プログラミング基礎演習レポート 2017-2

長谷川禎彦

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（Python ソース（.py）または Jupyter notebook（.ipynb））
- レポート本体（doc, docx, pdf, odt）．英語でも良い．

以下の点に留意せよ．なお，その他の注意点は第一回レポートに準ずる．

- 提出〆切り：2018/2/13 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 2」とすること（フォームからは 12/20 以降に提出可能となる）．
- 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．問題の拡張や手法の改良を行って良い．なお，Python の Pandas, Numpy, Scipy, Matplotlib 等（これらに限らず）ライブラリは基本的に自由に用いよ．ただし，ICA それ自体を行うライブラリを用いてはならない（例えば，scikit-learn は利用不可）．

カクテルパーティ効果とは，たくさんの人が異なる会話をしている時でも，自分が興味のある人の会話は聞き取ることができるというよく知られた効果である．このように，人間には複数の音源から必要な情報だけを抽出する機構が備わっている．これを模したアルゴリズムが独立成分分析（ICA : Independent Component Analysis）である．本レポートでは，ICA を Python を用いて実装してみる．

観測データを$\boldsymbol{x} \in \mathbb{R}^n$で表し，信号源データ（未知）を$\boldsymbol{y} \in \mathbb{R}^n$とする（どちらも列ベクトル）．実際のデータは$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_N\}$のように$N$個のデータベクトルによって与えられ，それに対応して信号源は$\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$とする．ここでは，観測データ$\boldsymbol{x}$は平均が$\boldsymbol{0} \in \mathbb{R}^n$になるように調整している（つまり$\boldsymbol{x} \leftarrow \boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]$）．なお，$\mathbb{E}[\boldsymbol{x}]$は平均（期待値）を表し，$\mathbb{E}[\boldsymbol{x}] = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$とする．

観測データは，信号源の重ね合わせのため，以下の線形関係があると仮定する

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y}. \tag{1}$$

ここで$\boldsymbol{A} \in \mathbb{R}^{n \times n}$の行列である．これより，もし$\boldsymbol{W} = \boldsymbol{A}^{-1}$が分かれば，信号源$\boldsymbol{y}$は

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}, \tag{2}$$

によって復元可能である．つまり，ICA は行列$\boldsymbol{W}$を求める問題に他ならない．

中心極限定理より，任意の独立な確率変数の和は正規分布に収束する．中心極限定理は無限の和の極限において成立する．しかし，たった二つの確率変数q_1とq_2を足した場合でも，足した後の$q_1 + q_2$の方が，q_1やq_2よりも正規分布に近い分布となる場合が多い．逆に言えば，正規分布から遠い分

布を持つデータは，データが足し合わさっていないことを意味する．この性質を用いると，$\boldsymbol{y}$ができるだけ正規分布から遠くなるように$\boldsymbol{W}$を決めれば，信号源$\boldsymbol{y}$が計算できそうである．実は，尖度（kurtosis）と呼ばれる統計量は，正規分布の時 0 になり，正規分布から乖離すると 0 から大きくずれる（尖度は正負の値をとれる）．平均が0の確率変数$y \in \mathbb{R}$の尖度は以下で定義される．

$$\text{kurtosis}[y] = \mathbb{E}[y^4] - 3\mathbb{E}[y^2]^2, \tag{3}$$

つまり，|kurtosis(y)|を最大化するように，$\boldsymbol{W}$を求めれば ICA が実現できることがわかる．

現実にはこれだけではうまくいかない．以下に述べる白色化と呼ばれる操作を，ICA の前段階に挟む必要がある．今，観測データの共分散行列を

$$\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top], \tag{4}$$

とする（$\boldsymbol{x}$は平均が 0 のように取っていることに注意せよ）．共分散行列$\boldsymbol{\Sigma}$は以下のように対角化させることができる

$$\boldsymbol{\Sigma} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\top, \tag{5}$$

ここで$\boldsymbol{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$，$\lambda_i$は$\boldsymbol{\Sigma}$の固有値である．$\boldsymbol{E}$は直交行列になっている．この時以下のように$\boldsymbol{V}$を導入する

$$\boldsymbol{V} = \boldsymbol{E}\boldsymbol{D}^{-1/2}\boldsymbol{E}^\top, \tag{6}$$

ここで，$\boldsymbol{D}^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \dots, \frac{1}{\sqrt{\lambda_n}}\right)$である．新しい確率変数$\boldsymbol{z}$を以下で導入する

$$\boldsymbol{z} = \boldsymbol{V}\boldsymbol{x}, \tag{7}$$

この時$\boldsymbol{z}$の共分散行列は単位行列となる，つまり$\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \mathbb{I}$．共分散を単位行列にする操作を白色化という（$\boldsymbol{z}$は白色化されている）．ICA では観測データ$\boldsymbol{x}$をまず白色化し，式(7)を用いて$\boldsymbol{z}$を計算する．$\boldsymbol{z}$を用いると，式(2)は以下のようになる

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{z}, \tag{8}$$

（この$\boldsymbol{W}$と式(2)の$\boldsymbol{W}$は当然一般には異なる）．式(8)において，信号源の一つに注目し，それを$\boldsymbol{y} \in \mathbb{R}$とする．この時，式(8)より

$$y = \boldsymbol{w}^\top \boldsymbol{z}, \tag{9}$$

ここで$\boldsymbol{w} \in \mathbb{R}^n$は列ベクトルである．$\boldsymbol{w}^\top$を縦に積んだものが$\boldsymbol{W}$になっている．計算の詳細は省くが，$\boldsymbol{y}$の尖度の絶対値を最大化させるには，以下の繰り返しアルゴリズムを適用することで可能である（詳細は参考文献参照）．

1. $\boldsymbol{w}$に対して初期値を適当に選ぶ．正規化する$\boldsymbol{w} \leftarrow \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$.
2. $\boldsymbol{w} \leftarrow \mathbb{E}[\boldsymbol{z}(\boldsymbol{w}^\top \boldsymbol{z})^3] - 3\boldsymbol{w}$を計算する
3. $\boldsymbol{w} \leftarrow \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$によって正規化する
4. 収束していなければ，2 に戻る．収束していれば終了する．

このアルゴリズムを収束するまで実行すると，情報源$\boldsymbol{y}$の尖度の絶対値が最大化され，信号源の一つが求まる．異なる信号源を求めるには，1.において異なる初期値$\boldsymbol{w}$でアルゴリズムを何度か実行すればよい（これが一番簡単）．ほかの方法としては，直交化を各ステップで使う方法もある（詳細は参考文献参照）．

1. 課題（必須課題）

ホームページの dat1.txt と dat2.txt は 2 つの信号源を異なる比率で混合したデータである（report2_kadai1_data1.zip, 図 1 参照）. Python で ICA を実装し, dat1.txt と dat2.txt に ICA を適用し, 2 つの信号源を同定せよ.

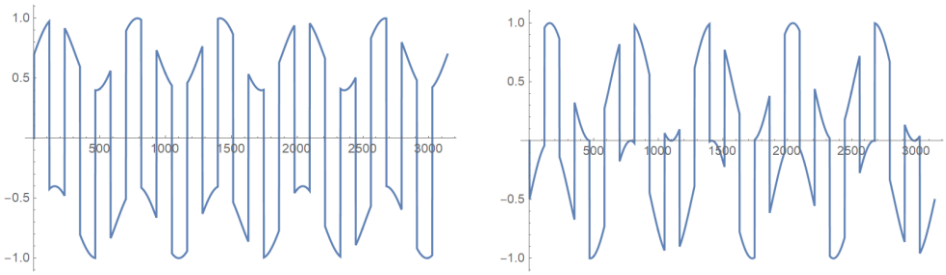

図1 ：テスト用のデータ．左から dat1.txt, dat2.txt.

2. 課題（自由課題）

ホームページに音声ファイルが入った圧縮ファイルが 3 つある. その内訳は以下のとおりである. これらのデータに作成した ICA を適用し, 音源を分離せよ. なお, 分離後の wav ファイルの提出は, 3 つのうちどれか一つでよい（report2_kadai2_data3 を提出する場合は mp3 等で圧縮してサイズを小さくすること．mp3 圧縮は https://online-audio-converter.com/ja/などで可能）.

- report2_kadai2_data1.zip
  - 二人の話者が同時に話したものを，（仮想的に）異なる場所で録音したデータが入っている．wav 形式，16bit，8000Hz，モノラル
- report2_kadai2_data2.zip
  - 同じ声の 3 人の話者が同時に話したものを，（仮想的に）異なる場所で録音したデータが入っている．wav 形式，16bit，8000Hz，モノラル
- report2_kadai2_data3.zip
  - 二つの曲を同時に演奏したものを，（仮想的に）異なる場所で録音したデータが入っている．wav 形式，16bit，44100Hz，モノラル

Python での wav ファイルの読み込み・書き込みは scipy.io.wavfile.read や scipy.io.wavfile.write を用いると簡単にできる（他のライブラリでももちろん良い）.

3. 課題（自由課題）

音声だけではなく, 画像に対しても ICA を適用することが可能である. ホームページに image1.png と image2.png が入ったファイルがある（report2_kadai3_data1.zip）．これは，二つの画像の異なる比率の重ね合わせである（図 2 参照）．これに作成した ICA を適用し，画像を分離せよ．なお画像はグレースケールの png フォーマットである.

4. 課題（自由課題）

課題で与えられたデータ以外（の面白いデータ）に対して，作成した ICA を適用してみよ.

図2 ：テスト用のデータ．左から image1.png, image2.png.

参考文献

- Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja. Independent component analysis. Vol. 46. John Wiley & Sons, 2004.
- [上の翻訳] Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja 著，根本幾，川勝真喜翻訳，独立成分分析～信号解析の新しい世界～，東京電機大学出版局