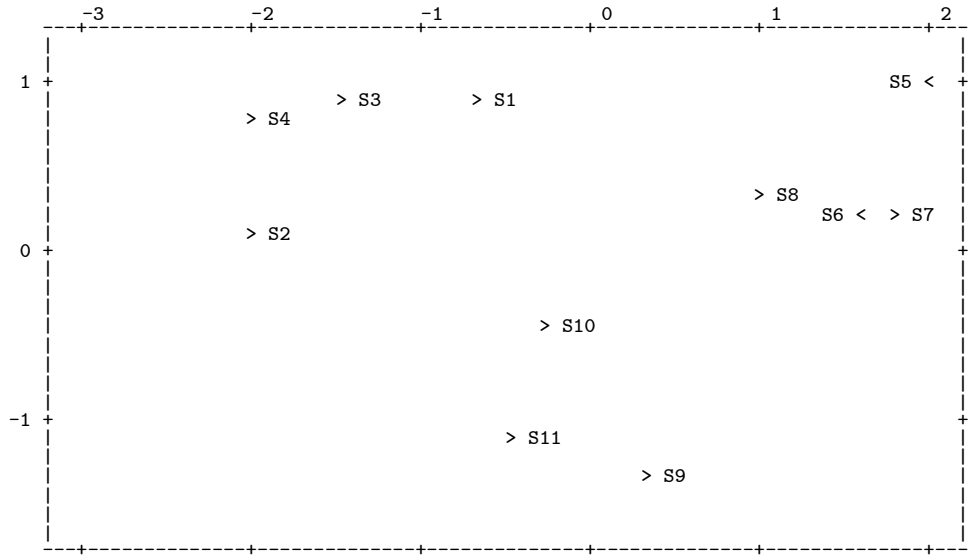

図 3: 多次元尺度法による個体の布置

表 1: 非類似度データ行列と樹状距離行列

樹状距離行列

	o1	o2	o3	o4	o5	o6	o7
o1	0	1	3	3	6	6	6
o2	1	0	3	3	6	6	6
o3	3	3	0	2	6	6	6
o4	3	3	2	0	6	6	6
o5	6	6	6	6	0	5	5
o6	6	6	6	6	5	0	4
o7	6	6	6	6	5	4	0

非類似度データ行列

	o1	o2	o3	o4	o5	o6	o7
o1	0.000	0.500	1.179	1.951	6.067	4.631	4.055
o2	0.500	0.000	1.344	1.841	5.588	4.206	3.563
o3	1.179	1.344	0.000	1.067	6.021	4.327	4.197
o4	1.951	1.841	1.067	0.000	5.169	3.371	3.532
o5	6.067	5.588	6.021	5.169	0.000	1.951	2.088
o6	4.631	4.206	4.327	3.371	1.951	0.000	1.667
o7	4.055	3.563	4.197	3.532	2.088	1.667	0.000

最小２乗樹状距離行列

	o1	o2	o3	o4	o5	o6	o7
o1	0.000	0.500	1.578	1.578	4.560	4.560	4.560
o2	0.500	0.000	1.578	1.578	4.560	4.560	4.560
o3	1.578	1.578	0.000	1.067	4.560	4.560	4.560
o4	1.578	1.578	1.067	0.000	4.560	4.560	4.560
o5	4.560	4.560	4.560	4.560	0.000	2.020	2.020
o6	4.560	4.560	4.560	4.560	2.020	0.000	1.667
o7	4.560	4.560	4.560	4.560	2.020	1.667	0.000

表 2: ツリー構造の細部を無視した場合の線形モデル

X	c7	c8	c9	c10	s2s1	s3s1	s3s2	s4s1	s4s2	s4s3	s6s5	s7s5	s7s6	s8s5	s8s6	s8s7
s2s1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
s3s1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
s3s2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
s4s1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
s4s2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
s4s3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
s5s1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s5s2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s5s3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s5s4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s6s1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s6s2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s6s3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s6s4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s6s5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s7s1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s7s2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s7s3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s7s4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s7s5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s7s6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s8s1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s8s2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s8s3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s8s4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s8s5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s8s6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s8s7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s9s1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s9s2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s9s3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s9s4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s9s5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s9s6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s9s7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s9s8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s10s1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s10s2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s10s3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s10s4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s10s5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s10s6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s10s7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s10s8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s10s9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s11s1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s11s2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s11s3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s11s4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
s11s5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s11s6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s11s7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s11s8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
s11s9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s11s10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A	c7	c8	c9	c10	s2s1	s3s1	s3s2	s4s1	s4s2	s4s3	s6s5	s7s5	s7s6	s8s5	s8s6	s8s7
c7	-1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c8	0	-1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
c9	0	0	-1	1	0	0	0	0	0	0	0	0	0	0	0	0
s2s1	0	0	1	0	-1	0	0	0	0	0	0	0	0	0	0	0
s3s1	0	0	1	0	0	-1	0	0	0	0	0	0	0	0	0	0
s3s2	0	0	1	0	0	0	-1	0	0	0	0	0	0	0	0	0
s4s1	0	0	1	0	0	0	0	-1	0	0	0	0	0	0	0	0
s4s2	0	0	1	0	0	0	0	0	-1	0	0	0	0	0	0	0
s4s3	0	0	1	0	0	0	0	0	0	-1	0	0	0	0	0	0
s6s5	0	0	0	1	0	0	0	0	0	0	-1	0	0	0	0	0
s7s5	0	0	0	1	0	0	0	0	0	0	0	-1	0	0	0	0
s7s6	0	0	0	1	0	0	0	0	0	0	0	0	-1	0	0	0
s8s5	0	0	0	1	0	0	0	0	0	0	0	0	0	-1	0	0
s8s6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	-1	0
s8s7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	-1