主成分分析†

講師: 長倉大輔

† この資料は私のゼミにおいて使用するために作成した資料です。ゼミのWEBページ上で公開しており、自由に参照して頂いて構いません。ただし、内容について、一応検証してありますが、もし間違いがあった場合でもそれによって生じるいかなる損害、不利益について責任を負いかねますのでご了承ください。間違いは発見次第、継続的に直していますが、まだ存在する可能性があります。間違いを見つけた場合 nagakura@z7.keio.jp までご連絡いただけると大変嬉しく思います。

内容

1. 準備
2. 主成分の定義
3. 主成分の計算
4. 分析例、主成分の解釈
5. 主成分分析その他

1. 準備

ここでは本スライドの内容に関連した事柄を復習する。

■ 固有値と固有ベクトル

$n \times n$ 正方行列 $\mathbf{A}$ の固有値と固有ベクトルとは関係式 $\mathbf{Ab} = \lambda \mathbf{b}$ を満たす $n \times 1$ ベクトル $\mathbf{b}$ と スカラー λ を言う。
$n \times n$ 正方行列 $\mathbf{A}$ に固有値は n 個存在し、行列が正則であるための必要十分条件は全ての固有値が 0 ではないことである。

■ 実対称行列(行列の要素が実数)の固有値

$n \times n$ 実対称行列 $\mathbf{A}$ の固有値は全て実数で、異なった固有値に対応する固有ベクトルは直交している。

1. 準備

■ 正値定符号行列

$n \times n$ 対称行列 $\mathbf{A}$ で 任意の $n \times 1$ ベクトル $\mathbf{x} \neq \mathbf{0}$ に対して $\mathbf{x}'\mathbf{A}\,\mathbf{x} > 0$ を満たすものを、正値定符号行列という。正値定符号行列の固有値は全て正である。

2. 主成分の定義

いくつかの教科書、資料を読むと、主成分分析とは「**多変量のデータの特徴を要約する**」もの、または「**多変量のデータの情報を残しつつ次元を縮約する方法**」というような書き方をよくしてある。

データの特徴の要約とは、例えば、**平均**なら**データの中心**、**分散**なら、データの**平均からの広がり具合**、を要約している、と言える。

主成分分析もそのようなものとして考えられる。

2. 主成分の定義

主成分分析では、データが与えられたときに、**主成分**というものを計算する。以下ではまず、この主成分の定義を述べる。

データとして、D 種類のデータが N 個あるとする。
これを

$$x_{id}, d = 1,\ldots, D,\ \ i = 1,.,..,N,$$

と表すとする。

2. 主成分の定義

このデータに対して、$H\,(\leq D)$ 種類の**主成分**

$$y_i^{(h)},\ h = 1, \ldots, H, \quad i = 1, \ldots, N$$

は以下のように定義される。

まず $y_i^{(h)}$ は

$$y_i^{(h)} = \beta_1^{(h)} x_{i1} + \beta_2^{(h)} x_{i2} + \ldots + \beta_D^{(h)} x_{iD}$$

と表されるとする。ここで $\beta_d^{(h)}$ は x_{id} , $d = 1, \ldots, D$ に対する重みを表す係数。この係数のベクトルを

$$\boldsymbol{\beta}^{(h)} = [\beta_1^{(h)}, \beta_2^{(h)}, \ldots, \beta_D^{(h)}]'$$

とする。

2. 主成分の定義

この時、$\boldsymbol{\beta}^{(h)}$は以下の問題の解として定義される。*

$$\max_{\{\boldsymbol{\beta}^{(h)}\}_{h=1}^{H}} \sum_{h=1}^{H} \mathrm{var}(y_i^{(h)})$$

$$\text{subject to} \quad \boldsymbol{\beta}^{(h)\prime}\boldsymbol{\beta}^{(k)} = \begin{cases} 1 & \text{if } h = k, \\ 0 & \text{if } h \neq k, \end{cases}$$

$$\text{and} \quad \mathrm{var}(y_i^{(1)}) \geq \mathrm{var}(y_i^{(2)}) \geq \cdots \geq \mathrm{var}(y_i^{(H)})$$

とする。ここで

$$\mathrm{var}(y_i^{(h)}) = N^{-1}\sum_{i=t}^{N}(y_i^{(h)} - \bar{y}^{(h)})^2, \quad \bar{y}^{(h)} = N^{-1}\sum_{i=1}^{N} y_i^{(h)}$$

である。つまり主成分の**分散(の和)が最大になるように**重みベクトル $\boldsymbol{\beta}^{(h)}$ を互いに直行するように決定する。このように $\boldsymbol{\beta}^{(h)}$ を決めた時、(分散が大きい順に) $y_i^{(1)}$を**第1主成分**、$y_i^{(2)}$を**第2主成分**と呼ぶ。

* この定義はあまり一般的でないと思うが、得られる結果は同じ。

3. 主成分の計算

ここでは、この重みベクターは x_{ik} の**分散共分散行列**の**固有ベクトル**として与えられることを確認する。

まず、$\mathrm{var}(y_i(h))$ は $y_i(h)$ の定義より

$$\mathrm{var}(y_i^{(h)}) = N^{-1}\sum_{i=1}^{N}(y_i^{(h)} - \bar{y}^{(h)})^2$$

$$= N^{-1}\sum_{i=1}^{N}(\beta_1^{(h)} x_{i1}^{*} + \cdots + \beta_D^{(h)} x_{iD}^{*})^2,$$

と書き換えられる。ここで

$$x_{id}^{*} = x_{id} - \bar{x}_d,\ \ \bar{x}_d = N^{-1}\sum_{i=1}^{N} x_{id}\ , d = 1, \ldots, D$$

である。

3. 主成分の計算

$\mathbf{x}_i^* = [x_{i1}^*, x_{i2}^*, \ldots, x_{iD}^*]'$ とすると、

$$\beta_1^{(h)} x_{i1}^* + \ldots + \beta_D^{(h)} x_{iD}^* = \boldsymbol{\beta}^{(h)\prime} \mathbf{x}_i^*$$

と表せるので(これは**スカラー**であることに注意）、
$\mathrm{var}(y_i^{(h)})$は

$$\mathrm{var}(y_i^{(h)}) = N^{-1} \sum_{i=1}^{N} (\boldsymbol{\beta}^{(h)\prime} \mathbf{x}_i^*)^2$$

$$= N^{-1} \sum_{i=1}^{N} \boldsymbol{\beta}^{(h)\prime} \mathbf{x}_i^* \mathbf{x}_i^{*\prime} \boldsymbol{\beta}^{(h)}$$

$$= \boldsymbol{\beta}^{(h)\prime} \boldsymbol{\Sigma}_x \boldsymbol{\beta}^{(h)}$$

と表せる。ここで $\boldsymbol{\Sigma}_x = N^{-1} \sum_{i=1}^{N} \mathbf{x}_i^* \mathbf{x}_i^{*\prime}$ は x_{ik} の分散共分散行列。$\boldsymbol{\Sigma}_x$ は通常、**正値定符号行列。**

3. 主成分の計算

先ほどと似ているが微妙に違う、次の最大化問題を考えよう。

$$\max_{\boldsymbol{\beta}^{(h)}} \ \mathrm{var}(y_i^{(h)}) \quad \text{subject to} \quad \boldsymbol{\beta}^{(h)\prime}\boldsymbol{\beta}^{(h)} = 1$$

前ページの表現を使うと

$$\max_{\boldsymbol{\beta}^{(h)}} \ \boldsymbol{\beta}^{(h)\prime}\boldsymbol{\Sigma}_x\boldsymbol{\beta}^{(h)} \quad \text{subject to} \quad \boldsymbol{\beta}^{(h)\prime}\boldsymbol{\beta}^{(h)} = 1$$

と書き直せる。この問題をラグランジュ乗数法を用いて解いてみよう。ラグランジュアン

$$L = \boldsymbol{\beta}^{(h)\prime}\boldsymbol{\Sigma}_x\boldsymbol{\beta}^{(h)} + \lambda_h(1 - \boldsymbol{\beta}^{(h)\prime}\boldsymbol{\beta}^{(h)})$$

の1階の条件は、

3. 主成分の計算

$$\frac{\partial L}{\partial \beta^{(h)}} = 2\boldsymbol{\Sigma}_x \boldsymbol{\beta}^{(h)} - 2\lambda_h \boldsymbol{\beta}^{(h)} = \mathbf{0} \Leftrightarrow \boldsymbol{\Sigma}_x \boldsymbol{\beta}^{(h)} = \lambda_h \boldsymbol{\beta}^{(h)}$$

および $$\frac{\partial L}{\partial \lambda_h} = 1 - \boldsymbol{\beta}^{(h)\prime} \boldsymbol{\beta}^{(h)} = 0 \Leftrightarrow \boldsymbol{\beta}^{(h)\prime} \boldsymbol{\beta}^{(h)} = 1$$

となる。最初の式は **$\boldsymbol{\beta}^{(h)}$の解は $\boldsymbol{\Sigma}_x$の固有値 λ_h に対する固有ベクトル**であることを示しており、次の式によってその固有ベクトルの長を 1 に基準化している。また、最大化された $\mathrm{var}(y_i^{(h)}) = \boldsymbol{\beta}^{(h)\prime} \boldsymbol{\Sigma}_x \boldsymbol{\beta}^{(h)}$ は、これらの条件を代入すると、

$$\boldsymbol{\beta}^{(h)\prime} \boldsymbol{\Sigma}_x \boldsymbol{\beta}^{(h)} = \lambda_h \boldsymbol{\beta}^{(h)\prime} \boldsymbol{\beta}^{(h)} = \lambda_h$$

であるので 固有値 λ_h に等しいことがわかる。

3. 主成分の計算

さらに分散共分散行列 Σ_x は**正値定符号行列**であるので、その固有値は全て正であり、対応する固有ベクトルは全て互いに直交するので、この解は

$$\boldsymbol{\beta}^{(h)\prime}\boldsymbol{\beta}^{(k)} = 0 \quad for \;\; h \neq k$$

という制約も自動的に満たす。よって固有値の大きい順に、それら固有値に対応する固有ベクトルを、$\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$, …, $\boldsymbol{\beta}^{(H)}$ とすれば、第1 主成分から第 H 主成分の係数ベクトルが求まり、主成分 $y_i^{(h)}$, h=1, …, H が計算できる。(これはやや乱暴な議論で本当はもう少し丁寧な議論が必要だが結果は同じ)。

3. 主成分の計算

また、$h \neq k$ に対して、$y_i^{(h)}$ と $y_i^{(k)}$ の(標本の)共分散は

$$\begin{aligned}
\mathrm{cov}(y_i^{(h)}, y_i^{(k)}) &= N^{-1} \sum_{i=1}^{N} (y_i^{(h)} - \bar{y}^{(h)})(y_i^{(k)} - \bar{y}^{(k)}) \\
&= N^{-1} \sum_{i=1}^{N} \boldsymbol{\beta}^{(h)\prime} \mathbf{x}_i^{*} \mathbf{x}_i^{*\prime} \boldsymbol{\beta}^{(k)} \\
&= \boldsymbol{\beta}^{(h)\prime} \boldsymbol{\Sigma}_x \boldsymbol{\beta}^{(k)} \\
&= \lambda_k \boldsymbol{\beta}^{(h)\prime} \boldsymbol{\beta}^{(k)} \\
&= 0
\end{aligned}$$

であるので(最後から2番目の等式は $\boldsymbol{\beta}^{(k)}$が $\boldsymbol{\Sigma}_x$の固有値 λ_k の固有ベクトルであることから得られる)、主成分どうしは無相関であることがわかる。

4. 分析例、主成分の解釈

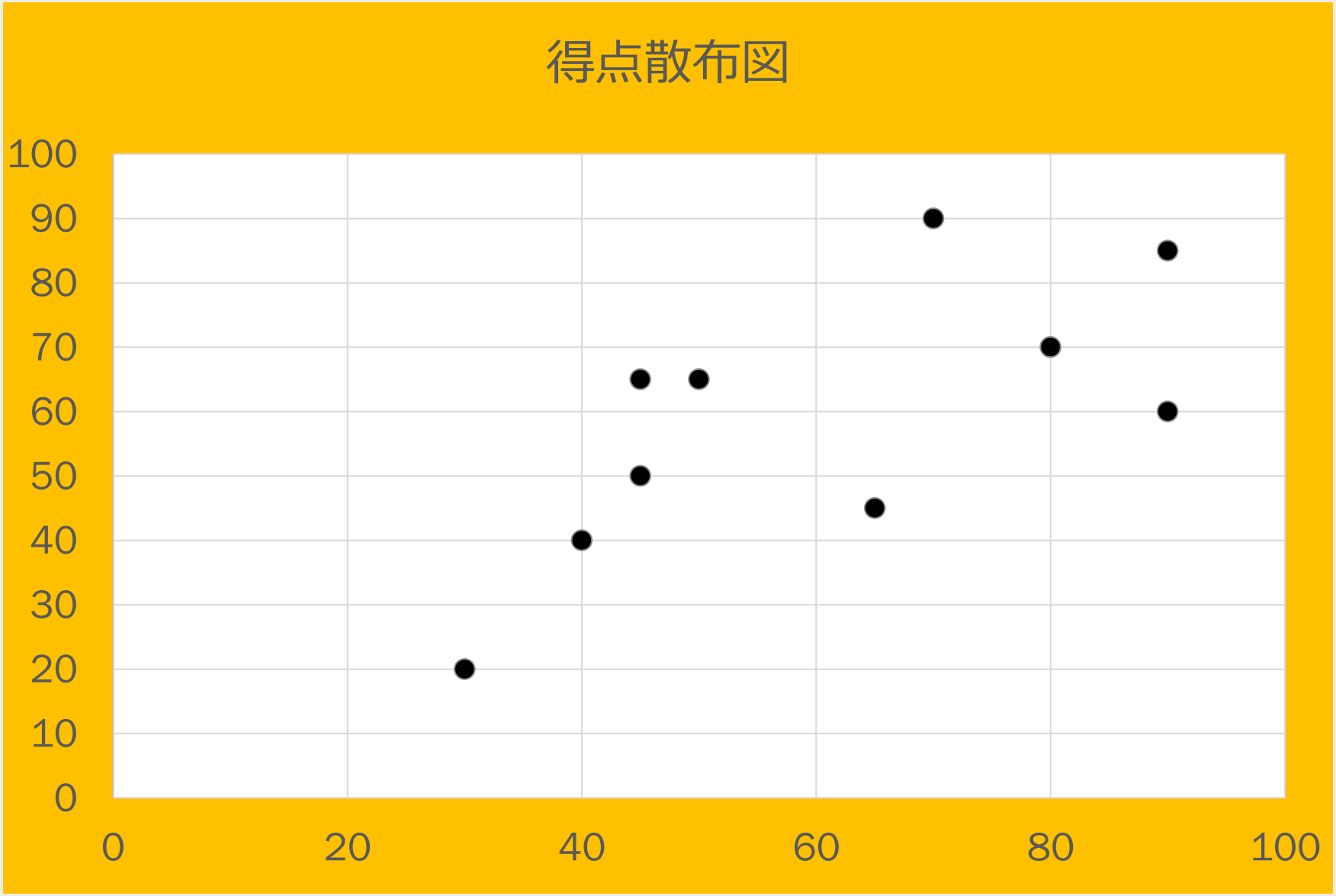
ここでは実際に主成分を計算してみて、それがどのように解釈されるか考える。

データとして、以下の数学と国語の10人分の試験の点数を考えてみよう（架空のデータ）

学生番号(i)	1	2	3	4	5	6	7	8	9	10
数学 (x_{i1})	50	40	45	70	90	45	30	65	80	90
国語 (x_{i2})	65	40	65	90	60	50	20	45	70	85

それぞれ数学の平均は 60.5 で 標準偏差は 20.4
国語の平均は 59 で標準偏差は 20.0である。
また相関は 0.69、散布図は

4. 分析例、主成分の解釈



4. 分析例、主成分の解釈

このデータに対して、主成分を求めてみると、第1主成分、第2主成分について、それぞれ

$$y_i^{(1)} = 0.7185103\, x_{i1} + 0.6955164\, x_{i2}$$
$$y_i^{(2)} = -\,0.6955164\, x_{i1} + 0.7185103\, x_{i2}$$

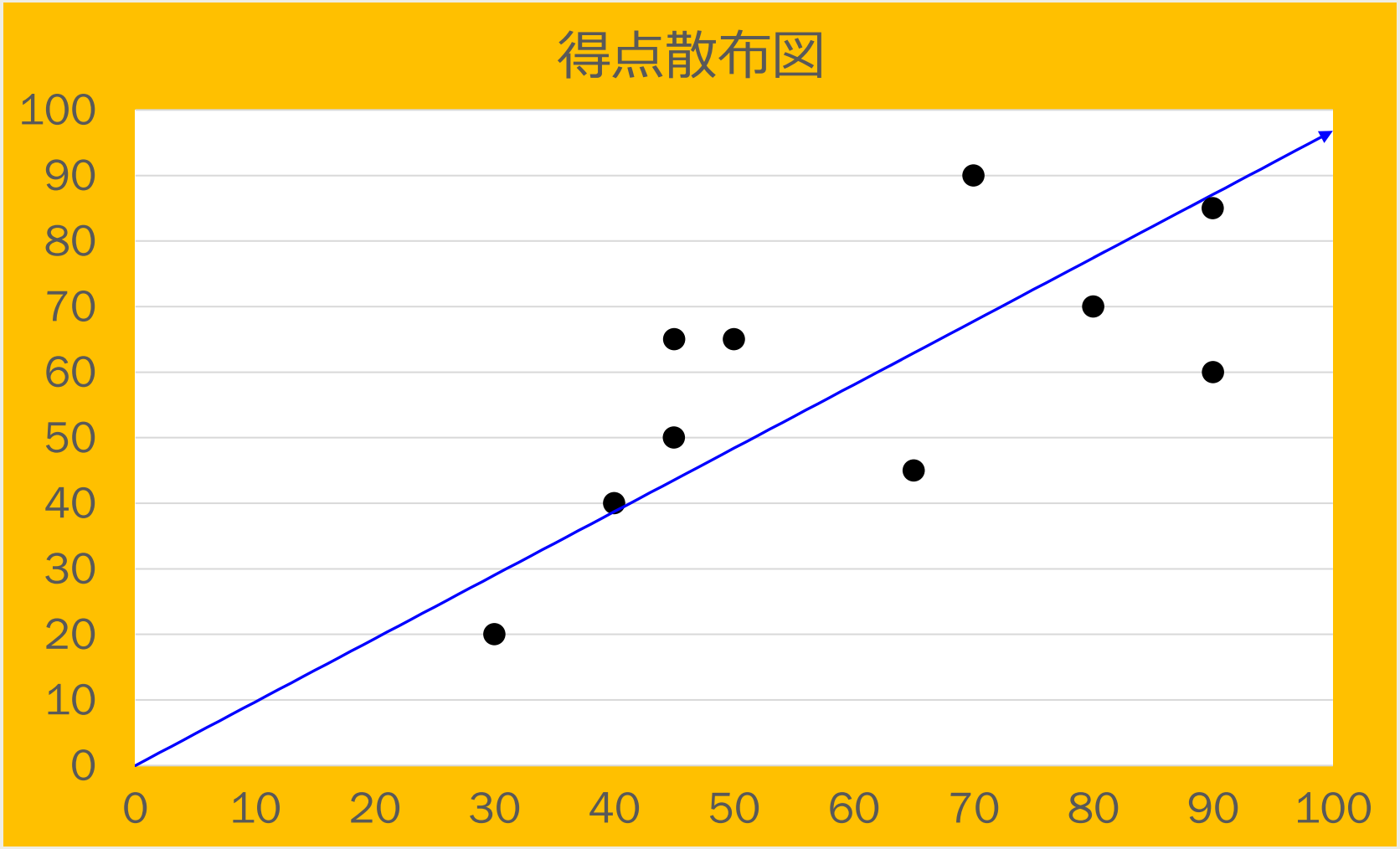
という式が求められ、それぞれの値は以下のようになった(小数点以下第1位で四捨五入している)。

学生番号(i)	1	2	3	4	5	6	7	8	9	10
第1主成分 ($y_i^{(1)}$)	81	57	78	113	106	67	35	78	106	124
第2主成分 ($y_i^{(2)}$)	12	1	15	16	–19	5	– 6	–13	– 5	– 2

この時、第1 主成分の分散は 668.7734 、第2主成分の分散は 127.4766である。さて、この**主成分の値**は何を表しているのだろうか？

4. 分析例、主成分の解釈

先ほどの散布図に第 1 主成分の**係数ベクトル** を書き入れた図は以下のようになる。

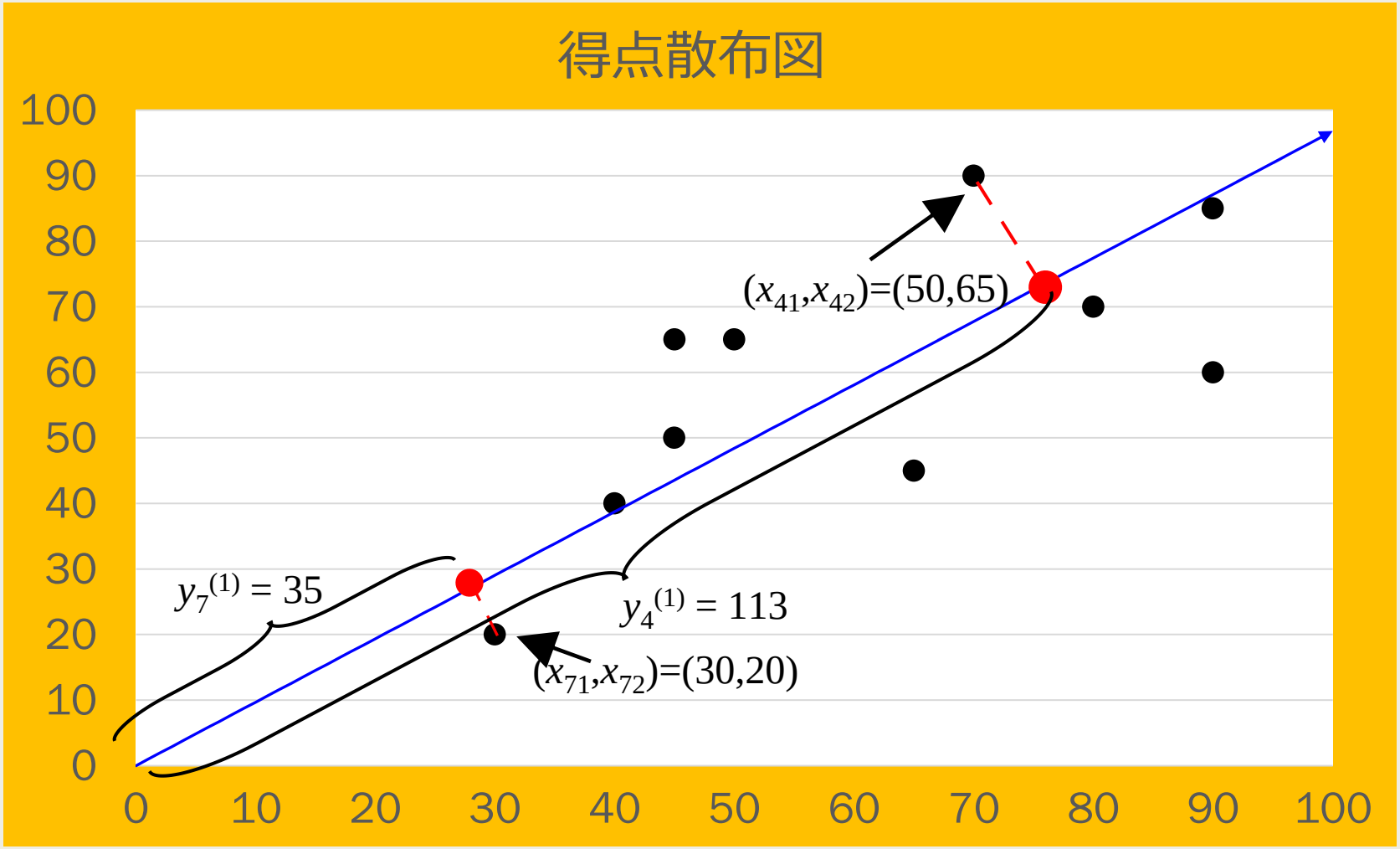


4. 分析例、主成分の解釈

この時、第1主成分 $y_i^{(1)}$ の値は、係数ベクトル (0.7185103, 0.6955164) と、データベクトル (x_{i1}, x_{i2}) の**内積**であるので、この図において、それぞれの点からこの直線に**垂線**を下した時のこの直線との交点と原点の距離に等しくなる(ベクトル $\vec{\mathbf{a}}$ と ベクトル $\vec{\mathbf{b}}$ の内積は $\vec{\mathbf{a}} \cdot \vec{\mathbf{b}} = |\vec{\mathbf{a}}||\vec{\mathbf{b}}| \cos\theta$ なので。ここで θ はこの2つのベクトルのなす角)。この結果は 係数ベクトルの長さを 1 に基準化したことに依存していることに注意。

例えば、$(x_{41}, x_{42}) = (70, 90)$ と $(x_{71}, x_{72}) = (30, 20)$ とに対応する第1主成分の値はそれぞれ $y_4^{(1)} = 113$と $y_7^{(1)} = 35$ であるが、これらは図の上では

4. 分析例、主成分の解釈



に対応している。ちなみに第2主成分は**この垂線の長さ**（符号付き）であることも同様の議論でわかる。

４分析例、主成分の解釈

主成分の幾何学的な意味付けはわかったが、ではこの値は一体何を意味しているのだろうか？

この値の意味は基本的に**分析者が恣意的に解釈するしかない**。例えば、先ほどの第1主成分の値を見ると、基本的に数学および国語の得点に比例して増えて行くことから、「総合的な頭の良さ」と解釈できるかもしれない。

また第2主成分は国語の成績が数学に比べて相対的に高いときに値が大きくなる傾向があるので、「文系的な頭の良さ」と解釈できるかもしれない。

5. 主成分分析その他

ここでは主成分分析に関連したいくつかの用語について説明する。

寄与率 :

第 h 主成分の寄与率とは、第 h 主成分の分散の、全ての主成分の分散の和への比率 (データが D種類の場合、最大 D 種類の主成分が得られるので この D 個の主成分の分散の合計)。第 h 主成分の分散を λ_h とすると、

$$\text{第 } h \text{ 主成分の寄与率} = \frac{\lambda_h}{\sum_{j=1}^{D} \lambda_j}$$

と計算される。

5. 主成分分析その他

定義により、寄与率は第 1 主成分が最も大きく、次いで、第2、第3 と単調に徐々に減少していく。

累積寄与率 :
寄与率を第 1 主成分から順に累積したもの。

主成分負荷量ベクトル:
重みベクトル $\beta^{(h)}$ のことを第 h 主成分の主成分負荷量ベクトルという。

例) 先ほどの得点の例だと、第1主成分の寄与率はおおよそ 0.84 である。この値をもって、この場合、第1主成分はデータの84%を説明している、などという。

5. 主成分分析その他

まとめと注意点

(1) 主成分は符号には意味がなく、また、相対的な位置関係にのみ意味がある。これは主成分の分散を最大にするように主成分負荷量ベクトルを決定しているが、分散はマイナスをかけても同じなので、負荷量ベクトルにマイナスをかけたものも解になることからもわかる(そもそも固有ベクトルだから符号には意味がないが)。また、長さを 1 に基準化することによって、幾何学的な意味が付与されているが、識別のためには別の(合理的な)基準化を用いてもよく、その場合、値の絶対値が変わる(相対的な位置関係は変わらない)ので、基本的に値の大きさもあまり意味はない (**と思うがどうなんだろう？**)

5. 主成分分析その他

[(1)の続き] 例えば、学生の得点の例で、第 2 主成分を「文系的な頭の良さ」としたが、マイナスをかけたものを用いるのであれば、「理系的な頭の良さ」と解釈しなおせばよい。

(2) 平均や分散と違い、主成分はその解釈が明確でない(**と思うがどうなんだろう？**)。

(3) 主成分の数をいくつまで増やすかという問題について、実際の分析においては累積寄与率が80%を超えるくらいまで、主成分の数を増やすよう。

5. 主成分分析その他

(4) 本スライドの話は以下の図に要約される(2次元)。