* 主成分分析 (principal component analysis : PCA)

・データセット $\mathcal{D} = \{x_1, \cdots, x_N\}$: given.

$$x_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,D} \end{pmatrix} \in \mathbb{R}^D \quad (i = 1, \cdots, N) \quad \text{とする.}$$

(サンプル数 Nコ, 特徴量 Dコ.)

・扱うサンプルは全て標準化されているとする.

$\boxed{\text{やりたいこと}}$ サンプルデータの情報が大きく失われない程度に

サンプルの特徴量をいい感じにまとめて,

特徴量の数をDコからKコに減らしたい.

(情報の圧縮, データの要約)

→ データの可視化や特徴量の選択に使える!

たとえば
5教科のテストの点数を
総合点一つにまとめるのも
一種のデータ要約.

Q. どんな感じでサンプルの特徴量をまとめる?

A. とりあえず特徴量の線形変換でいいのでは?

→ サンプル $x_i \in \mathcal{D}$ に対して, k番目の新特徴量 $(k = 1, \cdots, K)$ を

$$t_{i,k} := \sum_{j=1}^{D} w_{k,j} x_{i,j} \quad (w_{k,j} \in \mathbb{R},\ j = 1, \cdots, D)$$

と定めることにしよう!

$$\boldsymbol{w}_k := \begin{pmatrix} w_{k,1} \\ \vdots \\ w_{k,D} \end{pmatrix} \in \mathbb{R}^D \quad (k=1,\cdots,K)$$ とすると，先の式は

$t_{i,k} = \boldsymbol{x}_i^T \boldsymbol{w}_k$ と書ける．これを全サンプルについてまとめて書くと，

$$\tilde{\boldsymbol{t}}_k := \begin{pmatrix} t_{1,k} \\ \vdots \\ t_{N,k} \end{pmatrix} \in \mathbb{R}^N \quad (k=1,\cdots,K)$$ として，

$$\tilde{\boldsymbol{t}}_k = \begin{pmatrix} \boldsymbol{x}_1^T \boldsymbol{w}_k \\ \vdots \\ \boldsymbol{x}_N^T \boldsymbol{w}_k \end{pmatrix} = X\boldsymbol{w}_k$$ と書ける．

さらにこれを全ての新特徴量についてまとめて書くと，

$$T := \left.\Big(\overbrace{\tilde{\boldsymbol{t}}_1 \ \cdots \ \tilde{\boldsymbol{t}}_K}^{K} \Big)\right\} N \in M_{N,K}(\mathbb{R}),$$

$$W := \left.\Big(\overbrace{\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_K}^{K} \Big)\right\} D \in M_{D,K}(\mathbb{R})$$

として，

$$T = (X\boldsymbol{w}_1 \ \cdots \ X\boldsymbol{w}_K) = X(\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_K) = XW$$

と書ける．Tを

$$T = \begin{pmatrix} \tilde{\boldsymbol{t}}_1 & \cdots & \tilde{\boldsymbol{t}}_K \end{pmatrix} =: \begin{pmatrix} \boldsymbol{t}_1^T \\ \vdots \\ \boldsymbol{t}_N^T \end{pmatrix}, \quad \boldsymbol{t}_i \in \mathbb{R}^K$$

$$W^T = \begin{pmatrix} \boldsymbol{w}_1^T \\ \vdots \\ \boldsymbol{w}_K^T \end{pmatrix}$$

とも書くことにすると，$T = XW$ より $\boldsymbol{t}_i^T = \boldsymbol{x}_i^T W$ なので $\boldsymbol{t}_i = \underbrace{W^T \boldsymbol{x}_i}_{\boldsymbol{x}_i \text{の} W^T \text{による線形変換}}$．

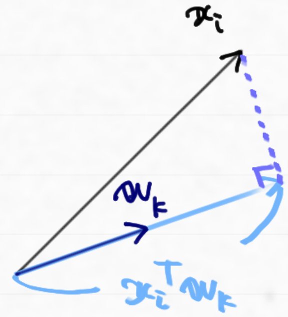
・ということで，$\boxed{\text{やりたいこと}}$がハッキリ具体的になった：

いい感じの線形変換W^Tをみつけて，データセットの情報量を

できるだけ残しつつ次元をおとしたデータセットを得たい．

○ ところで，線形変換W^Tは図形的にはどんなことをしているのか？

→ $t_{i,k} = w_k^T x_i$ について



$\|w_k\| = 1$ であれば $w_k^T x_i$ は

x_i の w_k 方向への正射影の長さ．

→ つまり，線形変換W^Tは，サンプルxを$w_1, \cdots, w_K$のそれぞれ

の方向に正射影する射影変換．

・そもそもw_kは$t_{i,k}$を作るために

$$t_{i,k} = \sum_{j=1}^{D} w_{k,j} x_{i,j} \quad (w_{k,j} \in \mathbb{R},\ j=1,\cdots,D)$$

と導入された量で，重要なのは

各jに対し$x_{i,j}$をどの程度足しあわせることで新しい特徴量$t_{i,k}$を
つくるか

ということ．つまり，w_kの各成分$w_{k,j}$の比率のみが重要！

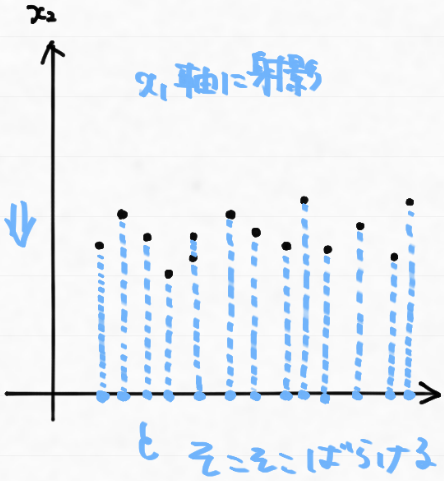
ベクトルの言葉でいえば，w_kの向きが大事で長さはどうでもよいということ．

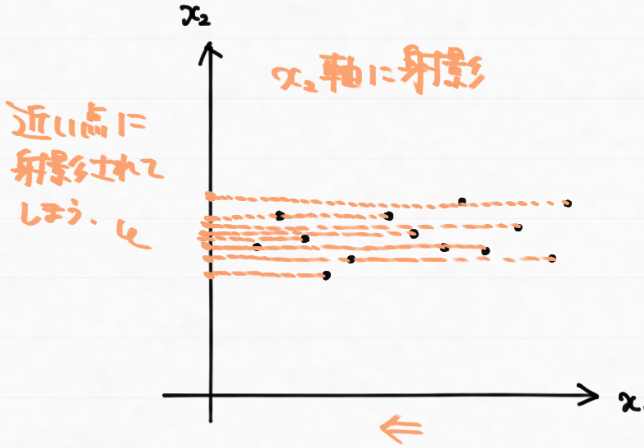
→ 以下では，$\|w_k\| = 1 \ (k=1,\cdots,K)$としておけばよい！

大きさ1のベクトルを正規ベクトルという．

Q. どの方向 w を選べば，データセットの情報量をできるだけ残せる？

A. 単純な例で考えてみよう．



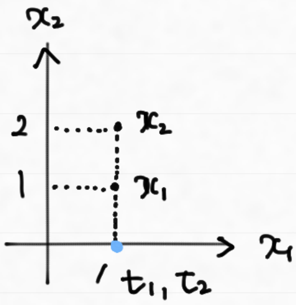


異なるサンプルが同じ値に射影されてしまうと，片方のサンプルのもつ情報は完全に失われてしまってかなしい．

たとえば，2点

$$x_1 = \begin{pmatrix}1\\1\end{pmatrix}, \quad x_2 = \begin{pmatrix}1\\2\end{pmatrix}$$

を x_1軸に射影すると
どちらも $t_1 = t_2 = 1$ となってしまい，
もともと片方のサンプルしかなかったときと同じ結果になる．



→ 異なるサンプルが同じ値に射影されにくいような方向 w をみつければデータセットの情報量をできるだけ保持できそう．

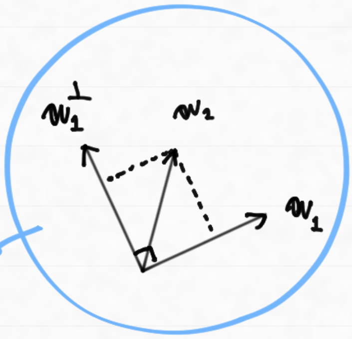
→ 射影先の値が大きくばらついていればいるほど，同じ値には射影されにくそう！

・値のばらつきを表す量 → 分散！

- 実際には、$w_1, \cdots, w_K$ の K コの方向を選ばなければならない。
  - w_1 は $\{t_{i,1}\}_{i=1}^N$ の分散が最大になるように選べばよい。
  - w_2 はどう選ぶ？
    - $w_2 \in \mathbb{R}^D$ を適当にきめたとする。

      → w_1 と直交する適当な $w_1^\perp \in \mathbb{R}^D$ を用いて

      $w_2 = a w_1 + b w_1^\perp$ と表すことができる。



データ x_i に対して、$t_{i,2} = a x_i^T w_1 + b x_i^T w_1^\perp = a t_{i,1} + b x_i^T w_1^\perp$.

→ $t_{i,2}$ に $t_{i,1}$ の項が入ってくると、1番目の新特徴量 $t_{i,1}$ との影響が2番目の新特徴量 $t_{i,2}$ にもでてきてしまう。

$t_{i,1}$ と $t_{i,2}$ の相関が0でないということ。

$t_{i,1}$ の情報は既にあるのだから、冗長ではないか？

→ $a=0$ となるように w_2 の方向を決める。つまり、w_2 は w_1 と直交する方向からさがすようにすればよさそう！

- しかも $\{t_{i,2}\}_{i=1}^N$ の分散ができるだけ大きくなるように選ぶ。

- 同様に、w_k $(k=3, \cdots, K)$ も、$\{w_1, \cdots, w_{k-1}\}$ と直交する方向から、$\{t_{i,k}\}_{i=1}^N$ の分散ができるだけ大きくなるように選べばよい。

・ 以上より，$\boxed{\text{やりたいこと}}$ がはっきりしたので，まとめる：

$\boxed{\text{やりたいこと}}$

"大きさが1で互いに直交する" ということ

正規直交するベクトル $\{w_1, \cdots, w_K\} \subseteq \mathbb{R}^D$ で，各 k に対して

$\{t_{i,k} = x_i^T w_k\}_{i=1}^N$ の分散が最大になるものを見つけたい．

○ やっていくことにする．

$\|w\| = 1$ なる $w \in \mathbb{R}^D$ を考える．$\{t_i = x_i^T w\}_{i=1}^N$ の標本分散

を求めたい．まず $\{t_i\}_{i=1}^N$ の標本平均 $\bar{t}$ は，

$$\bar{t} = \frac{1}{n}\sum_{i=1}^N t_i = \frac{1}{n}\sum_{i=1}^N x_i^T w = \left(\frac{1}{n}\sum_{i=1}^N x_i^T\right) w = \left(\frac{1}{n}\sum_{i=1}^N x_i\right)^T w$$

$$= 0^T w \quad (\because \text{標準化されているから } \bar{x} = 0)$$

$$= 0.$$

なので，標本分散 σ_w^2 は，

$$\sigma_w^2 = \frac{1}{n}\sum_{i=1}^N (t_i - \bar{t})^2 = \frac{1}{n}\sum_{i=1}^N t_i^2 = \frac{1}{n}\sum_{i=1}^N w^T x_i x_i^T w$$

$$= w^T\left(\frac{1}{n}\sum_{i=1}^N x_i x_i^T\right) w$$

$$= w^T\left(\frac{1}{n}\sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T\right) w \quad (\because \bar{x} = 0)$$

$$= w^T \Sigma w. \quad (\Sigma : \text{標本共分散行列}).$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T.$$

Σ は対称行列．

→ 函数 $f(w) = w^T \Sigma w$ を，制約条件 $\|w\| = 1$ のもとで

最大化する w を求める。

これを

$$
(P) \quad \begin{aligned} &\max \ f(w) = w^T \Sigma w \\ &\text{s.t.} \ \|w\| = 1 \end{aligned}
$$

← 等式制約つき最適化問題

"subject to"

と書く。

- $\|w\| > 0$ なので，$\|w\| = 1 \Leftrightarrow \|w\|^2 = 1$ であるから，

$$
(P') \quad \begin{aligned} &\max \ f(w) = w^T \Sigma w \\ &\text{s.t.} \ \|w\|^2 = 1 \end{aligned}
$$

とすると，問題(P')は(P)と等価。

Lagrangeの未定乗数法により

制約なし最適化問題

$$
(L) \quad \max \ L(w, \lambda) := f(w) - \lambda(\|w\|^2 - 1)
$$

とすると，(L)は(P')と等価．結局(L)をとけばよい．

- $L(w,\lambda)$の最適解の候補は，

$\dfrac{\partial L}{\partial w} = 0$，$\dfrac{\partial L}{\partial \lambda} = 0$ をみたす w, λ である．

（最適性の必要条件）

$\frac{\partial L}{\partial \lambda} = 0$ より $\|w\|^2 = 1 \iff \|w\| = 1$. → 大きさは1にする.

$\frac{\partial L}{\partial w} = 2\Sigma w - 2\lambda w$ （対称行列Aについて $\frac{\partial}{\partial w} w^T A w = 2Aw$ を用いた.）

となるので.

$$\Sigma w = \lambda w \quad -(*)$$

$\|w\| = 1$ より $w \neq 0$ なので. (*)は w が Σ の固有ベクトル

であることを意味する！

・ しかし固有ベクトルは $\dim \Sigma = D$ コある. どれを選ぶ？

→ (L)はもとの問題(P)と等価だったから. (P)の最適解も

(*)をみたす. よって, (L)の最適解を (w^*, λ^*) とすると, (P)の最適値は,

$$f(w^*) = w^{*T} \Sigma w^* = w^{*T}(\lambda^* w^*) \quad (\because (*) より)$$

$$= \lambda^* \|w^*\|^2$$

$$= \lambda^*. \quad (\because \|w^*\| = 1)$$

となる. つまり, (P)の最適値は Σ の最大固有値 λ^* であり

最適解は λ^* に対応する固有ベクトル w^* である！

- ここでもう一度 $\boxed{\text{やりたいこと}}$ を確認してみる：

$\boxed{\text{やりたいこと}}$

正規直交するベクトル $\{w_1, \cdots, w_K\} \subseteq \mathbb{R}^D$ で，各 k に対して

$\{t_{i,k} = x_i^T w_k\}_{i=1}^N$ の分散が最大になるものを見つけたい．

K コの正規直交するベクトル $\{w_1, \cdots, w_K\} \subseteq \mathbb{R}^D$ をみつける必要があるのだった．

- 線形代数の次の定理を思い出そう：

Thm. n 次実対称行列 $A \in M_n(\mathbb{R})$ の n 個の固有値は非負である．

さらに，n 個の固有ベクトルは正規直交するようにとれる． □

- Σ は D 次実対称行列なので，D 個の非負固有値がとれる．これを

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \ (\geq 0)$$

と降順に並べ，対応する固有ベクトルを $w_1, w_2, \cdots, w_D$ としよう．

→ 先の議論により，w_1 は変換後の座標の分散が最大 (λ_1) になる方向．次に分散が大きくなるのは w_2 の方向．以下同様．

→ 固有値を降順に並べ，対応する固有値を前から K コとった $\{w_1, \cdots, w_K\}$ が求めていたベクトル！

・まとめると，PCAは以下のようにできる：

Alg. (PCA)

Input：標準化済みのデータからなるデータ行列 $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix} \in M_{N,D}(\mathbb{R})$

正整数 $K \in \mathbb{Z}_{>0}$

Output：次元削減したデータからなる行列 $T = \begin{pmatrix} t_1^T \\ \vdots \\ t_N^T \end{pmatrix} \in M_{N,K}(\mathbb{R})$

1. 標本平均の計算：$\bar{x} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i$;
2. 標本共分散行列の計算：$\Sigma \leftarrow \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$;
3. 固有値，固有ベクトルの計算：

   Σの固有値，固有ベクトルを $\lambda_1 \geq \cdots \geq \lambda_D$ 対応 $\updownarrow$ $\updownarrow$ $w_1, \cdots, w_D$ とする；
4. 変換行列の計算：$W \leftarrow (w_1 \cdots w_K)$;
5. 変換：$T \leftarrow XW$;
6. return T. $\square$

こうしてx_iを変換して得られるt_iの第k成分を，第k主成分という $(k = 1, \cdots, K)$.

Q. D次元からK次元に次元削減したが、どの程度の情報が表現できているのだろうか？

A. 変換後の座標の分散が全分散に占める割合がその主成分軸で表現できている情報の割合を反映していると考えられる。

全分散は固有値の和で

$$V_{total} := \sum_{k=1}^{D} \lambda_k \quad (= \mathrm{tr}\, \Sigma)$$

となる。第k主成分が表現している情報の割合は

$$c_k := \frac{\lambda_k}{V_{total}} \qquad (k = 1, \cdots, D)$$

であり、これを第k成分の**寄与率**という。

第K主成分まで用いることで表現できる情報の割合は

$$r_K := \frac{1}{V_{total}} \sum_{k=1}^{K} \lambda_k$$

となり、第K成分までの**累積寄与率**という。

結局、K次元に次元削減することで、もとの情報の r_K 倍の情報が表現されていることになる。