データ解析　課題　第４回

概要：　「ノイズで汚された画像データ」を読み込み，ノイズを除去して画像復元をする．MCMC **法の一種であるギブスサンプラを実装する．**

提出方法：レポート（紙）とファイル（メール添付）の両方

締め切り：６月１６日（月）の 15:00 **必着（紙とメールの両方）**

1. **レポート（紙）はいつもどおりレポートボックスへ提出**
2. **ファイルはメール添付にて，講義中に説明した「データ解析」のアドレスへ送信．添付するのは，以下で説明する**"rep4-0412345.rda"**（復元画像）と**"rep4-0412345.R"**（スクリプト）です．メールの題名を「データ解析課題第４回** 0412345**」として，二つのファイルを添付します．ただし** 0412345 **を学籍番号におきかえます．**

4.1 ギブスサンプラ

m 次元ベクトルの系列 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ を次のアルゴリズムで生成する．このマルコフ連鎖の推移確率を決める条件付確率分布 $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ を求め，それが詳細釣合 $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)f(\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})f(\boldsymbol{x}_{t+1})$ を満たすことを示せ．ただし，以下の記法を用いる．m 次元ベクトル $\boldsymbol{x}$ の i 番目の要素を $\boldsymbol{x}[i]$ と書く．$\boldsymbol{x}$ から $\boldsymbol{x}[i]$ を取り除いた $m-1$ 次元ベクトルを $\boldsymbol{x}[-i]$ と書く．$\boldsymbol{x}[-i]$ の周辺確率分布を

$$h_i(\boldsymbol{x}[-i]) = \int_{-\infty}^{\infty} f(\boldsymbol{x})\, d\boldsymbol{x}[i]$$

と書き，$\boldsymbol{x}[-i]$ を与えたときの $\boldsymbol{x}[i]$ の条件付確率分布を

$$f_i(\boldsymbol{x}[i] \mid \boldsymbol{x}[-i]) = \frac{f(\boldsymbol{x})}{h_i(\boldsymbol{x}[-i])}$$

と書く．

1. **初期値 $\boldsymbol{X}_1$ を選び $t = 1$ とする．**
2. **添え字 i をランダムに選ぶ．**
3. **$V \sim f_i(V|\boldsymbol{X}_t[-i])$ を１個生成．**
4. **$\boldsymbol{X}_{t+1}[i] = V$，$\boldsymbol{X}_{t+1}[-i] = \boldsymbol{X}[-i]$ と代入する．**
5. **t の値をひとつ増やして 2へ戻る．**

4.2 「画像復元」のギブスサンプラ

２次元配列 $\boldsymbol{x}[i]$ と $\boldsymbol{y}[i]$, $i = (1,1), \ldots, (m,m)$ の各要素が $\{+1, -1\}$ のどちらかの値を取る．x を「真の画像」，y を「ノイズで汚された画像データ」とする．x の各要素に確率 ϵ $(0 < \epsilon < 1)$ で独立にノイズが入り反転するモデル (つまり $+1 \to -1$ または $-1 \to +1$) を考える．

$$f(\boldsymbol{y}[i]|\boldsymbol{x}[i]) = \begin{cases} 1-\epsilon & \boldsymbol{y}[i] = \boldsymbol{x}[i] \\ \epsilon & \boldsymbol{y}[i] \neq \boldsymbol{x}[i] \end{cases}$$

(i) **このとき，x を与えたときの y の条件付分布が**

$$f(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left(\lambda \sum_i \boldsymbol{x}[i]\boldsymbol{y}[i]\right)$$

とかけることを示せ．ただし $\lambda = \frac{1}{2}\log(\frac{1-\epsilon}{\epsilon})$，$\sum_i$ はすべての画像要素に関する和，$\propto$ は x, y に関して比例（それ以外は定数とみなす）である．

(ii) x **の事前分布を**

$$f(\boldsymbol{x}) \propto \exp\left(\gamma \sum_{(i,j)} \boldsymbol{x}[i]\boldsymbol{x}[j]\right)$$

で与える．ただし $\gamma > 0$ は定数，$\sum_{(i,j)}$ はすべての「隣接」に関する和（画像要素の４近傍を「隣接」と定

義），$\propto$ は x に関して比例とする．このとき，x の事後分布，すなわち y を与えたときの x の条件付分布が

$$f(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left(\lambda \sum_i \boldsymbol{x}[i]\boldsymbol{y}[i] + \gamma \sum_{(i,j)} \boldsymbol{x}[i]\boldsymbol{x}[j]\right)$$

とかけることを示せ．ただし $\propto$ は x に関して比例（y は定数とみなす）である．

4.3 画像の表示

R のバイナリ形式のデータ rep4-qustion.rda を読み込み，画像として表示・印刷して，レポートに含める．

```
> load("rep4-question.rda") # データファイルの読み込み
> dim(y) # 行列の次元
[1] 90 90
> y[1:5,1:5] # 一部を表示．２値画像です．(+1 と-1)
     [,1] [,2] [,3] [,4] [,5]
[1,]   -1    1   -1   -1   -1
[2,]   -1   -1    1   -1   -1
[3,]   -1   -1    1    1   -1
[4,]   -1   -1   -1   -1    1
[5,]   -1   -1   -1    1   -1
> bw <- rev(gray((0:64)/64)) # 64 階調のグレースケール
> image(y,axes=F,col=bw) # データのプロット
```

4.4 画像復元のプログラム作成

画像復元のプログラム（スクリプト）を作成せよ．

図 1　ノイズで汚された画像データ y

- プログラムは次の形式で作成してから実行する．下記のコメントに注意すること．

```
## データ解析 課題 第４回
## 東工太郎 0412345
set.seed(345) # 学籍番号の下３桁を指定する．
load("rep4-question.rda") # データファイルの読み込み
... 中略 ...
... 復元画像の結果はかならず x という名前のオブジェクトにする ...
save(x,file="rep4-0412345.rda") # 復元画像をバイナリ形式で保存
```

- スクリプトのファイル名は rep4-0412345.R の形式にする．
- 再現性のため，set.seed の行は必ず入れてる．ただし，指定する値は各自の学籍番号の下 3 桁にする．
- save の行も必ずいれる．オブジェクト名を x にすること，および，ファイル名を指定された形式（学籍番号を含める）にしてください．これが守られていないと，レポートが採点されないことがある．
- 得られた復元画像を image で表示するようにスクリプトを書く．
- アルゴリズムの概要や工夫した点をレポートに書くこと．またプログラムには読みやすいように適宜コメントをいれる．
- 講義で説明した方法にとらわれず，創意工夫のあるものを評価します．

4.5 プログラムの実行

上記で作成したプログラムを実行せよ．

- 復元画像のバイナリファイルをメール添付してください．「真の画像」からの誤差を成績に反映させます．もし正しい形式になっていないと，その得点が０になってしまいます．
- スクリプトを実行したときの，コンソール出力および復元画像を印刷してレポートに含める．

```
> load("rep4-0412345.rda") # 回答例：「復元画像」 x の読込
> mean(x!=y0)*100 # 「真の画像」 y0 からの誤差 (%)
[1] 5.851852
```