プログラミング基礎演習レポート 2017-1

長谷川禎彦

以下の点を守ってレポートを提出すること．

注意点

レポートは以下の２つを提出すること

- プログラムのソースコード（C 言語）
- レポート本体（doc, docx, pdf, odt）．英語でも良い．

プログラミングのレポートではプログラムソースや実行結果のみを送る人がいるが，プログラムのソースだけでは何をするものなのか分からないため，採点出来ない．プログラムの作成にあたってどのような工夫をして，それをどのように実現したのか，レポート本体に書くこと．レポートでは以下の点を書くことが一般的である．

- 導入
- 手法・結果
- 考察
- 参考文献（あれば）

レポート本体の提出フォーマットは Microsoft Word（docx, doc），PDF，OpenOffice Writer のどれかで提出する．なお，レポート本体のページ数に上限はないが，無意味な結果の羅列による水増しは却って印象を悪くする．

- 提出〆切り：2017/12/18 の 23:59
- 提出方法：作成したソースファイル（ファイル名は自由．一つのプログラムを複数ファイルに分割しても良い）とレポート本体（doc，docx，pdf，odt 等）を「学籍番号.zip」としてまとめる．作成した zip ファイルをホームページの提出フォームから提出する（「ファイル」と書かれた所から，ソースとレポート本体の zip ファイルを添付する）．その際，課題の選択を「レポート１」とすること（「小課題」ではない）．
- なお，全ての問題において大枠として題意を満たしていれば，方法・内容とも自由に変更して良い．また，問題の拡張や手法の改良を行って良い．

1. 課題（必須）

$\boldsymbol{A}$を$n \times n$の実対称行列する．また，$\boldsymbol{A}$は半正定値行列とする（$\boldsymbol{A}$の固有値は全て非負）．$\boldsymbol{A}$の固有値は特性多項式によって計算出来るが，nが非常に大きいときには固有値・固有ベクトルを計算することは難しい．一方で，多くの応用では，大きいいくつかの固有値・固有ベクトルのみを知れば十分な場合がある．そこで，Power method と呼ばれる方法で，簡単に最大の固有値・固有ベクトルを計算してみる．

今， $\boldsymbol{A}$と 0 以上の整数 tに対してn次元列ベクトル$\boldsymbol{x}(t)$を以下のように定義する．

$$\boldsymbol{x}(t) \equiv \boldsymbol{A}^t \boldsymbol{x}(0). \qquad (1)$$

ここで$\boldsymbol{A}^t$は行列$\boldsymbol{A}$のt乗である．この時，$\boldsymbol{x}(0)$は任意であるが，$\boldsymbol{A}$の固有ベクトル$\boldsymbol{v}_i$（列ベクトル）の

和に展開する（対称実行列の固有ベクトルは直交している）.

$$\boldsymbol{x}(0) = \sum_{i=1}^{n} c_i \boldsymbol{v}_i. \tag{2}$$

ここで，c_iは展開係数である．なお，後のために固有ベクトル$\boldsymbol{v}_i$は正規化されているとする．つまり$\|\boldsymbol{v}_i\| = 1$（各要素の二乗を足したものの平方根は1）．式(1)は，式(2)を用いると以下のようになる．

$$\boldsymbol{x}(t) = \boldsymbol{A}^t \boldsymbol{x}(0) = \boldsymbol{A}^t \sum_{i=1}^{n} c_i \boldsymbol{v}_i = \sum_{i=1}^{n} c_i \kappa_i^t \boldsymbol{v}_i = \kappa_1^t \sum_{i=1}^{n} c_i \left[\frac{\kappa_i}{\kappa_1}\right]^t \boldsymbol{v}_i. \tag{3}$$

ここでκ_iは固有値，特にκ_1は最大の固有値を表す．$\boldsymbol{v}_1$をκ_1に対応する固有ベクトルとすると$\kappa_1 > \kappa_i$（$i \neq 1$，ここでは縮退した場合は考えない）より，$t \to \infty$の極限で$\boldsymbol{x}(t) \propto \boldsymbol{v}_1$となる．つまり，適当な初期値$\boldsymbol{x}(0)$（例えば全ての要素が1）に行列$\boldsymbol{A}$をかけ続けると，固有ベクトル$\boldsymbol{v}_1$（の定数倍）に近づく（ただし$\boldsymbol{x}(0)$を運悪く$\boldsymbol{v}_1$と直交するものを選ぶとそうならない）．最大の固有値κ_1は，十分に大きいtに対して，以下の関係式が成り立つ：

$$\kappa_1 = \boldsymbol{v}_1^\top(t) \boldsymbol{A} \boldsymbol{v}_1(t). \tag{4}$$

このように，特性多項式を解くことなく固有値と固有ベクトルを計算できる．

これをもとにして，与えられた半正定値・実対称行列の最大固有値とそれに対応する固有ベクトル（正規化されている）を求めるプログラムを書け．与える行列はファイルにより与えよ．ホームページにテスト用の行列データがある（test_matrices.zip）．これには4，16，64次元の半正定値・実対称行列と固有値のデータが入っている．

2. 課題（自由）

課題1の$\boldsymbol{A}$に対して，二番目に大きい固有値とそれに対応する固有ベクトルを計算してみる．

実対称行列はスペクトル分解が可能なので，以下のように表すことが出来る（これは$\boldsymbol{A}$側に$\boldsymbol{V}$を集めれば対角化になっていることに注意）．

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top. \tag{5}$$

ここで対角行列$\boldsymbol{D} = \mathrm{diag}(\kappa_1, \kappa_2, \dots, \kappa_n)$ ($\kappa_1 > \kappa_2 > \cdots > \kappa_n$，ここでは縮退した場合は考えない)，$\boldsymbol{V}$は直交行列である．正規化された固有ベクトル$\boldsymbol{v}_i$を使えば，直交行列$\boldsymbol{V}$は

$$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n]. \tag{6}$$

と表される．式(5)はわかりやすく展開すると

$$\boldsymbol{A} = \kappa_1 \boldsymbol{v}_1 \boldsymbol{v}_1^\top + \kappa_2 \boldsymbol{v}_2 \boldsymbol{v}_2^\top + \cdots + \kappa_n \boldsymbol{v}_n \boldsymbol{v}_n^\top. \tag{7}$$

κ_1は課題1で計算出来たとする．この時行列$\boldsymbol{B}$を以下で定義する．

$$\boldsymbol{B} \equiv \boldsymbol{A} - \kappa_1 \boldsymbol{v}_1 \boldsymbol{v}_1^\top = \kappa_2 \boldsymbol{v}_2 \boldsymbol{v}_2^\top + \cdots + \kappa_n \boldsymbol{v}_n \boldsymbol{v}_n^\top. \tag{8}$$

この操作によって，$\boldsymbol{A}$において二番目の固有値κ_2を最大固有値に持つ新たな行列$\boldsymbol{B}$が出来る．つまり行列$\boldsymbol{B}$に対して，Power methodを適用すれば二番目の固有値が計算できる．同様に3番目以降も計算可能である．

これをもとにして，与えられた半正定値・実対称行列の二番目に大きい固有値とそれに対応する固有ベクトル（正規化されている）を求めるプログラムを書け．与える行列はファイルにより与えよ．

3. 課題（自由）

主成分分析（PCA：Principle Component Analysis）は観測データの特徴抽出に用いられる方法である．n次元のN個の観測データ$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$を考える．ここで観測点$\boldsymbol{x}_i$は$n$次元の列ベクトルである．観測データは図 1 に示すように，あるばらつきをもって分布している．このばらつきは完全なランダムではなく，図 1 の矢印で示すように，主要方向がある場合が多い．ここでは，主要方向は，分散が最大化させる方向であると考える．この主要方向ベクトルは，実はデータ$\boldsymbol{X}$の共分散行列$\boldsymbol{C}$の最大の固有値に対応する固有ベクトルになっている（つまり$\boldsymbol{v_1}$）．n番目の主要方向は$n-1$番目までの主成分に直交するが，$\boldsymbol{C}$のn番目の固有値に対応する固有ベクトルになっている．この導出の詳細は文献[Bishop]を参照せよ．

データ全体を行列$\boldsymbol{X}$で表した場合，$\boldsymbol{X}$のi行j列目の値X_{ij}は，j番目のデータの，i種類目のデータの値とする（つまり$\boldsymbol{X}$は$n\times N$行列）．この時，（不偏）共分散行列$\boldsymbol{C}$のa行b列目の要素C_{ab}は

$$C_{ab}\equiv\frac{1}{N-1}\sum_{l=1}^{N}(X_{al}-\overline{X_a})(X_{bl}-\overline{X_b}),\qquad \overline{X_a}\equiv\frac{1}{N}\sum_{l=1}^{N}X_{al}, \tag{9}$$

である．なお，共分散行列は一般に半正定値・実対称行列になる．

与えられたデータに対して PCA を適用し，第 1 主要方向ベクトル，第 2 主要方向ベクトル（可能ならさらに第n主成分）を計算するプログラムを書け．また，データ$\boldsymbol{X}$を各主要方向（$\boldsymbol{v}_1$や$\boldsymbol{v}_2$）に射影したものを主成分という．主成分を出力できるようにもしてみよ．

データはファイル入力によって与えるようにせよ．データは何でも良いが，例えば https://archive.ics.uci.edu/ml/datasets.html などから用いよ（この中で Iris などは次元が小さく使いやすい）．

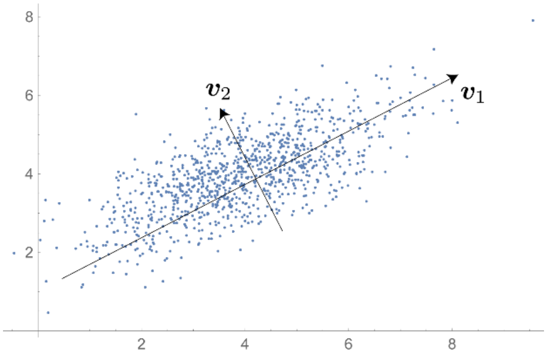

図1 ：データ（点）と最主要方向$\boldsymbol{v}_1$，第二主要方向$\boldsymbol{v}_2$．$\boldsymbol{v}_1$方向は，データの分散が最大化されている．$\boldsymbol{v}_2$は$\boldsymbol{v}_1$に直交する方向で，次に分散が大きい成分となる．

参考文献

[Bishop] C. M. Bishop, “Pattern recognition and machine learning”