

3章 固有値問題を用いたカーネル多変量解析

中川 哲

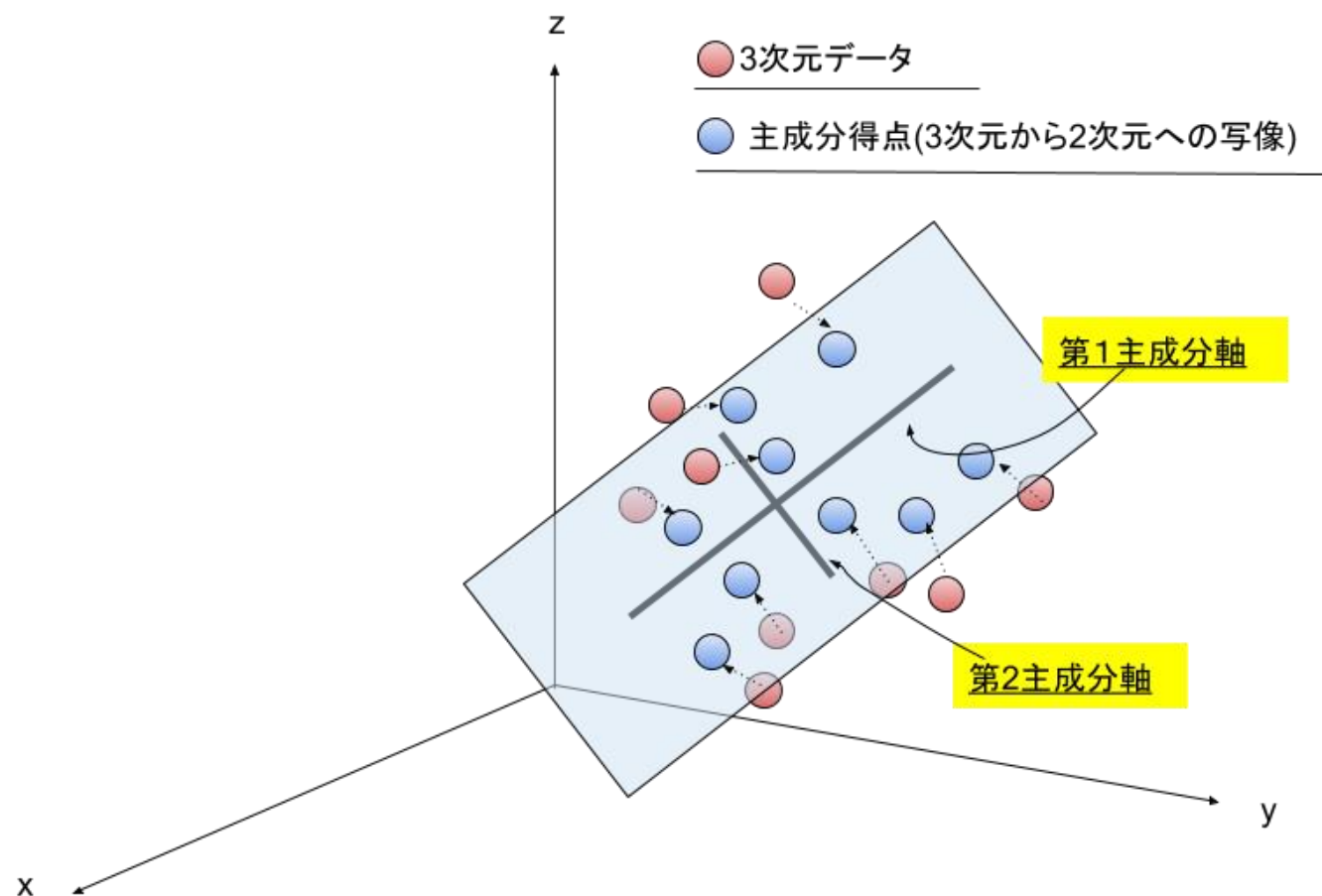
目次

- 3 主成分分析の復習
 - 3.1 カーネル主成分分析
 - 3.2 次元圧縮とデータ依存カーネル

3.主成分分析の復習

主成分分析とは？

高次元のデータを低次元のデータに圧縮すること



低次元構造を抽出するために、以下の二つの等価な最適化を行う

[1] 低次元に射影したときに、ばらつきができるだけ大きくなるようにする

[2] 縮約したデータをもとの近似データとみなしたとき、その近似誤差をできるだけ小さくする

仮にデータが正規分布に従うとすると、情報量は

$$p(x; \mu, \sigma^2) = \exp\left(-\frac{(x - \mu)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right)$$

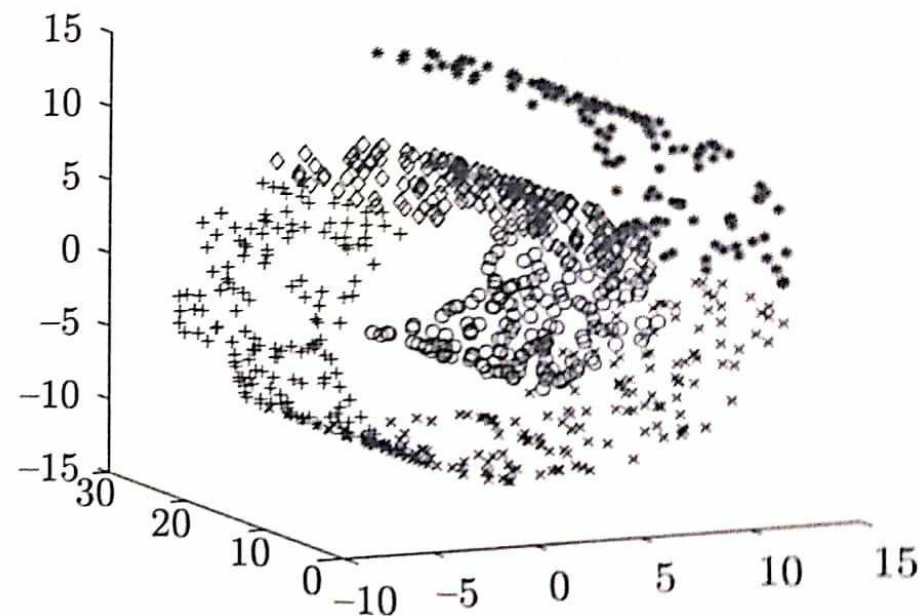
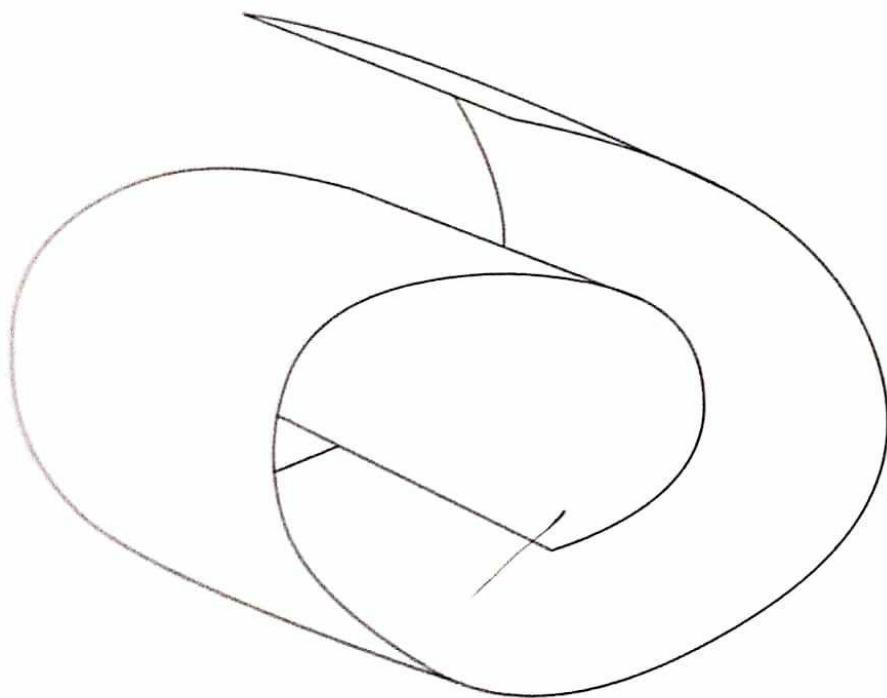
これを情報理論で用いられるエントロピーとして計算すると

$$-\int_{-\infty}^{\infty} p(x) \ln p(x) dx = \frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2) = \frac{1}{2}\ln \sigma^2 + \text{定数}$$

分散が大きい＝情報量が多い

3.1カーネル主成分分析

通常の主成分分析は非線形なデータ構造 に対して適用できない



スイスロールと呼ばれている3次元中の2次元構造

どうすれば良いのか？



高次元の特徴ベクトルに変換する

例 $\boldsymbol{x} = (x_1, x_2)^T$ を用いて作られる

$$a_1 x_1 + a_2 x_2^2 + a_3 x_2 + a_4 = 0$$

上の式は非線形な曲線であるが

$$\phi_1(\boldsymbol{x}) = x_1, \quad \phi_2(\boldsymbol{x}) = x_2^2, \quad \phi_3(\boldsymbol{x}) = x_2$$

という変換を行うと

$$a_1 \phi_1(\boldsymbol{x}) + a_2 \phi_2(\boldsymbol{x}) + a_3 \phi_3(\boldsymbol{x}) + a_4 = 0$$

$$a_1\phi_1(\boldsymbol{x}) + a_2\phi_2(\boldsymbol{x}) + a_3\phi_3(\boldsymbol{x}) + a_4 = 0$$



通常の主成分分析を使えばいい！

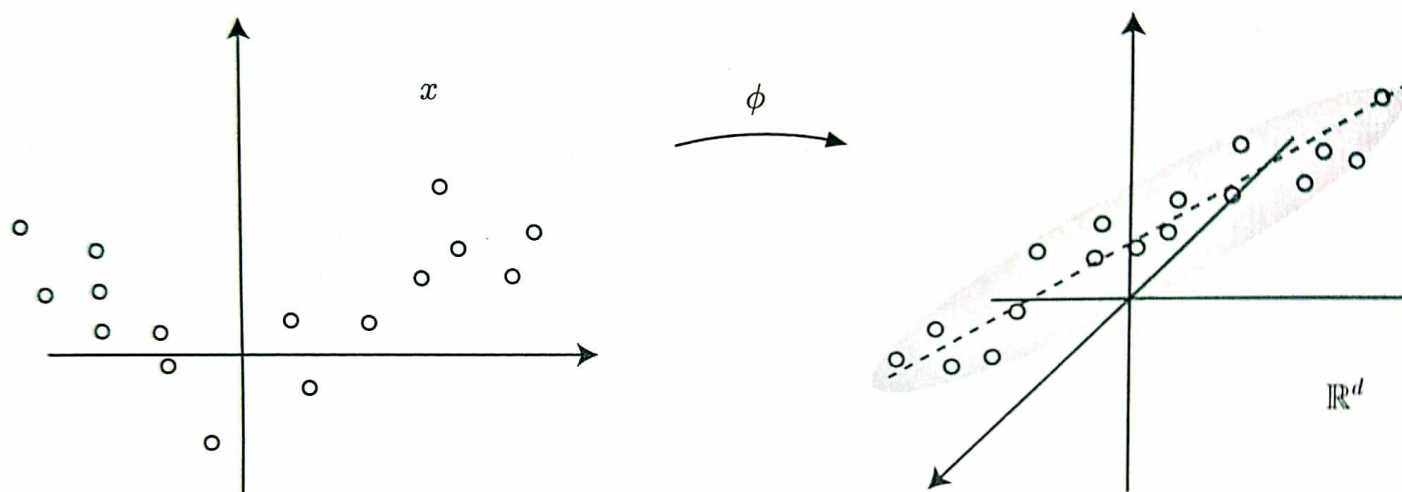


図 3.2 カーネル主成分分析の概念図．特徴抽出をした空間で分散が最大となる線形の部分空間を求める．

(1)平均0の場合 ($E_n[\phi(\boldsymbol{x})] = \frac{1}{n} \sum_{i=1}^n \phi(\boldsymbol{x}^{(i)}) = 0$)

特徴ベクトル $\phi(\boldsymbol{x})$ を1次元の直線上に射影し

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x})$$

という関数を考える、ただし \boldsymbol{w} は

$$\|\boldsymbol{w}\|^2 = 1$$

を満たす(単位ベクトル)

単位ベクトルに射影した点のサンプル分散は

$$\text{Var}_n[f(x)] = \frac{1}{n} \sum_{i=1}^n (w^T \phi(x^{(i)}))^2$$

これを単位ベクトルの制約式のもとで最大化
するという問題を解く



ラグランジュの未定乗数法を使う

分散の式に制約条件を加えた

$$L(w) = -\text{Var}_n[f(x)] + \lambda(\|w\|^2 - 1)$$

$L(w)$ を w で微分して0とおくと

$$-\frac{2}{n} \sum_{i=1}^n \left(w^T \phi(x^{(i)}) \right) \phi(x^{(i)}) + 2\lambda w = 0$$

$\alpha_i = w^T \phi(x^{(i)}) / (n\lambda)$ とおくことにより

$$w = \sum_{i=1}^n \alpha_i \phi(x^{(i)})$$

という形に書ける

よって $f(\mathbf{x})$ は

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}^{(i)}, \mathbf{x})$$

という形に書ける

ここで導入した α を使うとサンプル分散は

$$\begin{aligned} \text{Var}_n[f(\mathbf{x})] &= \frac{1}{n} \sum_{l=1}^n \left(\sum_{i=1}^n \alpha_i k(\mathbf{x}^{(i)}, \mathbf{x}^{(l)}) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \alpha_i \alpha_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(l)}) k(\mathbf{x}^{(j)}, \mathbf{x}^{(l)}) \\ &= \frac{1}{n} \boldsymbol{\alpha}^T \mathbf{K}^2 \boldsymbol{\alpha} \end{aligned}$$

$L(w)$ を α で書き直すと

$$L(\alpha) = -\frac{1}{n}\alpha^T K^2 \alpha + \lambda(\alpha^T K \alpha - 1)$$

α について微分し0とおくと

$$-\frac{2}{n}K^2\alpha + 2\lambda K\alpha = 0$$

K が正則と仮定すれば

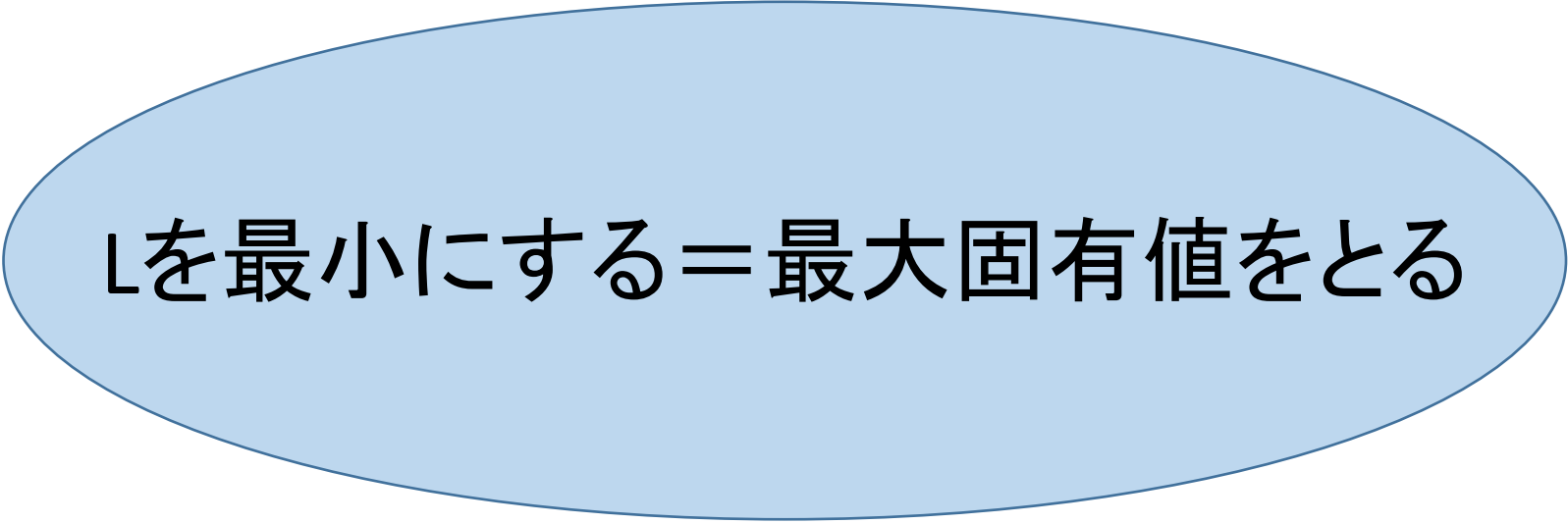
$$K\alpha = \lambda\alpha$$

これはグラム行列に対する固有値問題となる

上の式を使うと

$$L(\alpha) = -\lambda$$

となるので



Lを最小にする＝最大固有値をとる

2次元以上の空間に射影をとるときは、上からM個の固有値を取ってきて対応する固有ベクトルによって張られる空間に射影すれば良い

$$\lambda_1, \dots, \lambda_n$$

固有値

$$\alpha_1, \dots, \alpha_n$$

対応する固有ベクトル

$$f_j(\mathbf{x}) = \sum_{i=1}^n \alpha_{ji} k(\mathbf{x}^{(i)}, \mathbf{x}), \quad j = 1, \dots, M$$

$$\boldsymbol{\alpha}_j = (\alpha_{j1}, \dots, \alpha_{jn})^T$$

(2)一般の場合

サンプル平均が0と仮定しない場合は「分散＝二乗平均－平均の二乗」
を使ってサンプル分散を求める

サンプル平均とサンプル二乗平均はそれぞれ

$$\begin{aligned} \mathbb{E}_n[f(\boldsymbol{x})] &= \frac{1}{n} \sum_{l=1}^n f(\boldsymbol{x}^{(l)}) & \mathbb{E}_n[f(\boldsymbol{x})]^2 &= \frac{1}{n^2} (\boldsymbol{\alpha}^T K \mathbf{1})^2 \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{l=1}^n \alpha_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(l)}) & &= \frac{1}{n^2} (\boldsymbol{\alpha}^T K \mathbf{1}) (\mathbf{1}^T K \boldsymbol{\alpha}) \\ &= \frac{1}{n} \boldsymbol{\alpha}^T K \mathbf{1} & & \end{aligned}$$

よってサンプル分散は

$$\text{Var}_n[f(\boldsymbol{x})] = \frac{1}{n} \boldsymbol{\alpha}^T K J_n K \boldsymbol{\alpha} \quad J_n = I_n - \frac{1}{n} \mathbf{1} \mathbf{1}^T$$

となるので、最終的に

$$J_n K \boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}$$

という固有値問題を解くことに帰着される

カーネル主成分分析のまとめ

[1] データ点の集合 $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ からグラム行列 K を作る

[2] $J_n K \alpha = \lambda \alpha$ という固有値問題を解く

[3] 固有値に対応する固有ベクトルを取ってきて射影する

カーネル主成分分析は特徴ベクトルの世界で分散や二乗誤差を計算しているので特徴ベクトルの選び方(=カーネル関数の選び方)によって結果が変化してしまう

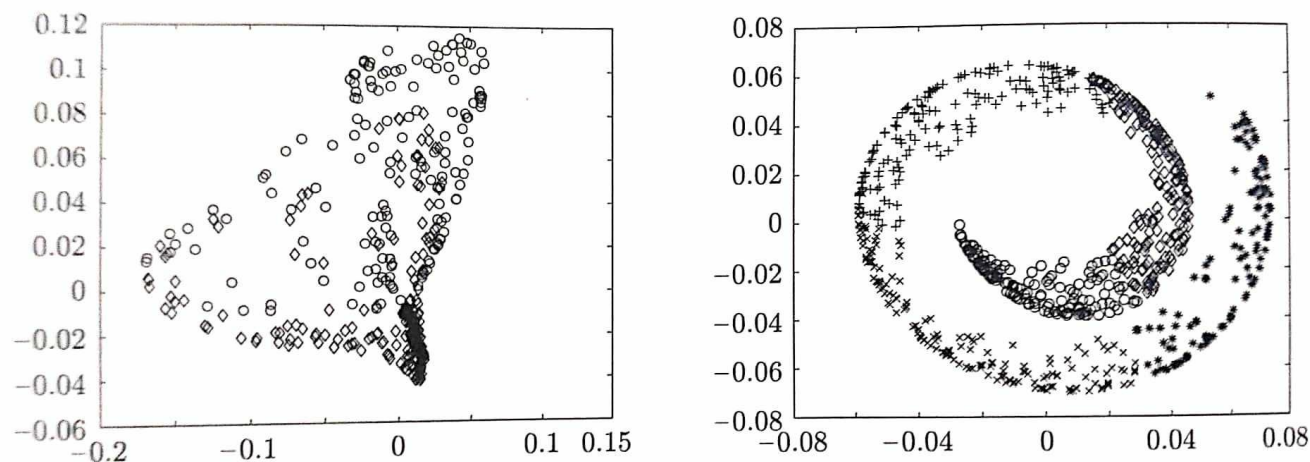
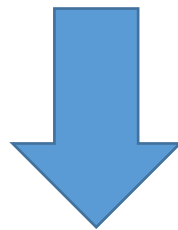


図 3.3 カーネル主成分分析をスイスロールのデータ(図 3.1)に適用して 2 次元空間に落とした例. カーネルはガウスカーネルを使った. β の値が左 0.1, 右 0.001 で, 値によって結果が大きく異なることがわかる. スイスロール状に带状に配置した記号(* \rightarrow * \rightarrow * \rightarrow * \rightarrow * \rightarrow *)は右側の図で一応その順番通りに並んでいる.

今まではカーネル関数がデータと無関係に決まっていた



今度は関数をデータに応じて変えてしまおう！

正定値行列とカーネル関数の等価性

サンプルと同じサイズの任意の行列 K が与えられたとき、
特徴ベクトルが存在し、それに対応するカーネル関数の
定めるグラム行列が K になる



データに応じて自分で正定値行列を設計する
ことが可能！

(詳しい説明は6章で)

類似度と正定値行列

2章でも述べた通りカーネル関数は特徴ベクトルの内積として定義されているので、特徴ベクトル同士の類似度・近さを表していると考えることができる

科学実験や社会調査などによって類似度を測定するようなアプリケーションでは、サンプル同士の類似度が直接与えられることもあるので、これを用いてカーネル関数を使うことができる

(詳しい説明は5章で)

3.2 次元圧縮とデータ依存カーネル

グラム行列とグラフ構造

サンプルデータをグラフの頂点に、カーネル関数は2つのデータ間から決まるので頂点と頂点を結ぶ枝と対応づける

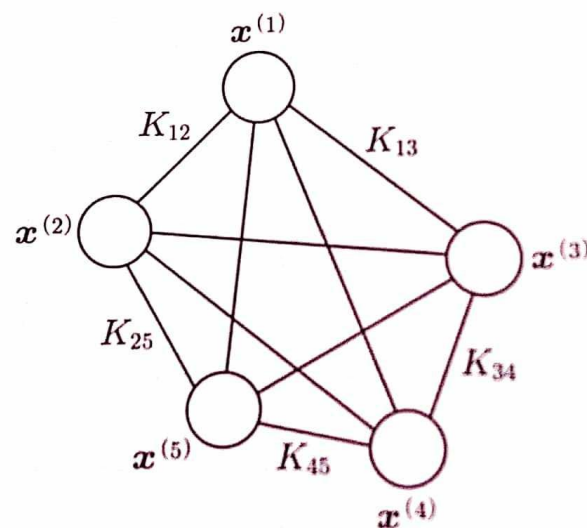


図 3.4 グラム行列とグラフ構造. 各サンプルをグラフのノード, サンプル間の関連性を枝の重みとして表現できる.

(a)ラプラシアン固有マップ法

サンプルに対応するグラフの枝に重みを付けることを考える。
互いに近いデータは大きな重みを、遠いデータは小さな重み
をとるようになる。

(例、データ空間が実数ベクトル空間ならガウスカーネルを重
みとしてとれる)

i と j を結ぶ枝の重み K_{ij} を成分とする行列を K として、サンプル
を1次元の値に縮約して表現することを考える

i番目のサンプル表現 β_i を決めるためにデータ間の重み付きの差を小さくすることを考える、つまり

$$\min_{\beta} \sum_{i,j} K_{ij} (\beta_i - \beta_j)^2$$

という問題を解く。2次形式で書いたものを

$$\sum_{i,j} (\beta_i - \beta_j)^2 K_{ij} = 2\beta^T P \beta$$

とおく。 β_i の自由度を除くために以下の制約を行う

$$\beta^T \Lambda \beta = 1$$

ラグランジュ関数は

$$L(\beta) = \beta^T P \beta - \lambda(\beta^T \Lambda \beta - 1)$$

となり、最適化問題になるので、 β で微分して0とおくと

$$P\beta = \lambda\Lambda\beta$$

という、一般固有値問題の最小固有値に対応する固有ベクトルを求めることに帰着される

(ただし固有値0で $\beta \propto 1$ という自明な解を除く)

小さい固有値に対応する固有ベクトルから順番に β_1, β_2, \dots と取っていき、各ベクトルの l 番目の成分 $\beta_{1l}, \beta_{2l}, \dots$ を集めたのがラプラシアン固有マップ法におけるサンプル $x^{(l)}$ の表現となる

カーネル主成分分析との関連性を見ると $P = \Lambda - K$ から

$$K\beta = \lambda' \Lambda \beta, \quad \lambda' = 1 - \lambda$$

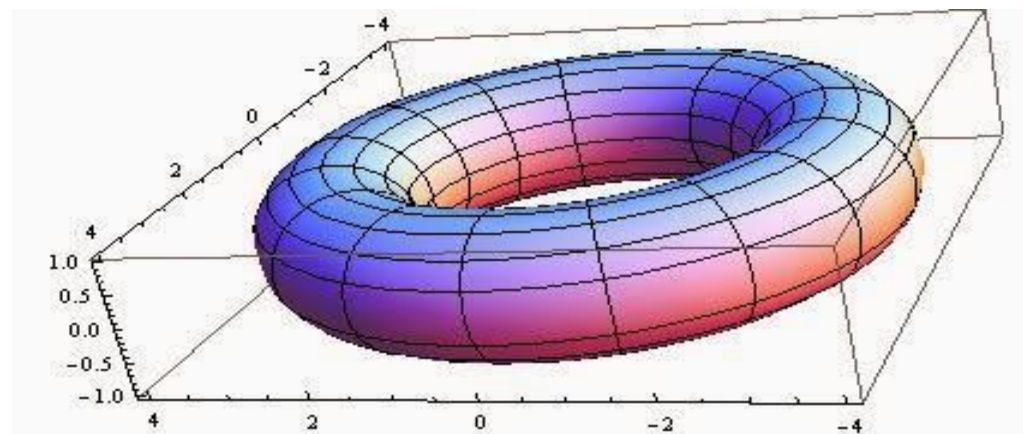
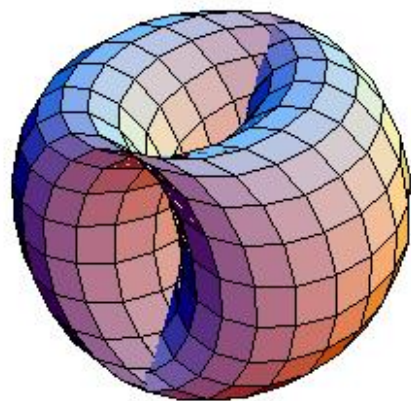
となり、 λ の最小化は、こちらの固有値問題では固有値 λ' の最大化となる

(b)ISOMAP:多様体上に基づく次元圧縮

(1) 次元圧縮と多様体あてはめ

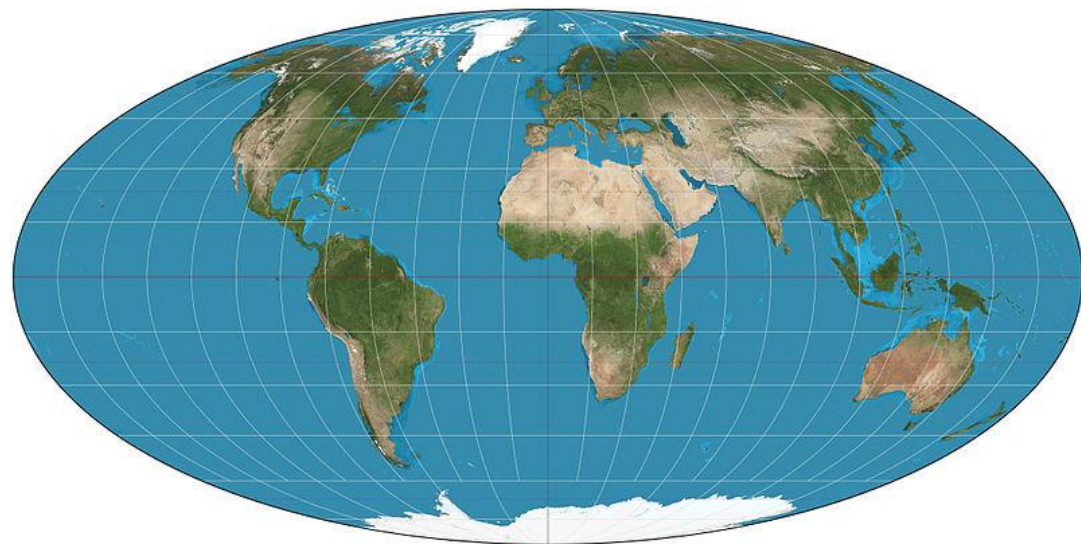
- 多様体について

多様体は非常に狭い範囲で見れば普通のユークリッド空間と同じ構造を持つが広い範囲で見ると一般には曲がった構造をしている



多様体は曲がっていてもその上に
適当な座標系を取ることができる
球面における経度・緯度がそれに
あたる。

経度は北極、南極で縮退してしまう
などユークリッド空間での直交座標系
とは異なる。これを回避するために、
複数のユークリッド空間を張り合わせて
多様体全体を覆えばよい



(2)多様体上の距離

グラム行列を適切に設定するためには「近さ」を決める必要があるが難しい



近さの反対概念の遠さを表す「距離」なら決められる！

データとして与えられているのはサンプル点だけなので
多様体をそのサンプル点を使って表現し、その上で最短
距離を近似的に求める必要がある



サンプル点を端点とする近傍グラフを作る

具体的には各サンプルどうしのユークリッド距離を測り、あらかじめ決めたしきい値 ϵ 以下、あるいは K 個の近傍について枝で結ぶ

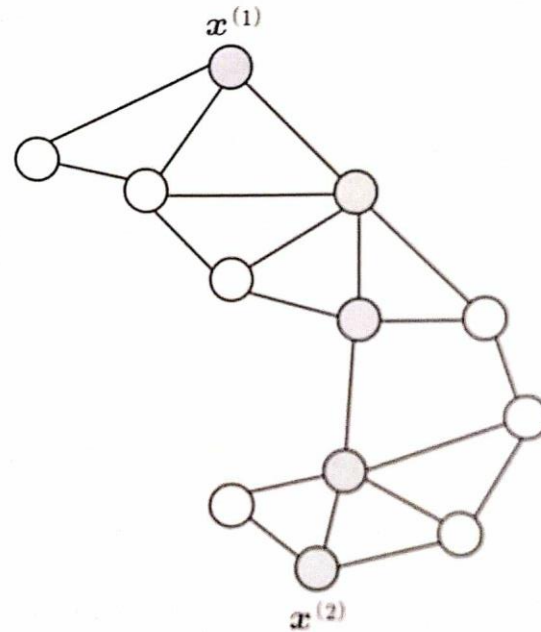


図 3.6 ISOMAP の概念図. サンプル点から近傍グラフを作り, 枝にはサンプル点の間のユークリッド距離の重みをおく. はなれた点(たとえば $x^{(1)}$ と $x^{(2)}$)への多様体上の距離はグラフの最短経路で近似する.

(3) 距離からカーネルへ

多様体上の距離は求まったがカーネル法を行うためには距離から類似度に変換しなければならない

簡単のため特徴ベクトルがあって、その間のユークリッド距離として距離が与えられているとする

特徴ベクトル $\phi(\mathbf{x})$ をユークリッド空間上の点と見なすと

$$\left\| \phi(\mathbf{x}^{(i)}) - \phi(\mathbf{x}^{(j)}) \right\|^2 = \left\| \phi(\mathbf{x}^{(i)}) \right\|^2 + \left\| \phi(\mathbf{x}^{(j)}) \right\|^2 - 2\phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)})$$

この式を変形するとカーネル関数は

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = -\frac{1}{2} \left(\left\| \phi(\mathbf{x}^{(i)}) - \phi(\mathbf{x}^{(j)}) \right\|^2 - \left\| \phi(\mathbf{x}^{(i)}) \right\|^2 - \left\| \phi(\mathbf{x}^{(j)}) \right\|^2 \right)$$

ただし、多くの問題ではサンプルどうしの距離だけしかわからない場合も多いので上の式の第2項、第3項は計算ができない

そこで、N個のデータ集合の要素の間の距離を要素として持つ行列Dが与えられる場合に、そこからグラム行列を計算する方法を考える。

Dの i, j 成分 D_{ij} が $\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ の間の距離を表すとする

$$\begin{aligned} D_{ij} &= \left\| \phi(\mathbf{x}^{(i)}) - \phi(\mathbf{x}^{(j)}) \right\|^2 \\ &= K_{ii} + K_{jj} - 2K_{ij} \end{aligned}$$

という関係式が得られる

特徴ベクトル全体を平行移動すると内積が変化してしまい
カーネルの値が定まらない



特徴ベクトルのサンプル平均を0に固定する

$$\sum_{i=1}^n \phi(\mathbf{x}^{(i)}) = \mathbf{0}$$

$\phi(\mathbf{x}^{(j)})$ との内積をとると

$$\sum_{i=1}^n \phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)}) = \sum_{i=1}^n K_{ij} = 0$$

また(3.38)を*i,j*について足し合わせると

$$\sum_{i=1}^n D_{ij} = \sum_{i=1}^n K_{ii} + nK_{jj}$$

$$\sum_{j=1}^n D_{ij} = nK_{ii} + \sum_{j=1}^n K_{jj}$$

全ての総和は

$$\sum_{i=1}^n \sum_{j=1}^n D_{ij} = 2n \sum_{i=1}^n K_{ii}$$

K_{ii} , K_{jj} を消去すると

$$2K_{ij} = \frac{1}{n} \left\{ \sum_{i=1}^n D_{ij} - \sum_{i=1}^n \frac{1}{n} \left(\sum_{i=1}^n D_{ij} - \sum_{i=1}^n k_{ij} \right) \right\} \\ + \frac{1}{n} \left(\sum_{i=1}^n D_{ij} - \sum_{i=1}^n k_{ij} \right) - D_{ij}$$

$\sum_{i=1}^n K_{ii}$ を消去すれば

$$-D_{ij} + \frac{1}{n} \sum_{i'=1}^n D_{i'j} + \frac{1}{n} \sum_{j'=1}^n D_{ij'} - \frac{1}{n^2} \sum_{i'=1}^n \sum_{j'=1}^n D_{i'j'} = 2K_{ij}$$

これは多変量解析で二重中心化と呼ばれている

(d)局所線形埋め込み法

前節にも述べた通り、多様体は狭い範囲で見れば線形空間と見なすことができる。この性質を利用し、以下のステップで多様体のあてはめを行う

[1] 狭い範囲の点だけを使い低次元の線形モデルをあてはめる

[2] そのような線形空間をなめらかにつなぎ全体の多様性を推定する

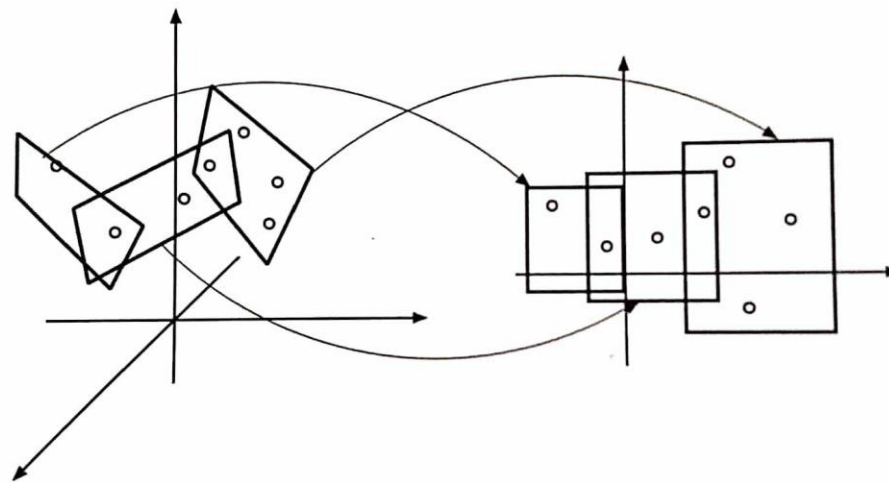


図 3.8 局所線形埋め込み法の概念図。複数の局所的な線形空間のつなぎあわせで曲がった部分構造を抽出する。

[1]で各サンプル点 $\boldsymbol{x}^{(i)}$ を近傍の点 $\{x^{(j)} \in \mathcal{N}_i\}$ で表現する
これを求めたのち

$$\min_W \left\| \boldsymbol{x}^{(i)} - \sum_{j \in \mathcal{N}_i} W_{ij} \boldsymbol{x}^{(j)} \right\|^2$$

という最小化問題を解く。ここで重み W_{ij} には

$$\sum_j W_{ij} = 1$$

という制約をおく

[2]で線形モデルをなめらかにつないでいく。簡単のために多様体上の座標は1次元で β とする

$x^{(i)}$ が含まれる近傍系のすべてについて、対応する多様体上の座標 β_i が

$$\min_{\beta} \sum_{i=1}^n \left(\beta_i - \sum_{j \in \mathcal{N}_i} W_{ij} \beta_j \right)^2$$

となるように $\beta = (\beta_1, \dots, \beta_n)^T$ を定める。さらに

$$\|\beta\|^2 = 1$$

という制約をおく

ラグランジュ未定乗数法を適用するとラグランジュ関数は

$$\|(I - W)\beta\|^2 - \lambda(\|\beta\|^2 - 1)$$

β で微分することで

$$(I - W)^T(I - W)\beta = \lambda\beta$$

という固有値問題に帰着され、最小固有値に対する固有ベクトルが解となる

$(I - W)^T(I - W)$ を展開すると $I - W - W^T + W^T W$ となり、この固有値最小化は

$$(W + W^T - W^T W)\beta = (1 - \lambda)\beta$$

と書けるので

$$\tilde{K} = W + W^T - W^T W$$

という固有値最大化問題と等価になる

ただし、これは正定値である保証がないので適当な正の数 c を用いて

$$K = \tilde{K} + cI_n$$

とすれば正定値となり、この行列に対するカーネル主成分分析と見なすことができる

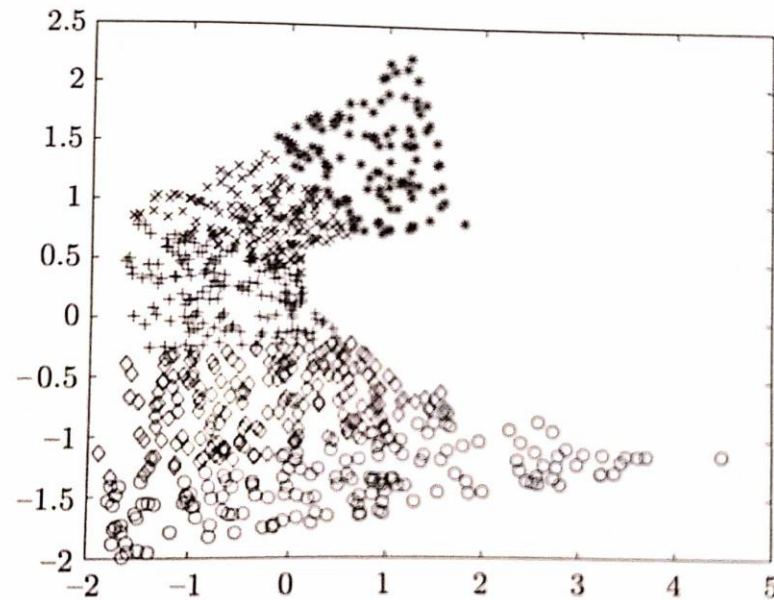


図 3.9 スイスロールデータに対する局所線形埋め込み法の
実行例。シート上に広がっている様子がわかり、 $* \rightarrow x \rightarrow$
 $+ \rightarrow \diamond \rightarrow \circ$ という順番も保存されている。