

2次元ガウス分布のグラフの作成

@ ANEMPTYARCHIVE

2022/02/06

多次元ガウス分布の定義式

多次元ガウス分布は、次の式で定義される確率分布です。

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$$

- D : 次元数
- \mathbf{x} : 確率変数の値
- $\boldsymbol{\mu}$: 平均ベクトル
- $\boldsymbol{\Sigma}$: 分散共分散行列

2次元のグラフで表現するため $D = 2$ とすると、それぞれ次の形状になります。

$$\mathbf{x} = (x_1, x_2), \boldsymbol{\mu} = (\mu_1, \mu_2), \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}$$

- σ_1^2 : x_1 の分散 : $\sigma_1^2 > 0$
- σ_2^2 : x_2 の分散 : $\sigma_2^2 > 0$
- $\sigma_{1,2} = \sigma_{2,1}$: x_1, x_2 の共分散

この2次元ガウス分布のグラフを作成します。

利用パッケージ

利用するパッケージを読み込みます。

```
# パッケージの読み込み  
library(tidyverse)  
library(mvnfast)
```

mvnfastは、多次元ガウス分布に関するパッケージです。

格子点の作成

等高線図を描画するのに格子状の点を利用します。
作図の前に、格子点の作成方法を確認します。

値の作成

seq() で、x軸とy軸の値を作成します。

```
# x軸の値を作成
x_vals <- seq(from = 1, to = 5, by = 1)

# y軸の値を作成
y_vals <- seq(from = 11, to = 15, by = 1)
x_vals; y_vals
```

```
## [1] 1 2 3 4 5
```

```
## [1] 11 12 13 14 15
```

- seq() で数列を作成
 - 第1引数from：最小値
 - 第2引数to：最大値
 - 第3引数by：値の間隔

第1引数(from)から第2引数(to)までの値を第3引数(by)に指定した間隔で作成します。

データフレームに格納

作成した値をデータフレームに格納します。

```
# 値をデータフレームに格納
x_vals_df <- tibble::tibble(
  x = x_vals, # x軸の値
  y = y_vals  # y軸の値
)
head(x_vals_df)
```

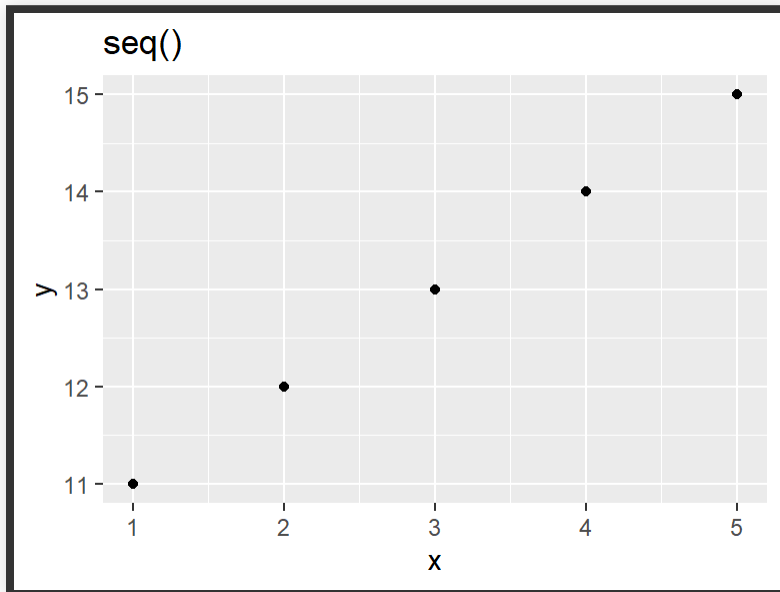
```
## # A tibble: 5 x 2
##       x     y
##   <dbl> <dbl>
## 1     1    11
## 2     2    12
## 3     3    13
## 4     4    14
## 5     5    15
```

ggplot2パッケージで作図するには、作図に用いる値をデータフレームにしておく必要があります。

散布図の作成

散布図を作成して値を確認します。

```
# 散布図を作成
ggplot() +
  geom_point(data = x_vals_df, mapping = aes(x = x, y = y)) + # 散布図
  scale_x_continuous(breaks = x_vals) + # x軸目盛
  scale_y_continuous(breaks = y_vals) + # y軸目盛
  labs(title = "seq()")
```



- `geom_point()` で散布図を描画

格子点に変換

`expand.grid()` で、格子点を作成します。

```
# 格子点を作成
x_grid_df <- expand.grid(
  x = x_vals, # x軸の値
  y = y_vals  # y軸の値
)
head(x_grid_df); tail(x_grid_df)
```

```
##      x      y
## 1 1 11
## 2 2 11
## 3 3 11
## 4 4 11
## 5 5 11
## 6 1 12
```

```
##      x      y
## 20 5 14
## 21 1 15
## 22 2 15
## 23 3 15
## 24 4 15
## 25 5 15
```

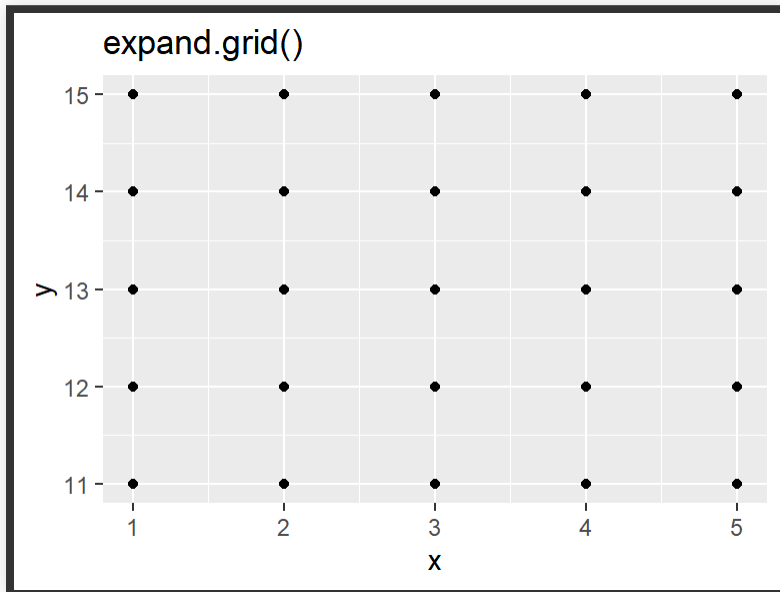
- `expand.grid()` で格子点(直交する点)を作成
 - 列名 = 値として指定
 - 引数に指定した値の全ての組み合わせを持つデータフレームを出力

各行が1つの点に対応し、「x_valsの要素数」掛ける「y_valsの要素数」個の点(組み合わせ)ができます。2つのベクトルの要素数が同じである必要はありません。

格子点の確認

作成した点を確認します。

```
# 散布図を作成
ggplot() +
  geom_point(data = x_grid_df, mapping = aes(x = x, y = y)) + # 散布図
  scale_x_continuous(breaks = x_vals) + # x軸目盛
  scale_y_continuous(breaks = y_vals) + # y軸目盛
  labs(title = "expand.grid()") # ラベル
```



x_vals, y_valsの全ての組み合わせになっているのを確認できます。

2次元ガウス分布の設定と計算

格子点の確認ができたので、2次元ガウス分布の作図を行います。
まずは、分布の設定を行い、確率密度を計算します。

点の設定

作図に利用する $\mathbf{x} = (x_1, x_2)$ の値を作成します。

```
# xの値を作成
x_vals <- seq(from = -5, to = 5, length.out = 51)

# xの点を作成
x_points <- expand.grid(x1 = x_vals, x2 = x_vals) %>%
  as.matrix()
head(x_points)
```

```
##      x1 x2
## [1,] -5.0 -5
## [2,] -4.8 -5
## [3,] -4.6 -5
## [4,] -4.4 -5
## [5,] -4.2 -5
## [6,] -4.0 -5
```

`seq()`で x_1, x_2 の値を作成して、`x_vals`とします。等高線が粗い(点の数が少ない)場合や処理が重い(点の数が多い)場合は、この設定を調整してください。

`expand.grid()`で \mathbf{x} の値(組み合わせ)を作成して、`x_points`とします。ただし、確率密度の計算するために、`as.matrix()`でデータフレームからマトリクスに変換しておきます。

パラメータの設定

平均ベクトル μ と分散共分散行列 Σ を作成します。

```
# 平均ベクトルを指定
mu_d <- c(0, 0)

# 分散共分散行列を指定
sigma_dd <- matrix(c(1, 0, 0, 1), nrow = 2, ncol = 2)
```

μ の値を指定してmu_dとします。

Σ の値を指定してsigma_ddとします。matrix()に $(\sigma_1^2, \sigma_{2,1}, \sigma_{1,2}, \sigma_2^2)$ の順に値を指定します。

分布の計算

設定したパラメータを使って確率密度を計算します。

```
# 2次元ガウス分布を計算
dens_df <- tibble(
  x1 = x_points[, 1], # x軸の値
  x2 = x_points[, 2], # y軸の値
  density = mvnfast::dmvn(X = x_points, mu = mu_d, sigma = sigma_dd) # 確率密度
)
head(dens_df)
```

```
## # A tibble: 6 x 3
##       x1      x2 density
##   <dbl> <dbl>   <dbl>
## 1    -5     -5 2.21e-12
## 2   -4.8     -5 5.89e-12
## 3   -4.6     -5 1.51e-11
## 4   -4.4     -5 3.71e-11
## 5   -4.2     -5 8.76e-11
## 6    -4     -5 1.99e-10
```

- mvnfastパッケージのdmvn()で多次元ガウス分布の確率密度を計算できます。
 - 変数の引数Xにx_points
 - 平均ベクトルの引数muにmu_d
 - 分散共分散行列の引数sigmaにsigma_dd

を指定します。

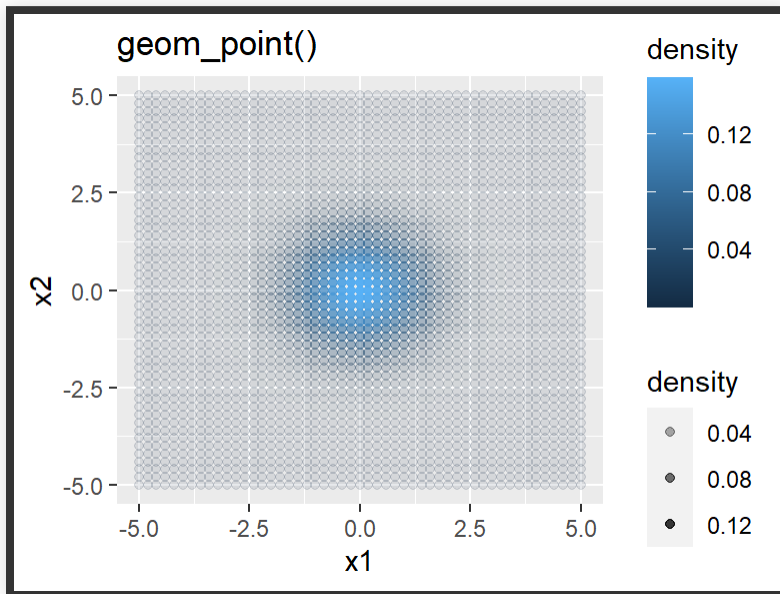
x_1, x_2 の値と、対応する確率密度(計算結果)をデータフレームに格納します。

2次元ガウス分布の作図

分布の設定と計算ができたので、2次元ガウス分布のグラフを作成します。

まずは、散布図による可視化

```
# 散布図を作成
ggplot() +
  geom_point(data = dens_df, mapping = aes(x = x1, y = x2, color = density, alpha = density)) + # 散布図
  labs(title = "geom_point()") # ラベル
```

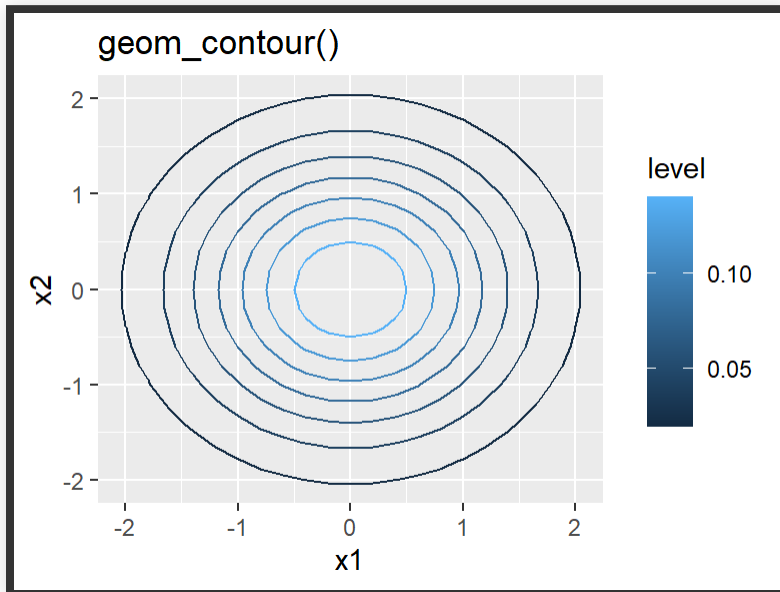


- `geom_point()` で散布図を描画
 - `color`引数：点の色
 - `alpha`引数：透過度

それぞれの引数に`density`列を指定して、確率密度に応じて点ごとに色と濃淡を付けます。

では、等高線図による可視化

```
# 等高線図を作成
ggplot() +
  geom_contour(data = dens_df, mapping = aes(x = x1, y = x2, z = density, color = ..level..)) + # 等高線
  labs(title = "geom_contour()") # ラベル
```

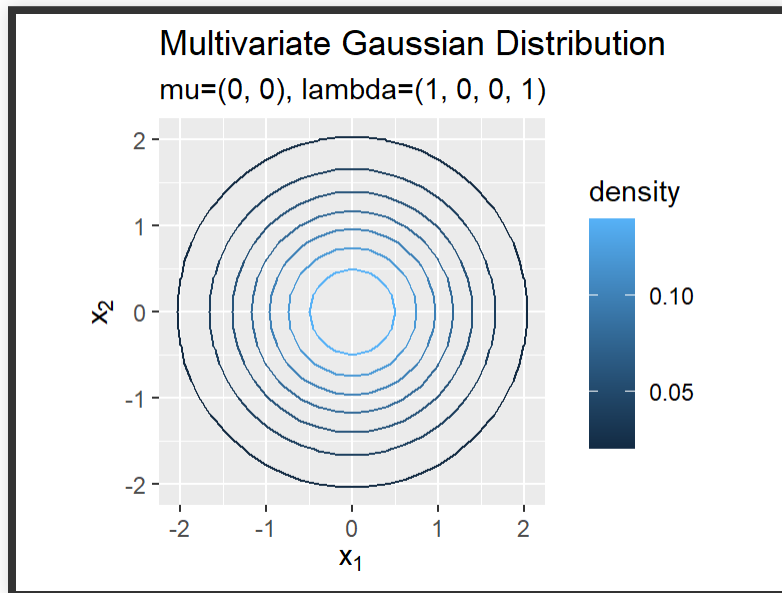


- `geom_contour()` で等高線図を描画
 - `z`引数：`z`軸の値
 - `color`引数：等高線の色
 - `..level..` を指定すると`z`引数の値に応じて色付け

先ほどの散布図に関して、確率密度が同じ点を線で結んだイメージです。

最後に、図の装飾

```
# 2次元ガウス分布を作図
ggplot() +
  geom_contour(data = dens_df, mapping = aes(x = x1, y = x2, z = density, color = ..level..)) + # 等高線
  # 図
  coord_equal() + # アスペクト比
  labs(title = "Multivariate Gaussian Distribution",
        subtitle = paste0("mu=(", paste(mu_d, collapse = ", "), ")",
                           ", lambda=(", paste(sigma_dd, collapse = ", "), ")",
        x = expression(x[1]), y = expression(x[2]),
        color = "density") # ラベル
```



Enjoy!