## work-010

PCAの簡単な導出を行う。データ集合  $\{ {m x}_n \}$   $n=1,\ldots,N$  を考える。 ${m x}$  は D次元のデータ(特徴量がD個)であり、D < N を仮定する。

$$\mathbf{x}_i = x_{i1}\mathbf{e}_1 + x_{i2}\mathbf{e}_2 + \cdots + x_{iD}\mathbf{e}_D$$

ここで、 $\{e_i\}$  は正規直交基底とする。 以下で第一主成分を求める。

まず、D次元単位ベクトル  $m u_1$  を考える。第一主成分とは、「データの分散が最大となる方向」であるので、 $m u_1^Tm x$  の分散が最大となるような $m u_1$ を求めればよい。

1.  $oldsymbol{u}_1^Toldsymbol{x}$  はどのような量か。説明せよ。

分散は

$$ext{Var}(oldsymbol{u}_1^Toldsymbol{x}) = rac{1}{N}\sum_i^N \{oldsymbol{u}_1^Toldsymbol{x}_i - oldsymbol{u}_1^Toldsymbol{\overline{x}}\}^2 = oldsymbol{u}_1^Toldsymbol{\Sigma}oldsymbol{u}_1$$

となる。

ここで、 $\overline{x}$ 、 $\Sigma$  はそれぞれ x の平均、共分散行列を表す。

- 2.  $\overline{x}$ 、 $\Sigma$  を書け。
- 3.  $\operatorname{Var}(\boldsymbol{u}_1^T\boldsymbol{x}) = \boldsymbol{u}_1^T \boldsymbol{\Sigma} \boldsymbol{u}_1$  の計算過程を書け。

 $\operatorname{Var}(\boldsymbol{u}_1^T\boldsymbol{x})$  を最大にする  $\boldsymbol{u}_1$ を求める。ラグランジュの未定乗数法を用いると $\boldsymbol{u}_1$ の満たすべき条件式が以下になることがわかる。

$$\Sigma u_1 = \lambda_1 u_1$$

- 4. ラグランジュの未定乗数法を用いて上式を導け。
- 5. 求めた $u_1$ の各成分はどのような意味を持つか?
- 6. **u**<sub>1</sub> 方向の分散を求めよ。

以上が、第一主成分を求める手続きである。上記のような方法を繰り返すことで、第二、第三主成分も同様に求めることができる。

求めた主成分を用いて各 $oldsymbol{u}_i$ 方向にデータ $oldsymbol{x}_n$ を射影する。具体的には以下のように表現される。

$$egin{aligned} y_{n1} &= oldsymbol{u}_1^T oldsymbol{x}_n \ y_{n2} &= oldsymbol{u}_2^T oldsymbol{x}_n \ dots \ y_{ni} &= oldsymbol{u}_i^T oldsymbol{x}_n \ dots \ y_{nd} &= oldsymbol{u}_d^T oldsymbol{x}_n \end{aligned}$$

行列とベクトルで表現すれば、

$$oldsymbol{y}_n = U^T oldsymbol{x}_n$$

となる。

7. 上式はどのような意味を持つか。説明せよ。