宇宙物理研究室 http://www.is.oit.ac.jp/~shinkai/

主成分分析を用いた競馬のデータ分析

Astrophysics Group, OIT 邉見広大

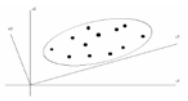
目的：過去のデータから勝ち馬を予想する
概要：良く使われる指標がどれだけ有効なのか統計的に分析する

分析の流れ

- データは過去４年分の有馬記念の出馬６１頭と、過去４年分の中山競馬場2500mの勝ち馬42頭である。
- データは
  右回り
  左回り
  中山競馬場
  ＧＩ出走
  の勝率を用いて主成分分析を行う。
- ４種のデータの相関を調べ、勝ち馬との対応を分析する。

主成分分析 ①合成変量

- 例えば、統計データ (x_1, x_2) が与えられたとき、１次結合 $z = a_1x_1 + a_2x_2$ で新たな変量を構成する。
- データの分布を図のような楕円と見なす時、長軸・短軸方向に相当する軸の向きを見つける。
  $a_1 = cos\theta$
  $a_2 = sin\theta$
  に対応する



- ｚを使うことにより、もとの変量の個数よりも少ない個数の変動の動きにまとめることである。

主成分分析 ②合成変量の分散

- 合成変量zの分散を求めると、
  $$Var(z) = a_{12}v_{11} + 2a_1a_2v_{12} + a_{22}v_{22} \quad -(1)$$
  となる。ラグランジュ乗数法を用いることにより、
  $$V\ \alpha = \lambda\ \alpha \quad -(2)$$
  になる。Ｖは分散共分散行列、 λは固有値、 α固有ベクトルである。
- 式（２）の固有値は分散に相当する。分散の最大のものが楕円の長径なので、固有値が最大のものを求める。
- 一般にｎ個のデータに対しても、同様に拡張できる。

主成分分析 ③主成分得点

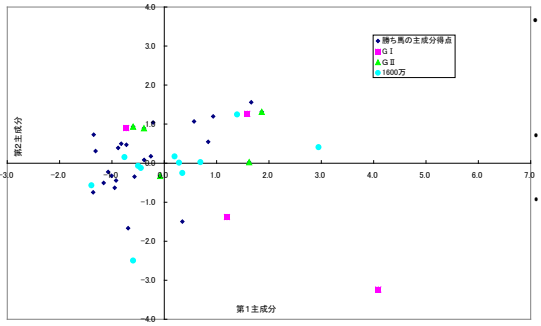
- ｎ個の変数に対し
  $$z_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{n1}x_2 \quad -(3)$$
  のことを第１主成分とよび、個々の観測値に当てはめた時に得られる値
  $$z_{1\,i} = a_{11}x_{1i} + a_{21}x_{2i} + \cdots + a_{n1}x_{2i} \quad -(4)$$
  を主成分得点という。
- これによって新しい変数を導き出す。
- この時、固有値と固有ベクトルはｎ個出てくる。その中でもウエイトの高いものを見つけるために寄与率を調べ、変数の数を減らす。

主成分分析 ④寄与率と累積寄与率

- ｎ個の固有値から、ｉ番目の固有値の寄与率を
  $$\mu_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} * 100 \quad -(5)$$
  とする。
- 累積寄与率を
  $$\nu_i = \sum_{k=1}^{i} \mu_i = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} * 100 \quad -(6)$$
  とする。これはｉ個の固有値が全体の何パーセントを占めるか説明できる。
- 累積寄与率の値を、あらかじめ決めておきその値を超えたら主成分と決める。

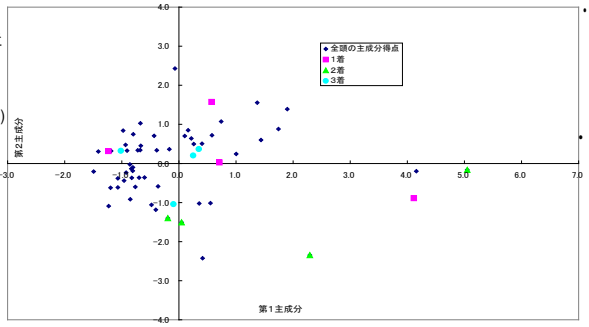
実際のデータで分析する

- 右回り、左回り、中山競馬場、ＧＩ出走の成績から勝率を出す。
- ４種のデータは性質が異なるが、平均と分散を共通になるように標準化した。
- 標準化は
  $$y_{1i} = \frac{x_{1i} - \overline{x_i}}{\sqrt{Var(x_1)}}$$
  で求まる。
- ここでは累積寄与率を８０％として、必要な成分数を与えた。



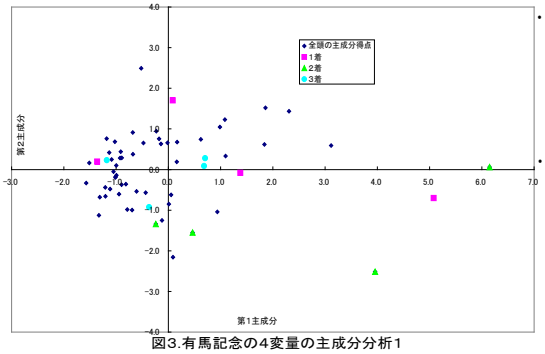
- 中山競馬場２５００ｍで、勝った馬を主成分分析した分布図。
  ２成分で８２％
  第１主成分　総合力
  第２主成分　利き手 (右、左)
- 負けた馬がいないので、勝った馬の力関係を図示している。
- レベルの高いレースほど、第１主成分が高くなっているが、いくつか例外もある。

図1.中山競馬場の勝ち馬の主成分分析



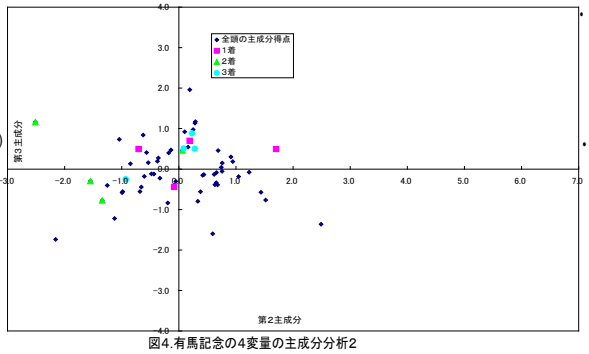
- 有馬記念で走った馬６１頭の右回りと左回りと中山成績の成績を、第１主成分と第２主成分で図示している。
  ２成分で８４％
  第１主成分　総合力
  第２主成分　利き手 (右、左)
- グラフの左上の点の馬が勝つ確率が高いと考える。

図2.有馬記念の3変量の主成分分析



- 有馬記念で走った馬６１頭の右回りと左回りと中山成績とＧＩ出走の成績を、第１主成分と第２主成分で図示している。
  ３変量で９０％
  第１主成分　総合力
  第２主成分　利き手 (右、左)
- ＧＩという強さの指標を入れたので３変量のグラフより第１主成分の方向に広がった。

図3.有馬記念の４変量の主成分分析1



- 有馬記念で走った馬６１頭の右回りと左回りと中山成績の成績を、第２主成分と第３主成分で図示している。
  第２主成分　利き手 (右、左)
  第３主成分　右回りＧＩの強さ
- 右回りのＧＩが強くても３着以内に入らない傾向がある。

図4.有馬記念の４変量の主成分分析2

まとめ

- ＧＩではほとんどレベルが同じ馬が走っている為、図が原点によりやすい。
- 第２主成分で、右回りが得意、左回りが得意よりも、両方得意な軸辺りの点の方が３着に入りやすかった。
- 有馬記念の分布図より、第１、第２成分がプラスの馬が勝ちやすいと考えたが必ずしもそうでないとわかった。
- 合成変量の寄与率がすでに高ければ、ほとんど新たな結果が得られなかった。これは主成分分析の効果が見られていると考える。
- 現段階で、どの馬券を買ったら良いかの判断は、まだ尚早である。

問題点と今後の方針

- 走った回数が０回の馬と、走った事はあるが、勝った事の無い馬を同じ扱いにしているので、走った事のある方に重みをもたし議論したい。
- 勝率のみでしていたので、３着以内率を出して議論していきたい。
- これまでの結果に騎手の要素を増やして、分析を続けていきたい。
- データ数の不足を感じるので、データの数を増やしていきたい。
- 各データに重みを加えるか、他の因子分析を試み、より確度の高い予想が可能か調べる。

2008/11