線形代数で眺める確率統計

@minami106

2013 年 2 月 4 日

確率統計では, データの特徴を代表して表すさまざまな指標が出現します. それらはわりとごちゃごちゃしためんどくさい式によって定義されるのですが, ベクトルや, ベクトルの演算を用いると, びっくりするくらい簡潔かつ明快にいろんなことが記述できることにふと気づいたので, まとめてみたいと思います.

1　いろんなベクトルの定義

実際に観測されたデータを $x_1, x_2, \cdots, x_n (\in \mathbb{R})$ とし, データから求めた算術平均を $\overline{x}$ とします.

定義 1.1. データベクトル

$\boldsymbol{x} \in \mathbb{R}^n$ を次のように定義し, **データベクトル (data vector)** と呼ぶ.

$$\boldsymbol{x} \stackrel{\text{def}}{=} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

定義 1.2. 平均ベクトル

$\overline{\boldsymbol{x}} \in \mathbb{R}^n$ を次のように定義し, **平均ベクトル (mean vector)** と呼ぶ. $\overline{x}$ は, データから求めた算術平均である.

$$\overline{\boldsymbol{x}} \stackrel{\text{def}}{=} \begin{pmatrix} \overline{x} \\ \overline{x} \\ \vdots \\ \overline{x} \end{pmatrix}$$

定義 1.3. 偏差ベクトル

$\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ を次のように定義し, **偏差ベクトル (deviation vector)** と呼ぶ.

$$\tilde{\boldsymbol{x}} \stackrel{\text{def}}{=} \boldsymbol{x} - \overline{\boldsymbol{x}} = \begin{pmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix}$$

2　標本分散, 標本標準偏差

定義 2.1. 標本分散

次のように定義される $V_{\boldsymbol{x}}$ を, $\boldsymbol{x}$ の**標本分散 (sample variance)** という.

$$V_{\boldsymbol{x}} = \frac{1}{n}||\tilde{\boldsymbol{x}}||^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

偏差ベクトルは, **データベクトルが平均（ベクトル）からどの向きにどのくらいずれているか？**ということを表すベクトルと捉えることができます. このことから, データの標本分散というのは, データベクトルの、平均ベクトルからの**ズレの大きさ**を, **偏差ベクトルのノルム**という形で測ったものに他ならないということが分かります (図 1) [*1].

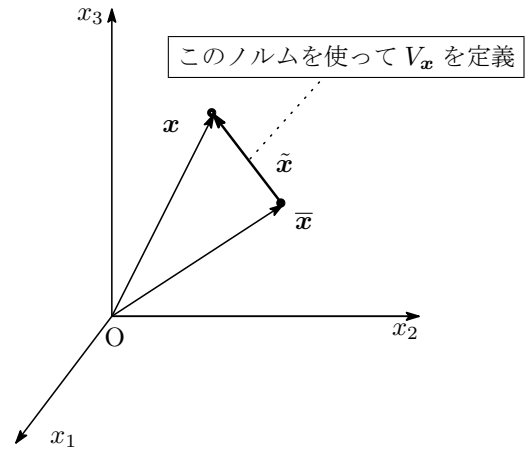


図 1　標本分散の定義のイメージ (データ数 $n = 3$ の場合)

定義 2.2. 標本標準偏差

$\sigma_{\boldsymbol{x}} \stackrel{\text{def}}{=} \sqrt{V_{\boldsymbol{x}}}$ と定義し, $\sigma_{\boldsymbol{x}}$ を**標本標準偏差**と呼ぶ.

なんか, この, ノルムによる定義 (ノルムは自分自身との内積でかけるので, 内積を使った定義でもある) に, ベクトルの内積の性質を適用して, 標本分散や標本標準偏差に関する性質をいろいろ導き出せないですかね. やってみたら面白いかもですね.

[*1] データ数 n で $||\tilde{\boldsymbol{x}}||^2$ を割っているのは, データ数をたくさんにするとノルムが際限なく増えてしまうので, 正規化するためにかけている定数です. データ数で割ることにより, **ひとつのデータあたり, 大体どのくらい平均からズレているか**を考えていると思うとわかりやすいかもです.

3　標本共分散, 相関係数

今度は, ふたつのデータ列 $x_1, \cdots, x_n \in \mathbb{R}, y_1, \cdots, y_n \in \mathbb{R}$ を考えます. ふたつのデータ列[*2]に対して, それぞれのデータベクトルを $\boldsymbol{x}, \boldsymbol{y}$, 偏差ベクトルを $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$ と置いた時, **標本共分散**を定義することができます.

定義 3.1. 標本共分散

次のように定義される $\mathrm{Cov}\,(\boldsymbol{x}, \boldsymbol{y})$ を, $\boldsymbol{x}, \boldsymbol{y}$ の**標本共分散 (sample covariance)** と呼ぶ

$$\mathrm{Cov}\,(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\mathrm{def}}{=} \frac{1}{n}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

定義 3.2. 相関係数

次のように定義される $r(\boldsymbol{x}, \boldsymbol{y})$ を, $\boldsymbol{x}, \boldsymbol{y}$ の**相関係数 (correlation coefficient)** と呼ぶ.

$$r(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\mathrm{def}}{=} \frac{(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})}{||\tilde{\boldsymbol{x}}||||\tilde{\boldsymbol{y}}||} = \frac{\mathrm{Cov}\,(\boldsymbol{x}, \boldsymbol{y})}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}} = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \cos\theta.$$

ただし, θ は, $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$ が $\mathbb{R}^n$ でなす角である.

まず, $\cos\theta$ の性質から, 相関係数の次の性質がすぐに分かります.

命題 3.1. (相関係数の値)

$$-1 \leq r(\boldsymbol{x}, \boldsymbol{y}) \leq 1.$$

さて, 相関係数がどんなものかってことは, 統計学の授業なんかで何となく習ったことがある方も多いのではなかろうかと思いますが, なぜそういう性質を持つのかってことは知らない人が結構いるのではないかと. 実は, 相関係数の意味は, **相関係数は偏差ベクトルのなす角 θ の余弦 $\cos\theta$ であるという視点に立つと**, 相関係数とは何たるかがものすごくスッキリわかるのです.

3.1　相関係数の意味

さて, 偏差ベクトルは, **データベクトルが平均どっち向きに, どのくらいずれているか**を表すベクトルでした (図 2). いま, 偏差ベクトル $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$ と, なす角の余弦 $\cos\theta$ について, 図 3 のような関係が成立します.

図 3 のように,

- 偏差ベクトルどうしが**同じ向き** のとき, $r = \cos\theta$ **は最大 (+1)**.
- 偏差ベクトルどうしが**そっぽ向き**のとき, $r = \cos\theta$ **は 0**.
- 偏差ベクトルどうしが**逆向き**のとき, $r = \cos\theta$ **は最小 (-1)**,

[*2] たとえば, x_i が出席番号 i 番の学生の国語の点数, y_i が数学の点数だったりするわけです.

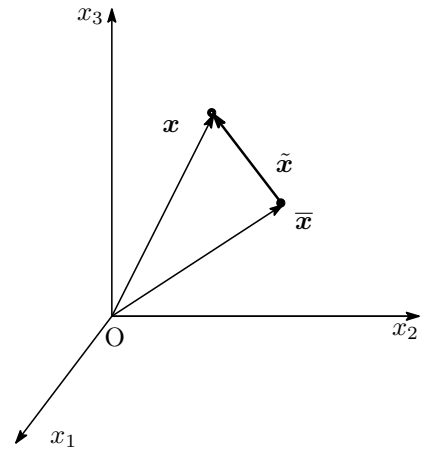
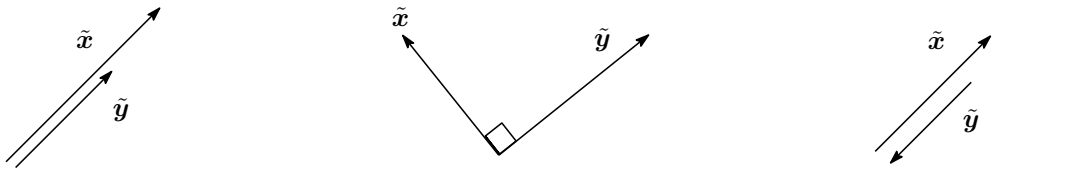


図 2　偏差ベクトル ($\tilde{\boldsymbol{y}}$ についても同様のイメージ)



$\theta = 0$
$\Rightarrow \cos\theta = +1$(最大)

$\theta = \pi/2$
$\Rightarrow \cos\theta = 0$

$\theta = \pi$
$\Rightarrow \cos\theta = -1$(最小)

図 3　$\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$ の向きの関係と $\cos\theta$ の値

ということが言えます. つまり $\cos\theta$ は, ベクトルが**どのくらい同じ向きを向いているか**の指標といえるのです. さて, これから, 相関係数の 3 つの値のパターンについて, データはどのような挙動を見せるのか, 考えてみましょう.

1. $0 < r \leq +1$ のとき（**相関係数が正**）.
   このとき, $\cos\theta > 0$ なので, 偏差ベクトル $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$ は, **比較的同じ方向を向いている**ということが言えます. つまり, $\boldsymbol{x}$ **が平均からズレたのと大体同じ方向に,** $\boldsymbol{y}$ **もズレる傾向がある**ということです (図 4). こんな傾向にのっとって, データが生起しているということを頭に入れて, データをプロットしまくっ

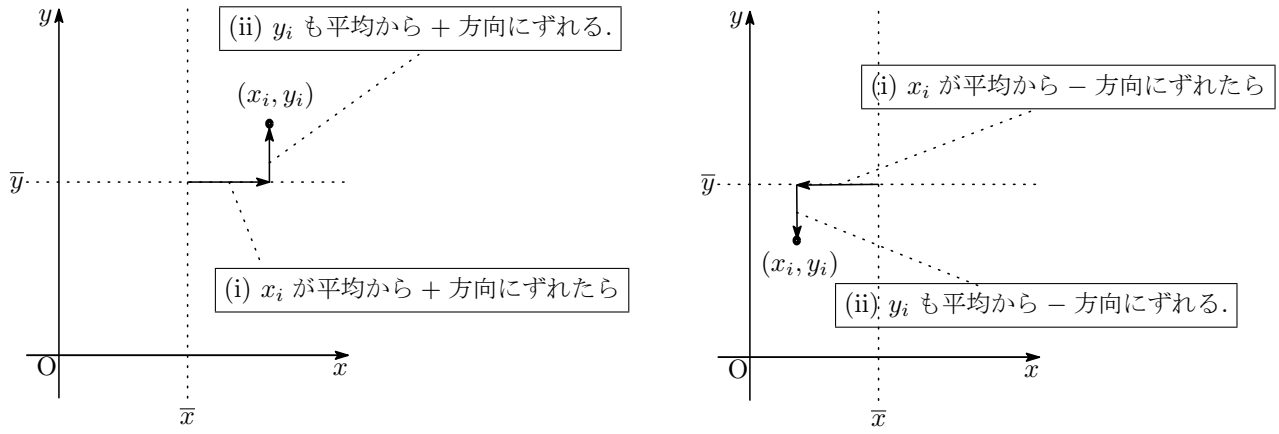


図 4　$r = \cos\theta > 0$ のときのデータの傾向.

てみると, データはおおよそ図 5 のように分布するだろうということが分かります. データが x_i, y_i が

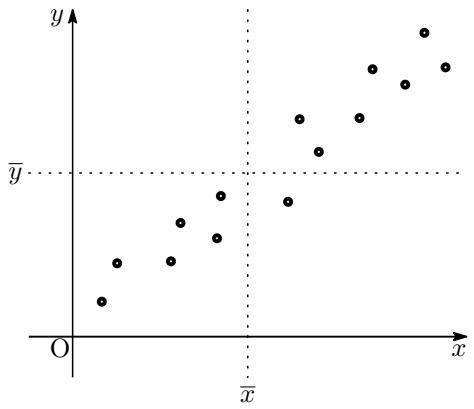


図 5　$r = \cos\theta > 0$ のときのデータの分布の感じ

こんな感じに分布することを, **正の相関**があるといいます.

定義 3.3. 正の相関
$r(\boldsymbol{x}, \boldsymbol{y}) > 0$ のとき, データ $\boldsymbol{x}, \boldsymbol{y}$ は**正の相関 (plus correlation)** を持つといいます.

2. $-1 \leq r < 0$ のとき（**相関係数が負**）.
このとき, $\cos\theta < 0$ なので, 偏差ベクトル $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}$ は, **比較的逆の方向を向いている**ということが言えます. つまり, $\boldsymbol{x}$ **が平均からズレたのと大体逆の方向に,** $\boldsymbol{y}$ **もズレる傾向がある**ということです (図 6). こんな傾向にのっとって, データが生起しているということを頭に入れて, データをプロットしまくっ

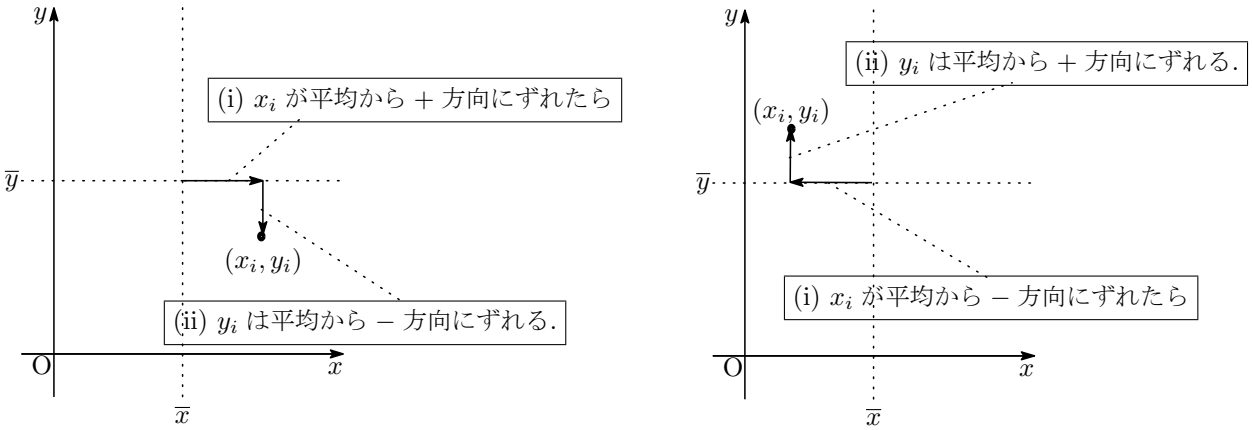


図 6　$r = \cos\theta < 0$ のときのデータの傾向.

てみると, データはおおよそ図 7 のように分布するだろうということが分かります. データが x_i, y_i がこんな感じに分布することを, **負の相関**があるといいます.

定義 3.4. 負の相関
$r(\boldsymbol{x}, \boldsymbol{y}) < 0$ のとき, データ $\boldsymbol{x}, \boldsymbol{y}$ は**負の相関 (minus correlation)** を持つといいます.

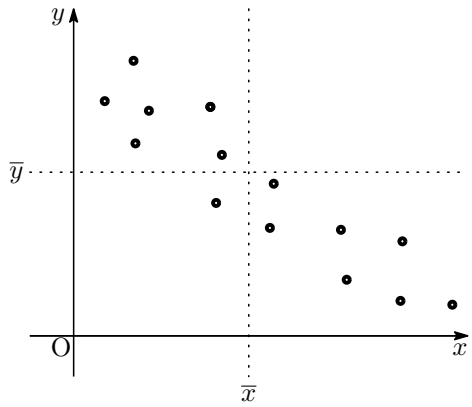


図 7　$r = \cos\theta < 0$ のときのデータの分布の感じ

3. $r = 0$ のとき **(相関係数が 0)**

このとき, $\cos\theta = 0$ ということなので, $\boldsymbol{x}$ がどっち向きに平均からズレようが, $\boldsymbol{y}$ がどっち向きにずれるのかは分かりません. このとき, データの分布は図 8 のようになります.

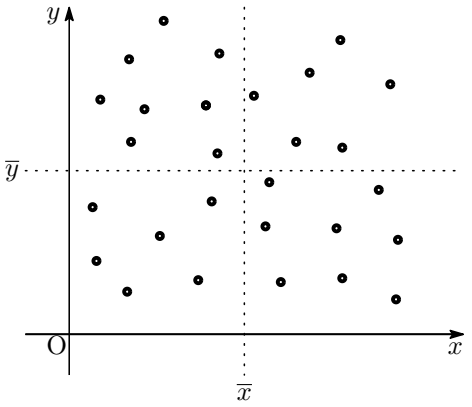


図 8　$r = 0$ のときのデータの分布の感じ.

データが x_i, y_i がこんな感じに分布することを, **相関を持たない**といいます.

定義 3.5. 相関を持たない

$r(\boldsymbol{x}, \boldsymbol{y}) = 0$ のとき, データ $\boldsymbol{x}, \boldsymbol{y}$ は**相関を持たない**といいます.

こんなふうに, 相関係数を**偏差ベクトルのなす角の余弦**と捉えることによって, 相関係数が持つ意味がこんなにもスッキリと見えてしまいました. 以下に, 相関係数の性質をまとめましょう.

命題 3.2. 相関係数の意味

相関係数 $r(\boldsymbol{x}, \boldsymbol{y})$ について,

$$\begin{cases} 0 < r \leq 1 & \Rightarrow \text{データ } \boldsymbol{x}, \boldsymbol{y} \text{ は}\textbf{正の相関}\text{を持つ.} \\ r = 0 & \Rightarrow \text{データ } \boldsymbol{x}, \boldsymbol{y} \text{ は}\textbf{相関を持たない}\text{.} \\ -1 \leq r < 0 & \Rightarrow \text{データ } \boldsymbol{x}, \boldsymbol{y} \text{ は}\textbf{負の相関}\text{を持つ.} \end{cases}$$

こんなふうに, 線形代数的な視点で確率統計で出てくる重要な指標を眺めてみると, 随分明快な理解が得られるのだなあと実感してしまいます. この資料は相関係数のところについて書きたかったという理由で書いてみたものですが, 他にも確率統計を線形代数のレンズで調べてみたときに得られる面白い結果というのは色々あるのでしょうか. 気になるところですね〜. 面白い話があったら是非是非教えてほしいです.

参考

- Wikipedia - 相関係数
  http://ja.wikipedia.org/wiki/%E7%9B%B8%E9%96%A2%E4%BF%82%E6%95%B0