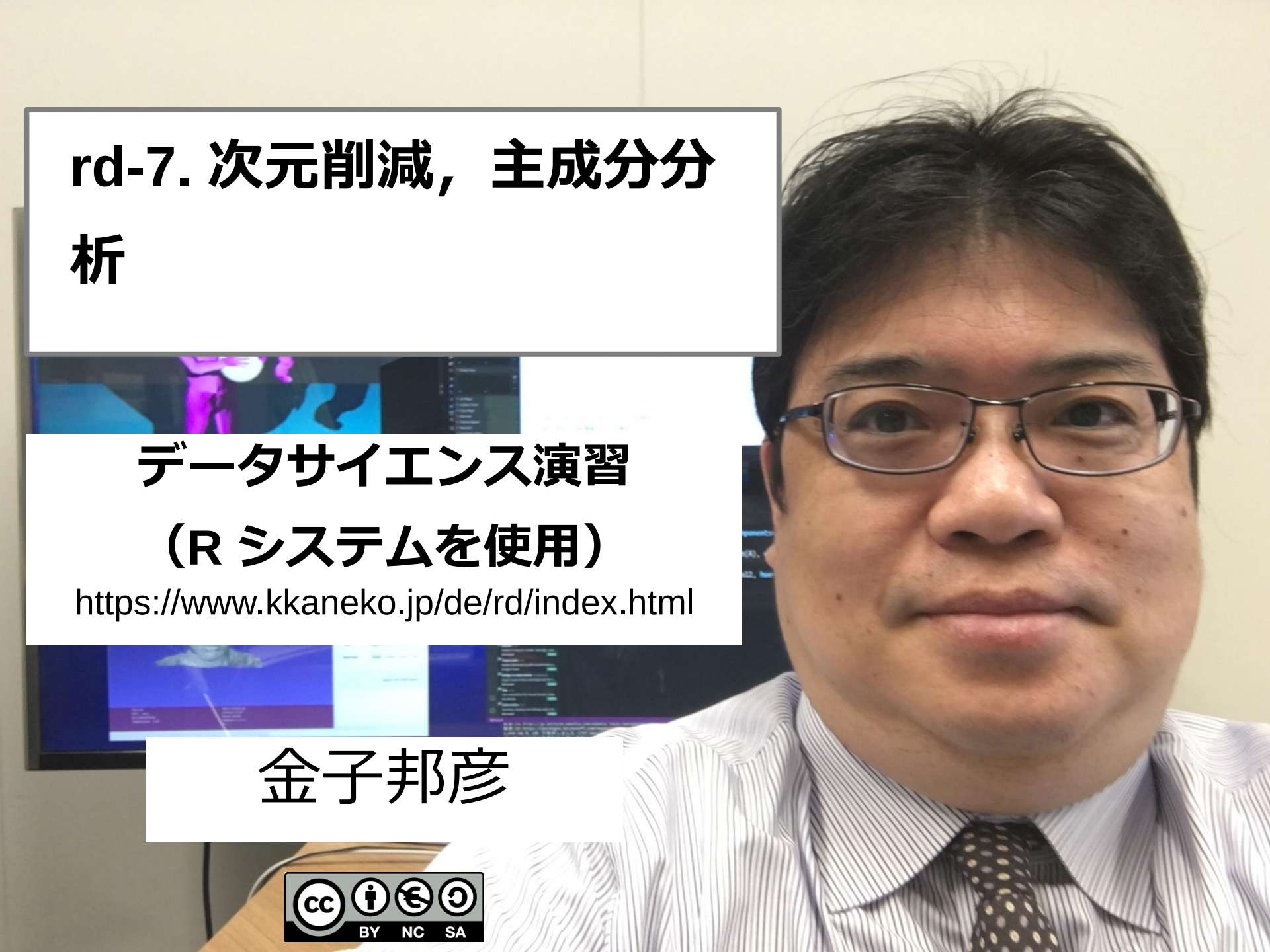
rd-7. 次元削減，主成分分析



データサイエンス演習

（R システムを使用）

https://www.kkaneko.jp/de/rd/index.html

金子邦彦

アウトライン

7-1. 主成分分析と次元削減

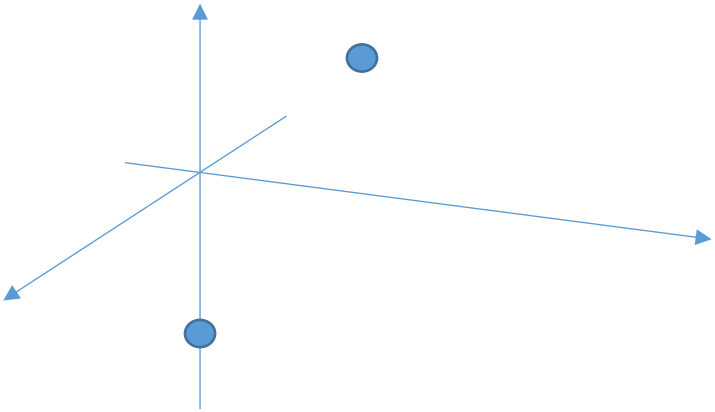
7-2. Rシステムでの主成分分析の実行

7-3. ロバストな主成分分析

7-1 主成分分析と次元削減

データの次元

データの**次元**：**データの表現に必要な最小の情報の数**



空間の中の２つの点

x	y	z
0	-20	0
10	20	0.1

２つのデータ「0 -20 0」と「10 20 0.1」

次元数：３

データの次元

さまざまなデータについて、**次元**を考えることが可能

温度	湿度
25	60
28	70

2 つのデータ「25 60」と
「28 70」

次元数：2

次元削減

データの**次元削減**を考えることが可能

属性 z を削除

x	y	z
0	-20	0
10	20	0.1

x	y
0	-20
10	20

元データ：　次元数は**3**　　次元数は**2**

次元削減の効果

- **可視化のため**

データを散布図などのグラフにするとき、データを２次元や３次元に次元削減

- **本質でない情報の除去のため**

データにノイズが含まれていたり、分析のために不要なものが含まれている場合、次元削減を行う

- **計算の効率化のため**

次元削減によりデータ全体のサイズが少なくなり、計算の効率化ができる

次元削減の手法①

次元削減の単純な方法

- 属性の削除

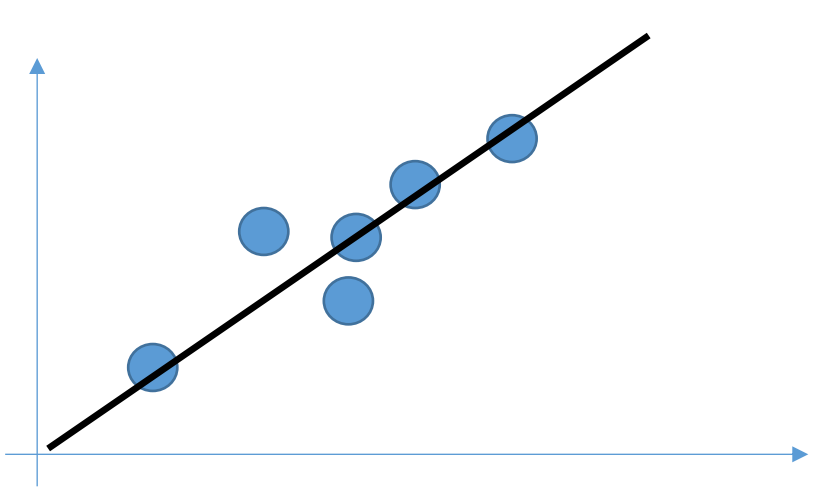
x	y	z
0	-20	0
10	20	0.1

次元数は**3**

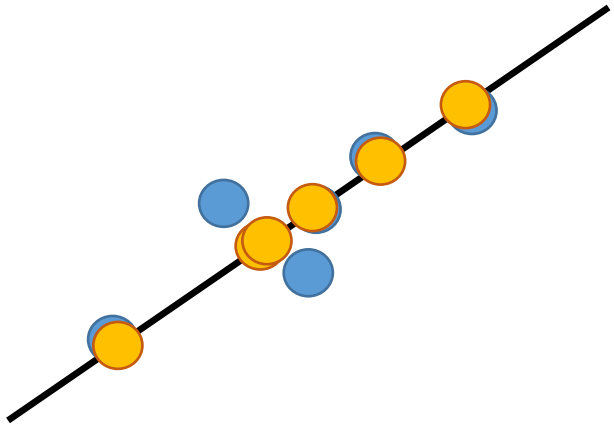
x	y
0	-20
10	20

次元数は**2**

- 近似直線への投影



次元数は**2**



次元数は**1**

次元削減の手法②

次元削減にはさまざまなアプローチがある

主成分分析（PCA），ロバストな主成分分析

データの分散が最大になる方向に軸を見つける（**データの特徴**を最もよく表す**軸**を見つける）．

- **T-SNE**

  データ間の距離を用いる．距離を保つようにしながら次元削減.

- **Linear Discriminant Analysis**

  教師有りの機械学習の技術を使用.

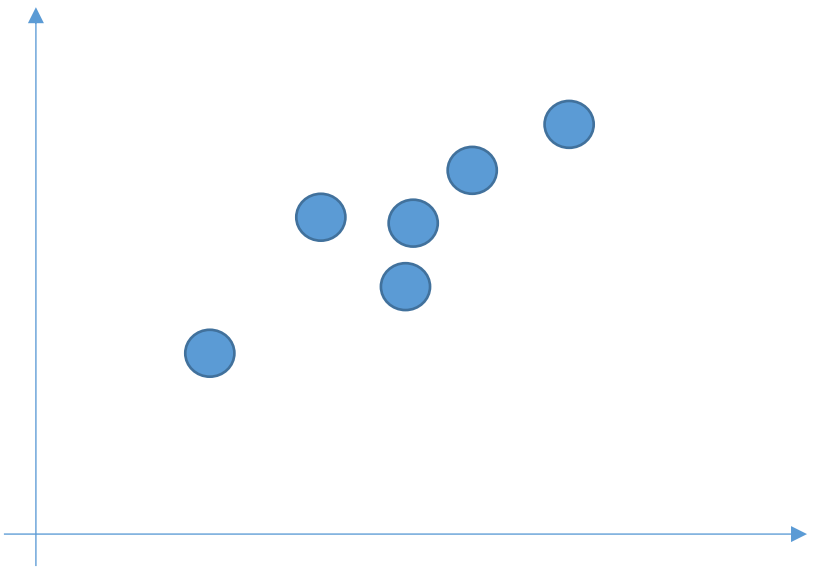
  「データは、種類ごとの正規分布」，「各種類のデータは同じ形状の分布である（共通の共分散行列を持つ）．ただし，平均は違ってもよい」という性質の成り立つデータに有効

- **QDA**

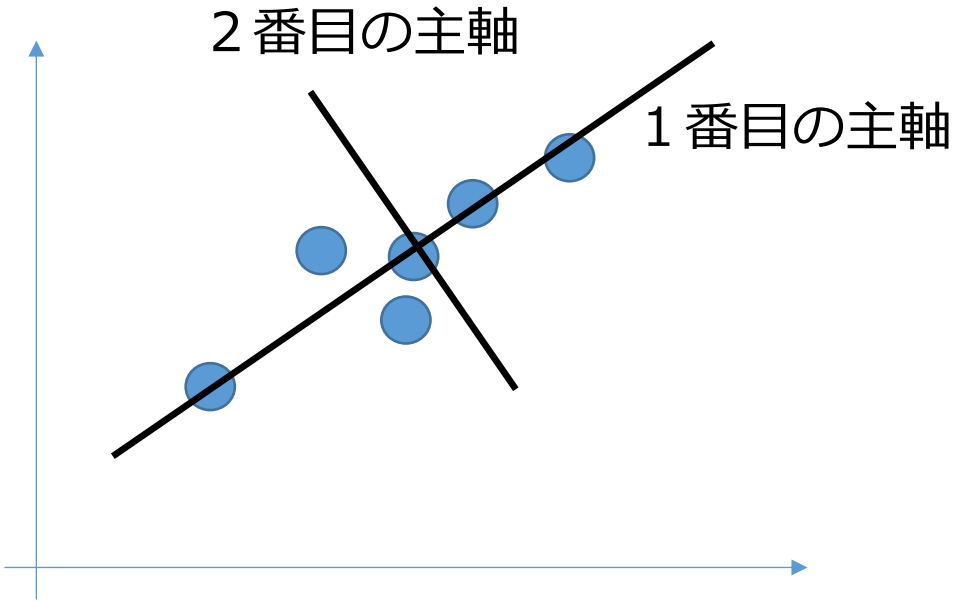
  LDAの改良．「共通の共分散行列を持つ」という仮定を置かない

主成分分析と主軸

- **主成分分析**では、**データの分散が最大となる方向の軸**を**主軸**という
- **主成分分析**では、**元のデータの次元数**と**同じ数**の**主軸**を作成できる



次元数は**2**



主軸を**２**つ

主成分分析と主軸

主成分分析では、**元のデータの次元数**と**同じ数**の**主軸**を作成できる

① 1番目の**主軸**は，データの分散が最大になるような方向の軸

② 2番目の**主軸**は，1番目の**主軸**とは異なる方向で、その方向における分散が最大

1番目の**主軸**に対する**データの成分を取り除いた**残りのデータから行う

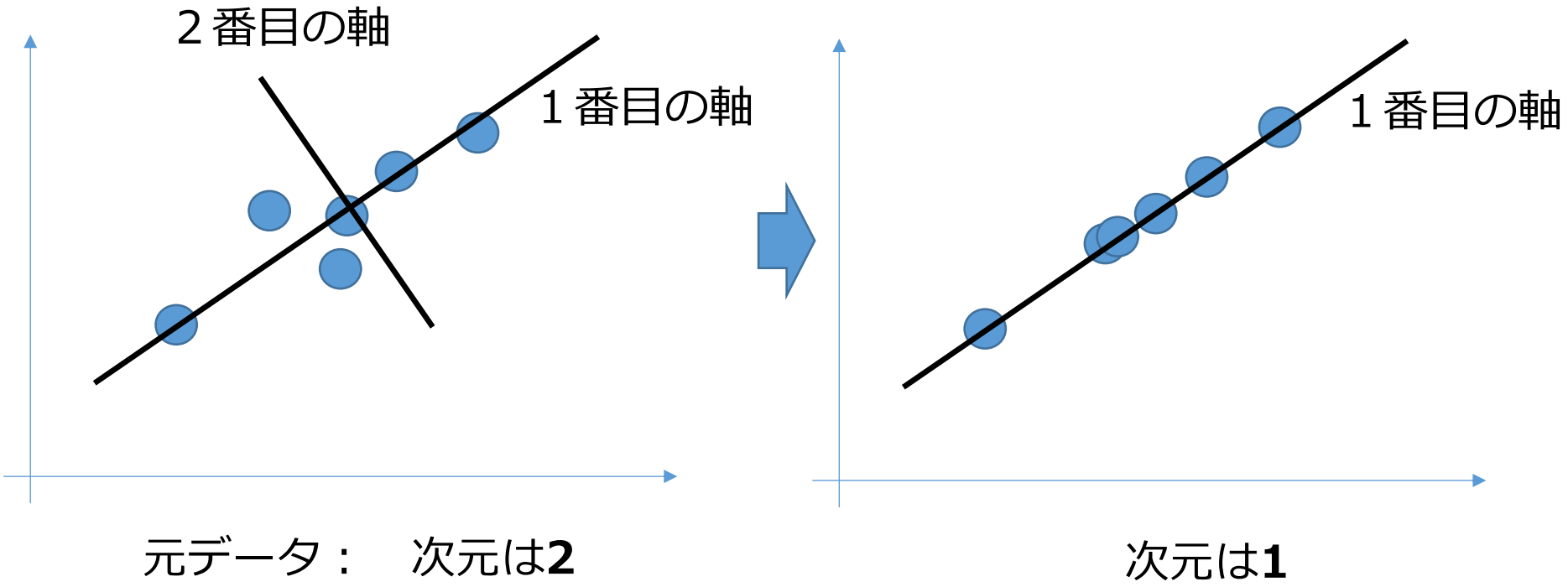
③ 3番目以降も同様

3番目の**主軸**は1番目と2番目の**主軸**に対する**成分を取り除いた**残りのデータから、その方向における分散が最大となるように選ぶ

得られた各主軸は互いに直交（各段階で「成分を取り除く」ので）

主成分分析

- **主成分分析**では，得られた**主軸**の中から、上位の**主軸**を**選び**，下位の**主軸**を削除.
- **選ばれた主軸**に，元データを投影することで，次元削減を行う．この投影は線形変換である.

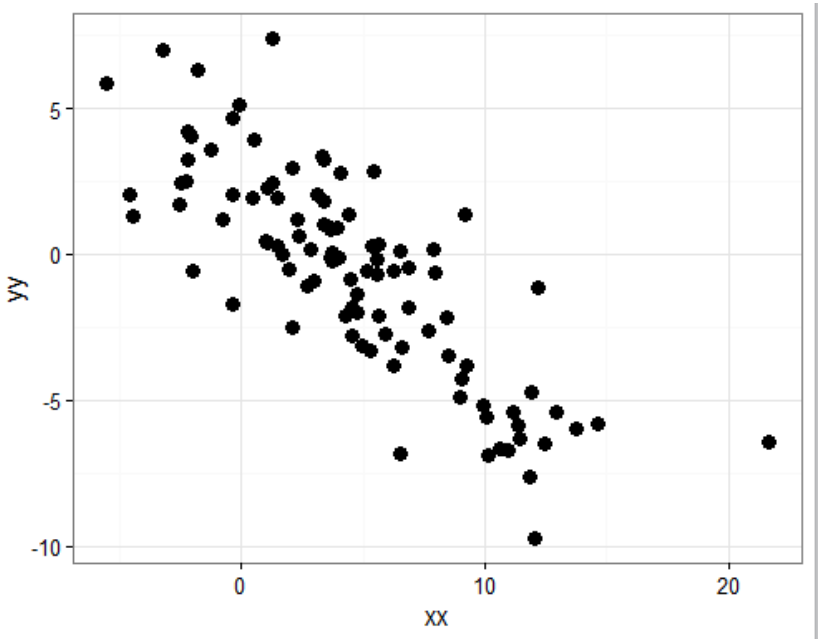


元データ：　次元は**2**

次元は**1**

主成分分析の例

元データ



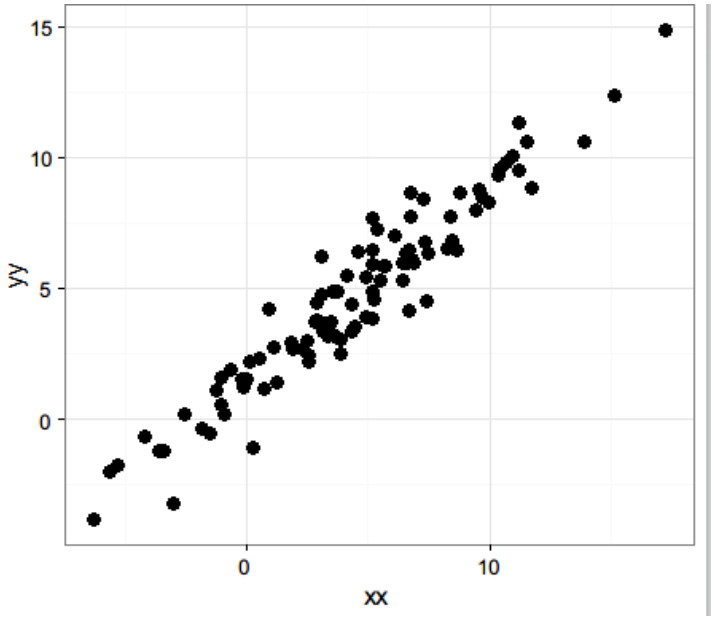

```
> print(afrotation)
             PC1       PC2
xx -0.8229231 0.5681528
yy  0.5681528 0.8229231
>
```

1番目の主軸　　2番目の主軸

主成分分析の結果

主成分分析の例

元データ




```
>
```

	PC1	PC2
xx	-0.7954104	0.6060712
yy	-0.6060712	-0.7954104

1番目の主軸

2番目の主軸

主成分分析の結果

7-2 Rシステムでの主成分分析の実行

パッケージの設定

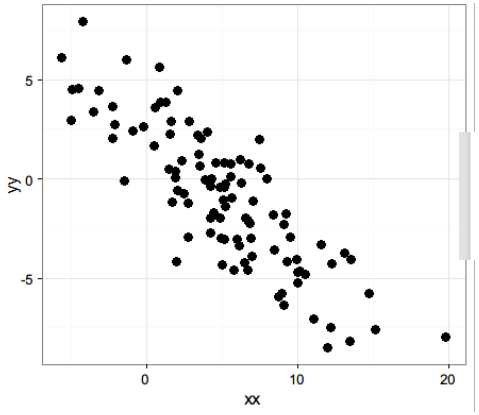
- 次の手順で，必要なパッケージをインストール
- パッケージをインストールするのにインターネット接続が必要
- install.packages("**ggplot2**") を実行
- install.packages("**pcaPP**") を実行

合成データからランダムに100個選び，主成分分析を実施

合成データ

サイズ１００
のデータを２セッ

タイプ：数値（整数化しない）
サイズ：100,000




```
> print(a$rotation)
          PC1       PC2
xx -0.8229231 0.5681528
yy  0.5681528 0.8229231
>
```

主成分分析

```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
n <- floor( runif(100, 1, 100000+1) )
d8 <- data.frame( xx=x[n],  yy=y[n] )
d8$yy <- d8$yy - (d8$xx + d8$yy) * 0.6
library(ggplot2)
ggplot(d8, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
a <- prcomp(d8)
print(a$rotation)
```

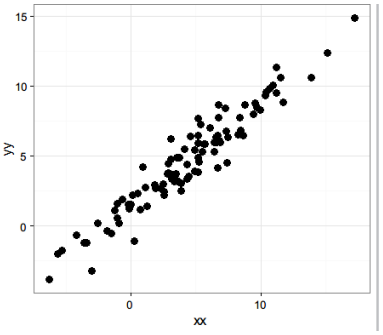
合成データの生成

合成データに
相関関係をもたせる

この２行が主成分分析

合成データからランダムに100個選びデータを作る

サイズ１００
のデータを２セ




```
          PC1        PC2
xx -0.7954104  0.6060712
yy -0.6060712 -0.7954104
>
```

主成分分析

タイプ：数値（整数化しない）
サイズ：100,000

```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
n <- floor( runif(100, 1, 100000+1) )
d9 <- data.frame( xx=x[n],  yy=y[n] )
d9$yy <- d9$yy + (d9$xx - d9$yy) * 0.8
library(ggplot2)
ggplot(d9, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
a <- prcomp(d9)
print(a$rotation)
```

合成データの生成

合成データに
相関関係をもたせる

この２行が主成分分析

7-3 ロバストな主成分分析

主成分分析

- **主成分分析**は，便利であるが，万能ではない
- ノイズを含むデータについて，**ノイズがランダム**であれば，**主軸**に**影響はない**
- **ノイズがランダムでない場合**，ノイズが主軸に影響を及ぼし，**次元削減に悪影響**

ノイズの他にも、**外れ値**（他の値と比べて、異常に離れた値）、**計測漏れ**（データが空，データが０）も、次元削減に悪影響

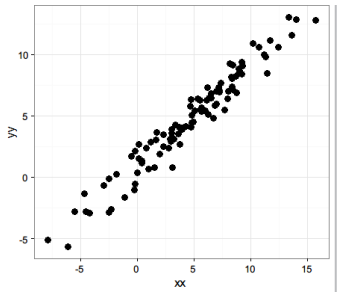
外れ値や**計測漏れ**などの**不適切なデータ**は、手作業や、適切な分析手法で**取り除く**必要あり

- **外れ値の検出**：データが**正規分布**に従っていると**仮定**し，「**平均から何倍の標準偏差以上離れているデータを外れ値**」とするなど

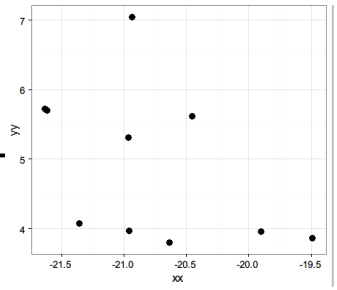
- **欠損データの取り扱い**：**無視**したり，**平均値や中央値で補完**する．欠損データにパターンがある場合には，欠損データの原因を調べる手が仮になる場合がある．

- 異常検出：統計，機械学習の手法がある

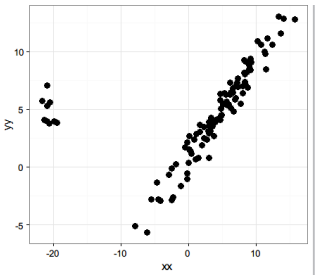
外れ値を含むデータの合成の例



d9 ＋



d10
外れ値

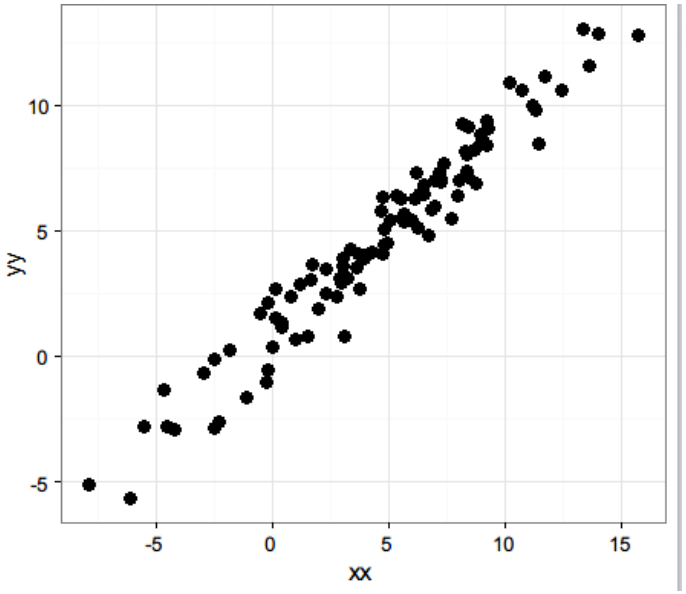


d11
外れ値が混入

```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
n <- floor( runif(100, 1, 100000+1) )
d9 <- data.frame( xx=x[n],  yy=y[n] )
d9$yy <- d9$yy + (d9$xx - d9$yy) * 0.8
d10 <- data.frame( xx=rnorm(10, mean=-20, sd=1),
  yy=rnorm(10, mean=5, sd = 1) )
d11 <- rbind( d9, d10 )
library(ggplot2)
ggplot(d11, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
```

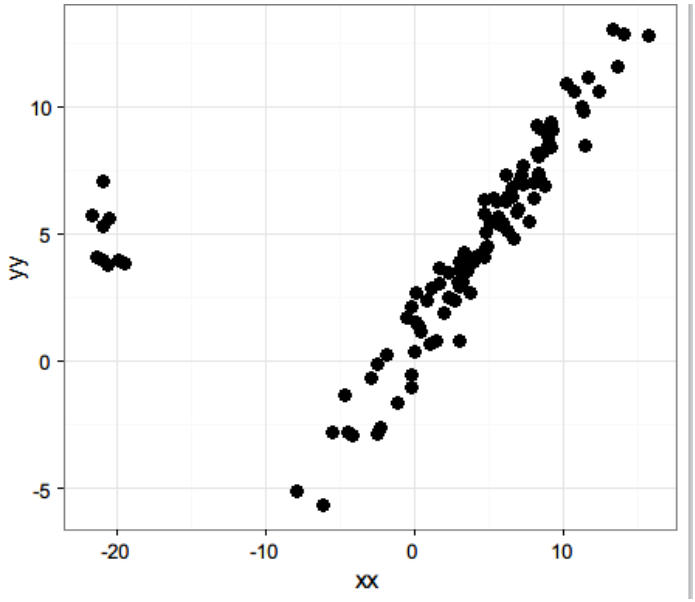
外れ値の混入

主成分分析は外れ値に弱い



d9

```
> a <- prcomp(d9)
> print(a$rotation)
           PC1        PC2
xx -0.7638487 -0.6453954
yy -0.6453954  0.7638487
>
```

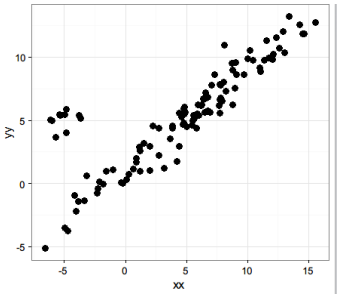


d11
外れ値が混入

```
> a <- prcomp(d11)
> print(a$rotation)
           PC1        PC2
xx -0.9681328 -0.2504373
yy -0.2504373  0.9681328
>
```

全データから
忠実に軸を
算出

主成分分析は外れ値に弱い



d11

```
> a <- prcomp(d11)
> print(a$rotation)
            PC1        PC2
xx -0.9681328 -0.2504373
yy -0.2504373  0.9681328
>
```

主成分分析

```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
n <- floor( runif(100, 1, 100000+1) )
d9 <- data.frame( xx=x[n],  yy=y[n] )
d9$yy <- d9$yy + (d9$xx - d9$yy) * 0.8
d10 <- data.frame( xx=rnorm(10, mean=-20, sd=1),
  yy=rnorm(10, mean=5, sd = 1) )
d11 <- rbind( d9, d10 )
library(ggplot2)
ggplot(d11, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
a <- prcomp(d11)
print(a$rotation)
```

外れ値の混入

ロバストな主成分分析

【基本は】

外れ値や**計測漏れ**などの**不適切なデータ**は、手作業や、適切な分析手法で**取り除く**必要あり

【ロバストな主成分分析】

ノイズがあったり，一部，外れ値や計測漏れが混入していたとしても，それらの悪影響が少ないように改良された主成分分析

主成分分析のバリエーション①
ロバスト主成分分析（robust PCA ）

- **主成分分析**は**データの外れ値の影響を受けやすい**

- 外れ値（他のデータから大きく離れたデータ）は、大きく離れているからこそ、データの分散値に大きな影響を与える可能性がある

- **ロバスト主成分分析**は、この問題を克服するための一手法

文献

Algorithms for Projection-Pursuit Robust Principal Component Analysis" by C. Croux, P. Filzmoser, and M. Oliveira, published in Chemometrics and Intelligent Laboratory Systems, Vol. 87, pp. 218-225, 2007

主成分分析のバリエーション②
Principal Component Pursuit

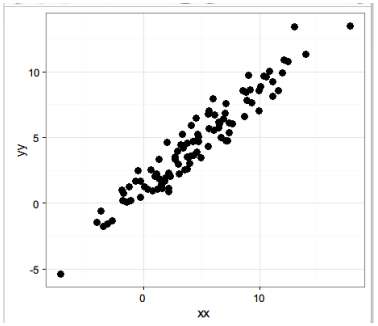
- ロバストな主成分分析の一手法

文献

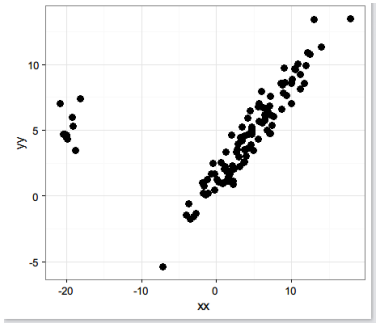
Robust principal component analysis?” by E. J. Candes, X. Li, Y. Ma, and J. Wright, published in Journal of the ACM (JACM), 58(3), 11, 2011

robust PCA の実行結果例

pcaPP パッケージを使用



d9



d11
外れ値が混入

- PCA

```
> a <- prcomp(d9)
> print(a$rotation)
           PC1        PC2
xx -0.7937832 -0.6082008
yy -0.6082008  0.7937832
> |
```

```
> a <- prcomp(d11)
> print(a$rotation)
           PC1        PC2
xx -0.9776248 -0.2103561
yy -0.2103561  0.9776248
> a <- prcomp(d9)
```

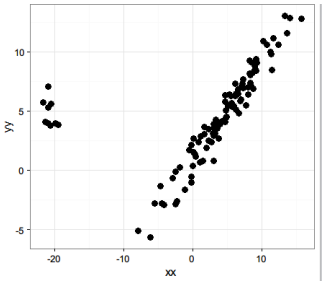
- Robust PCA

```
> a2 <- PCAgrid(d9)
> print(a2$loadings)

Loadings:
   Comp.1 Comp.2
xx  0.725 -0.688
yy  0.688  0.725
```

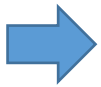
```
> a2 <- PCAgrid(d11)
> print(a2$loadings)

Loadings:
   Comp.1 Comp.2
xx  0.805 -0.594
yy  0.594  0.805
```

外れ値に対して
ある程度の
耐性がある

pcaPP パッケージを用いて robust PCA



d11



```
> library(pcaPP)
> a2 <- PCAgrid(d11)
> print(a2$loadings)

Loadings:
   Comp.1 Comp.2
xx  0.849 -0.529
yy  0.529  0.849
```

```
x <- rnorm(100000, mean=5, sd=5)
y <- rnorm(100000, mean=5, sd=5)
n <- floor( runif(100, 1, 100000+1) )
d9 <- data.frame( xx=x[n],  yy=y[n] )
d9$yy <- d9$yy + (d9$xx - d9$yy) * 0.8
d10 <- data.frame( xx=rnorm(10, mean=-20, sd=1),
  yy=rnorm(10, mean=5, sd = 1) )
d11 <- rbind( d9, d10 )
library(ggplot2)
ggplot(d11, aes(x=xx)) +
  geom_point( aes(y=yy), size=3 ) + theme_bw()
library(pcaPP)
a2 <- PCAgrid(d11)
print(a2$loadings)
```

外れ値の混入

全体まとめ

- **主成分分析と次元削減**：**主成分分析（PCA）**は**次元削減**の一手法．**多次元データを低次元データに変換**．
- **主成分分析の特徴**：**主軸**は，データの分散を最大限に保つように選ばれる．**元のデータセットの情報を可能な限り保持**するように**次元削減**を行う．
- **ロバストな主成分分析**：**主成分分析**では、**外れ値**が分散に大きな影響を与，得られる主軸が不適切になるという問題がある．**ロバストな主成分分析**は、**この問題を克服**することを目的としている．