3 主成分分析

次元の大きいデータをうまくまとめられるよう，データの特徴がよくわかる座標軸を作る．

3.1 簡単な説明

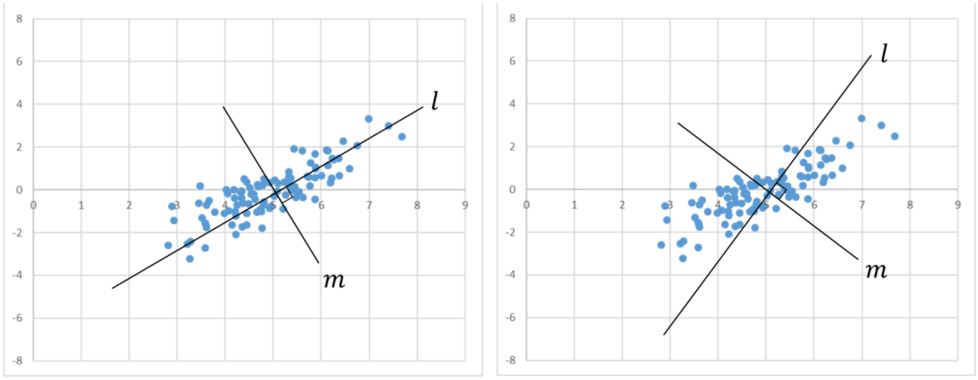


図 6

例えば，データが図 6 の左のようなとき，このデータの特徴は各データの (x, y) 座標という 2 次元の情報を使わなくても，各データが直線 l と m の交点から l に沿って (右上左下方向に) どのくらい離れているかという 1 次元の情報である程度把握できる．したがって，もとのデータの特徴を十分もった新しい 1 次元データを作ることができる．その新しい 1 次元のデータは，各点の l 軸上の座標 (l, m の交点と，各データから l におろした垂線の足の距離) である．このように，うまく直線 l のような直線を選んでそれを座標軸として使えば 2 次元データを 1 次元データに変換できる (図 6 の右はうまく直線を選べなかったときの例)．

3.2 定式化

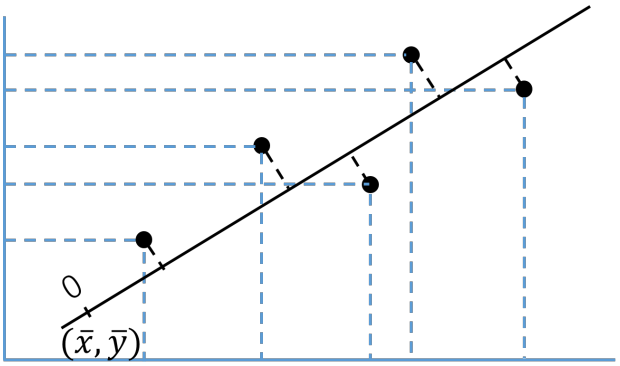
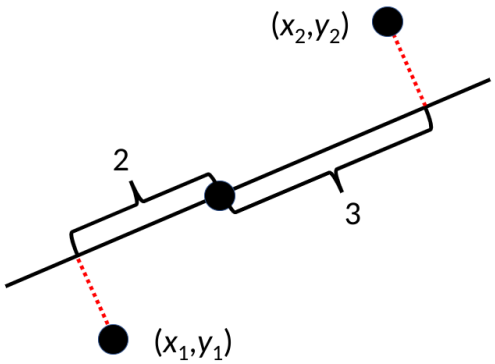


図 7

2 次元のデータ (図 7 の黒点) を直線に射影する．直線は $(\bar{x}, \bar{y})$ を通っているものとする．各データは，その垂線の足と $(\bar{x}, \bar{y})$ との距離の値に置き換わる (1 次元データ)．ただし $(\bar{x}, \bar{y})$ より右が正の値，左が負の値とする．$(\bar{x}, \bar{y})$ の新しい値は 0.



たとえば，上の図ならデータは次のように 1 次元に変換される．

$$(x_1, y_1) \to z_1 = -2, \qquad (x_2, y_2) \to z_2 = 3$$

直線を選ぶ基準はできるだけ「**情報の損失**を最小」にするというもの．失う情報とは，点をずらしてしまった距離だと考えられる (その距離はデータが変わってしまった量なので)．したがって，情報の損失は図 8 の赤い破線の大きさである．

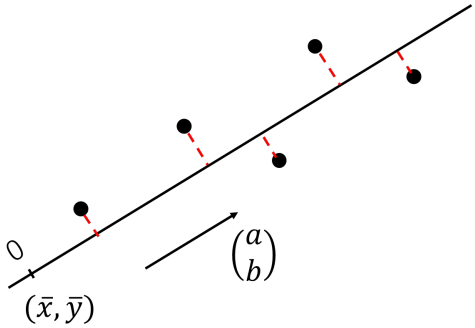


図 8

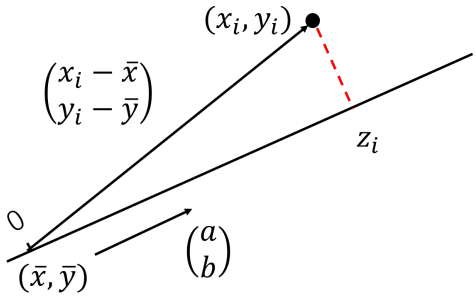


図 9

最適化問題として定式化する．

求める直線の方向ベクトルを $(a, b)^\top$ とする．ただし，ベクトルの長さは 1，つまり，$a^2 + b^2 = 1$ とする．

データ (x_i, y_i) の垂線の足の新しい座標軸での座標 ($(\bar{x}, \bar{y})$ からの距離) を z_i とするとそれは，$a^2 + b^2 = 1$ なので，$(a, b)^\top$ と $(x_i - \bar{x}, y_i - \bar{y})^\top$ という 2 つのベクトルの内積と等しく

$$z_i = a(x_i - \bar{x}) + b(y_i - \bar{y})$$

と計算される．その理由は，2 つのベクトルのなす角を θ とすると，

$$\begin{aligned} z_i &= ((x_i - \bar{x}, y_i - \bar{y})^\top \text{の長さ}) \times \cos\theta \\ &= ((a, b)^\top \text{の長さ}) \times ((x_i - \bar{x}, y_i - \bar{y})^\top \text{の長さ}) \times \cos\theta \\ &= (a, b)^\top \text{と} (x_i - \bar{x}, y_i - \bar{y})^\top \text{の内積} \end{aligned}$$

だから．

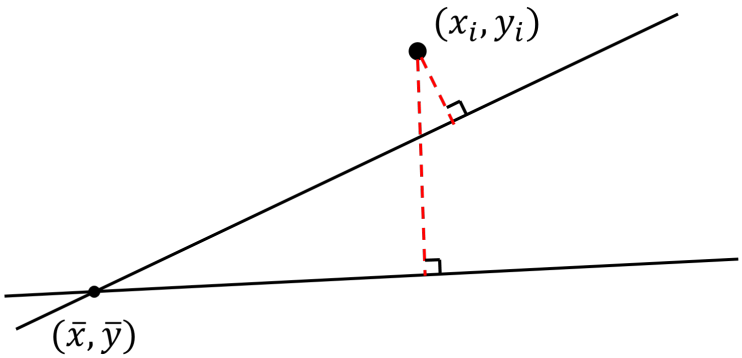


図 10

求めたいのは「情報の損失を最小」にする直線だが，それは「z_i たちの大きさをできるだけ大きく」する直線でもある (ピタゴラスの定理)．図 10 の 2 つの直線をくらべると，傾き

の大きな直線の方が傾きの小さな直線より情報の損失 (赤破線の大きさ) が小さいが，それによって対応する z_i の大きさは逆の大小関係になっている．

そこで「情報の損失の最小化」のかわりに z_i たちの大きさの総和 (であると同時にバラツキ具合)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} z_i^2 &= \frac{1}{n}\sum_{i=1}^{n}\{a(x_i-\bar{x})+b(y_i-\bar{y})\}^2 \\ &= a^2 S_{xx} + 2abS_{xy} + b^2 S_{yy}\end{aligned}$$

を最大化する．ただし，記号の意味は，

$$S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2, \quad S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),$$
$$S_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

次の最大化問題を解けばよい．

$$\begin{aligned}\max_{a,b} \quad & a^2 S_{xx} + 2abS_{xy} + b^2 S_{yy} \\ \text{subject to} \quad & a^2 + b^2 = 1\end{aligned}$$

ラグランジュ乗数法で解く．

ラグランジュ関数は

$$L(a,b,\lambda) = a^2 S_{xx} + 2abS_{xy} + b^2 S_{yy} + \lambda(1-a^2-b^2)$$

それぞれの変数で偏微分したものがゼロになるという連立方程式を解く．

つまり，次を満たす a^*, b^*, λ^* を探す．

$$\begin{aligned}\frac{\partial}{\partial a}L(a^*,b^*,\lambda^*) &= 2a^*S_{xx} + 2b^*S_{xy} - 2\lambda^* a^* = 0 \\ \frac{\partial}{\partial b}L(a^*,b^*,\lambda^*) &= 2a^*S_{xy} + 2b^*S_{yy} - 2\lambda^* b^* = 0 \\ \frac{\partial}{\partial \lambda}L(a^*,b^*,\lambda^*) &= 1-(a^*)^2-(b^*)^2 = 0\end{aligned}$$

すなわち

$$a^*S_{xx} + b^*S_{xy} = \lambda^* a^*$$
$$a^*S_{xy} + b^*S_{yy} = \lambda^* b^*$$
$$(a^*)^2 + (b^*)^2 = 1$$

行列とベクトルで表す.

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix}, \quad \boldsymbol{a} = \begin{pmatrix} a^* \\ b^* \end{pmatrix}$$

とする．この S はデータの分散共分散行列でもある．すると，解くべき連立方程式は

$$S\boldsymbol{a} = \lambda^* \boldsymbol{a}, \qquad \boldsymbol{a}^\top \boldsymbol{a} = 1$$

つまり，求めたい方向ベクトル $\boldsymbol{a} = (a^*, b^*)^\top$ は分散共分散行列 S のある**固有値** λ^* の**固有ベクトル**であることがわかる.

しかし，S は 2 次正方行列なので，固有値は 2 つある可能性がある．では，λ^* はどの固有値か.

連立方程式の 1 つ目の式に左から $\boldsymbol{a}^\top$ をかける.

$$\boldsymbol{a}^\top S\boldsymbol{a} = \lambda^* \boldsymbol{a}^\top \boldsymbol{a} = \lambda^*$$
$$(a^*)^2 S_{xx} + 2a^*b^*S_{xy} + (b^*)^2 S_{yy} = \lambda^*$$

左辺はいま最大化したもの (z_i たちのバラツキ)．なので，右辺の λ^* も固有値のうちで最大のものだとわかる．そうでないとすると，より大きな固有値 λ が存在し，それに対応してより大きな $\lambda = \boldsymbol{a}^\top S\boldsymbol{a}$ を与える固有ベクトル $\boldsymbol{a}$ が存在することになり，$\boldsymbol{a}^\top S\boldsymbol{a}$ が最大であることに矛盾する.

したがって，このデータの特徴をもっともよく表す座標軸の方向 $\boldsymbol{a} = (a^*, b^*)^\top$ はデータの分散共分散行列 S の固有値で最大のもの $\lambda_1 (= \text{上の} \lambda^*)$ に対応する長さ 1 の固有ベクトルであることがわかった.

この座標軸をデータの**第 *1* 主成分**と呼ぶ.

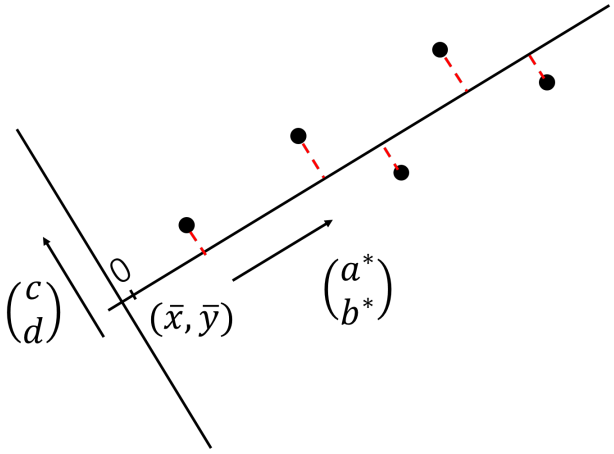


図 11

では，データの特徴を 2 番目によく表す座標軸は何か．(2 次元なので $\boldsymbol{a}$ に直交する方向なのは明らかだが，次元が増えると明らかではない．)

他の主成分は，第 1 主成分とは異なる座標軸なので，第 1 主成分の方向 $\boldsymbol{a}$ に直交するベクトルのうち，データを射影したときに一番バラツキが大きくなる方向を選べばよい．

最適化問題は

$$\begin{aligned} \max_{c,d} \quad & c^2 S_{xx} + 2cdS_{xy} + d^2 S_{yy} \\ \text{subject to} \quad & c^2 + d^2 = 1 \\ & a^* c + b^* d = 0 \end{aligned}$$

一番最後の制約式が増えたが，それは第 1 主成分の方向 $(a^*, b^*)^\top$ と直交する $(c, d)^\top$ から選択しないといけないという制約を表している．

この問題の最適解 $(c^*, d^*)^\top$ は第 1 主成分のときと同じような計算で分散共分散行列 S の 2 番目に大きな固有値 λ_2 の固有ベクトルだとわかる．それが**第２主成分**．

ここで重要なのは，固有値 λ_1, λ_2 は第 1，第 2 主成分にデータを射影したもののバラツキ具合を表しているということ．