次元縮約によるマウスの行動実験データの解析

山口瑶子　(指導教員：瀬々潤)

1　研究背景

一卵性双生児は、遺伝子上は全く同一でありながら、互いに個性が違ったり、運動能力、知能などに差ができることが有る。本研究では、同一の系統（一卵性双生児）のマウス 41 匹に主として知能に関連する行動のテストを行い、この観測結果を元にして、知能の高いマウス、低いマウスを見つけ出す。41 匹のマウスに対して行う行動テストは 1 日数回、1 週間以上に渡って行われるのもの多く、データとしては高次元のデータとなっているため、マウス間の差異や共通性を理解するのは容易ではない。そこで、ここでは次元縮約を用いデータを低次元で観測する事で、マウス間にどのような行動の差異や共通性があるかを発見し、ひいては、知能の高いマウスを発見する事を目指す。

2　行動実験データの説明

今回、次の 3 種類の実験を行い、マウスの行動データを採取した。

バーンズ迷路 (BM) 円盤に 12 個の穴があり、そのうち 1 つの穴の下にマウスが好む暗い箱 (ターゲット) があり、マウスがその穴に入るまでの時間を計測する。ただし, 毎日円盤を 90 度回転させることで、ターゲットの位置を変えることとする。1 日 3 回、7 日間行われる。この実験では空間記憶、参照記憶を調べることができる。

8 方向放射状迷路 (RM) 8 方向にのびたアームの先端に餌を置き, 中央に置いたマウスが餌を取ってまわる様子を観察し、全ての餌を取るまでに間違ったアームに入った回数を計測する。ただし, 隣り合ったアームを順にまわっていくことができないよう、ドアが自動的に開閉する。全部で 28 回の実験を行い、23 回目からは、4 つの正解のあと、中央での待ち時間を入れる事でテストを難しくした。この実験では作業記憶を調べることができる。

恐怖条件付けテスト (FZ) 1 日目に音刺激と電気ショックを組み合わせて条件付けを行い、2 日目に電気ショックをかけたのとは違う箱で音のみ聞かせ、フリージングの割合を測定した。1 分間ごとに区切って 6 分間フリージングの時間を計測する。この実験では恐怖に対する記憶が調べられる。

各実験は複数のトライアル (実験) を含み、それぞれ数値で計測されている。今回用いたマウスの行動実験データの模式図を表 1 に示す。1 列目にマウスの番号、1 行目にトライアル番号 (実験番号) が記載されている。41 匹のマウスに対して実験を行っているので,ID は 1〜41 番まで、BM,RM,FZ の各実験はそれぞれ 21 個,28 個,6 個のトライアルを含む、55 次元のデータある。本研究では、このような高次元のデータの傾向を変えずに、できるだけ低次元で表現することで、各マウスの特徴づけを目指す。

mouseID	BM	RM	FZ
01	···	···	
···	21trials	28trials	6trials
41	···	···	···

表 1: 行動実験データの概要

3　手法

3.1　標準化

はじめに、各トライアルで計測の単位や条件が異なっているため、これらを均一に扱えるよう、データの標準化を行う。標準化は、各トライアル独立に平均 0, 分散 1 となるように変換した。

3.2　主成分分析

標準化したデータを基に、3 つの異なる実験の内、どの実験がマウスを特徴づける実験として適切かを選択したい。ここでは、次元縮約の手法として一般的な主成分分析を用いてデータの分布を良く表す低次元空間を自動的に見つけることで、マウスを特徴づける実験の選択を行う。

主成分分析は、与えられた多次元空間上の点を射影したとき、分散が最も大きくなるような軸を選択する方法である。$\mathbf{X}$ を n 匹のマウスを対象として、m 回のトライアルを行った実験データとすると、n 個の点 $\mathbf{X}$ を $\mathbf{a}$ に沿って射影したとき、分散 σ^2 が最大となるような $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$ を満たすような $\mathbf{a}$ 求めよう。

$$\sigma^2 = (\mathbf{Xa})^{\mathrm{T}}(\mathbf{Xa}) = \mathbf{a}^{\mathrm{T}}\mathbf{Sa}$$

(ここで $\mathbf{S}$ は $\mathbf{X}$ の共分散行列)。この分散が最大になる点を、ラグランジュ未定乗数法を用いて求めると、次式の解が分散を最大化するベクトル $\mathbf{a}$ となることが分かる。

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = 0$$

この式は $\mathbf{S}$ の固有値、固有ベクトルを求める式と同一である。$\mathbf{S}$ の固有値を $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ とすると、主成分分析では、λ_i に対応する固有ベクトルを第 i 主成分と呼び、より大きな λ_i に対応する主成分の方が、射影した点の分散を大きくする軸となっている。

3.3　実験の選択

記憶力を調べる 3 つの実験 BM,RM,FZ からマウスの差異が最も表れた実験を選び出すために、表 1 を標準化したものを主成分分析した。図 3 に、主成分分析後の散布図 (第 1,2 主成分) を示す。主成分分析の結果かた、累積寄与率が 10%を超える主成分 1, 主成分 2, 主成分 3 に着目する。傾きを表す方向ベクトルの大きさが 0.3 を超える次元を調べると RM,BM,FZ それぞれ 67%,33%,0% という割合になった。以上より、3 つの実験 RM,BM,FZ の中でマウスの差異を調べるのに最も有効な実験は RM と考えられるため、RM を取り上げて解析を行った。

3.4 サンプル選択

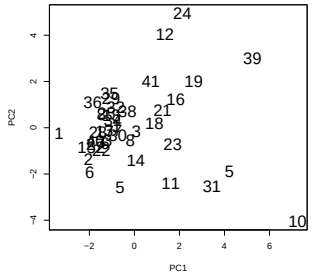

図 1: 主成分 1 , 主成分 2 による RM データのマウスの散布図

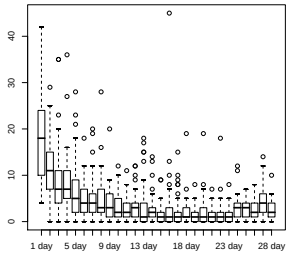

図 2: RM のトライアル毎のマウスの分布を表す箱髭図

RM のデータを主成分分析し第 1 主成分と第 2 主成分の散布図を作成した (図 1)。図 1 を観察すると、多くのマウスが一部に集中して固まっている一方で、固まっている付近から離れた位置に散布しているマウスも見ることができ、この外れ値となっているマウスの原因を解析した。各トライアルについて、値の分布を箱髭図で表示した (図 2)。主成分 1 が最も大きく傾いていた 4day のトライアルを見てみると、ID-04 のマウスが 27 回,ID-07 のマウスが 36 回という実験結果で、トライアルの平均値から離れた値をとっていることが確認できる。図 1 上で、ID-04 と ID-07 のマウスの位置を確認してみると、横軸 (主成分 1) に関して大きな値をとっていることが確認できる。また、主成分 2 が大きく傾いているトライアル 16day では ID-9 のマウスが分散の 1.5 倍、トライアル 18day では ID-20 のマウスでは 5.4 倍、ID-33 のマウスでは 1.5 倍平均値から離れており、外れ値となっていることが確認できる。図 1 上でこれらのマウスの位置を確認してみると、縦軸 (主成分 2) に関して大きな値をとっており、主成分の向きに大きな寄与を与えてしまっていると考えられる。そこで、以上で挙げた ID-04,07,09,20,33 のマウスを取り除いたデータを用い解析をすすめる。

3.5 トライアル選択

RM の実験のみでも 28 コの実験があるため、トライアルの選択をすることで低次元でマウスの特徴を表したい。ここでは、RM のトライアルの中から結果に差異が表れたと考えられるトライアルを探る。

結果に差異が現れたトライアルとは、単純には分散が大きなトライアルと考えられるが、外れ値の影響により分散が不当に大きくなっている場合もあり、分散だけでは判断ができない。また、結果に差異が現れたトライアルとは、平均値付近には値がなく、平均値前後に値が集まる、2 コブ型の分布を示していることが望ましい。このような分布を見つけるために、次の指標を導入する。

トライアル t のマウス i の観測値を x_{ti} とし、トライアル t の平均値を $\bar{x}_t$ とする。いま、見つけたい 2 コブ型の分布では、分散 $\sum_i (x_{ti}-\bar{x}_t)^2$ は、大きくなってほしい。また、外れ値検出のために、平均値よりも大きく外れた値が存在すると、3 次のモーメント $\sum_i (x_{ti}-\bar{x}_t)^3$ が大きくなる事を利用すると、$\sum_i (x_{ti}-\bar{x}_t)^3 / \sum_i (x_{ti}-\bar{x}_t)^2$ が小さいほど、外れ値の無い分布になる可能性が高い。

ところが、上記の指標では、単に一様な分布をしている場合でも、値が小さくなってしまう問題があるので、平均値付近に値が存在しない事を確認する指標として、1 次のモーメント $\sum_i |x_{ti}-\bar{x}_t|$ を利用する。この値は、平均付近に値が集中すると小さくなる。よって、指標 $\sum_i (x_{ti}-\bar{x}_t)^2 / \sum_i |x_{ti}-\bar{x}_t|$ を考えれば、同一の分散を持つ 2 つのトライアルがあるとしたら、この指標は平均付近に値が無いほど、小さくなる。

以上の 2 つの指標を合成した、以下の指標 R をなるべく小さくするトライアルを選択することで、2 コブ型の分布のパラメータを抽出することが可能となる。

$$R = \left(\frac{\sum_i (x_{ti}-\bar{x}_t)^3}{\sum (x_{ti}-\bar{x}_t)^2}\right) \cdot \left(\frac{\sum_i (x_{ti}-\bar{x}_t)^2}{\sum |x_{ti}-\bar{x}_t|}\right) = \left(\frac{\sum_i (x_{ti}-\bar{x}_t)^3}{\sum |x_{ti}-\bar{x}_t|}\right)$$

RM のトライアルにおいて、R を計算した所、最も小さい 2 つのトライアルは、trial4 と trial25 であった。つまり、この 2 つのトライアルでは、結果が良かったグループとそうでなかったグループとにはっきりと分かれているということができる。

3.6 散布図の比較

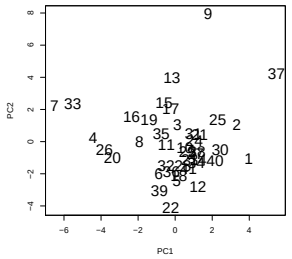

図 3: BM,RM,FZ のデータをまとめて主成分分析を行った散布図

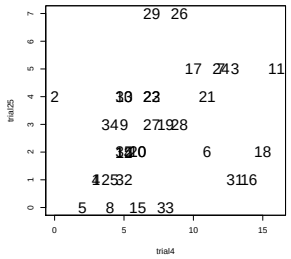

図 4: RM の trial4 と trial25 の散布図

図 3 と図 4 の散布図を比較してみると、図 3 ではマウスが密集している部分があるのに対し、図 4 では、マウスは密集することなく散らばっている事が分かる。2 次元での分散の大きさが図 3 では約 1.82 となり図 4 では約 3.74 となり図 4 のほうが分散が大きいと言える。次元縮約で主成分分析を用いる場合、求めた主成分で高次元のデータを低次元で観察する。本手法では、主成分を単に軸として選ぶのではなく、主成分が傾いた向きを調べることで、分散が大きく、かつどのトライアルに着目すればマウスの記憶力、学習能力の差が表れるかを選択し、高次元のデータを低次元で表す事ができた。

4 今後の課題

今回は様々な閾値を手動で選択しているが、今後手法の改良と共に閾値の自動選択を行えるよう、改善したい。主成分分析以外のデータを解析する方法を考案し、引き続きデータ解析を行う。記憶力、学習能力、運動能力などそれぞれの能力間に相関関係があるか探る。そして、マウスの行動実験を解析することで得られた行動の共通性や差異の情報を元にして、遺伝子の発現量データを解析し、マウスの行動に関係していると思われる遺伝子を探る。