主成分分析

大竹用

Greg Nishihara

2019 Sep 24

Contents

List of Tables

List of Figures

1 パッケージの読み込み

```
library(tidyverse)
library(readxl)
library(FactoMineR) # このパッケージで PCA, RDA 解析ができます。
library(factoextra)
```

2　データの読み込み

```
d3 = read_xlsx("主成分分析_190924.xlsx", sheet = "Raw data_full year")
d3 = d3 %>% drop_na() # 欠損値を外す
```

PCA はここであてはめる。

```
d3out = PCA(d3, graph = FALSE)
```

PCA 細かい情報はここで抽出します。まず，主成分に対する各要因の貢献度 (Contribution)。

```
tmp1 = get_pca_var(d3out) %>% pluck("contrib") %>%
  as_tibble(rownames = "Variable") %>%
  gather("Dim", "Contribution", -Variable)
```

つぎに，主成分に対する各要因の相関係数。

```
tmp2 = get_pca_var(d3out) %>%  pluck("cor") %>%
  as_tibble(rownames = "Variable") %>%
  gather("Dim", "Correlation", -Variable)
```

この２つの情報を結合して，貢献度の評価をします。N 要因の貢献度が等しいとき，貢献度は $100/N$ です。つまり，貢献度が $100/N$ 以上のとき，その要因が主成分に強く影響しています。

```
d3_x = full_join(tmp1, tmp2) # 結合
N = length(unique(tmp1$Variable)) # 要因の数
```

貢献度の評価はここで行います。

```
d3_x = d3_x %>%
  mutate(C = ifelse(Contribution > 100 / N,
                    "Above average contribution",
                    "Below average contribution"))
```

さらに，主成分分析の図をつくるためのデータをもとめます。

```
pcacoord = get_pca_var(d3out) %>%
  pluck("coord") %>%
  as_tibble(rownames = "Variable")
pcacoord = pcacoord %>%
  mutate(d1 = 1.1*sqrt(Dim.1^2 + Dim.2^2) * sin(atan2(Dim.1, Dim.2)),
         d2 = 1.1*sqrt(Dim.1^2 + Dim.2^2) * cos(atan2(Dim.1, Dim.2)),
         d3 = 1.1*sqrt(Dim.3^2 + Dim.4^2) * sin(atan2(Dim.3, Dim.4)),
         d4 = 1.1*sqrt(Dim.3^2 + Dim.4^2) * cos(atan2(Dim.3, Dim.4)),
         a1 = atan2(Dim.2, Dim.1),
         a2 = atan2(Dim.4, Dim.3)) %>%
  mutate(adj1 = ifelse(abs(a1) > pi/2, a1 - pi, a1),
         adj2 = ifelse(abs(a2) > pi/2, a2 - pi, a2))
```

固有値があれば，主成分がどの程度全体の分散を説明しているかがわかります。

```
eval = d3out %>% get_eig()
```

```
cap = "主成分１に対すして，相関の高い順でならべた要因です。
黒丸は貢献度が 100/14 を超えた要因です。
主成分１に貢献する要因は８つあります。"
```

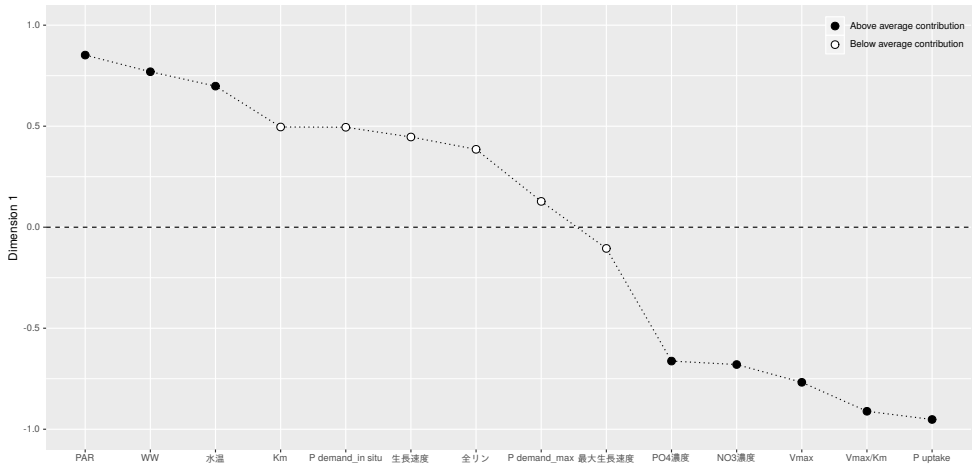

Fig. 1　主成分１に対すして，相関の高い順でならべた要因です。黒丸は貢献度が 100/14 を超えた要因です。主成分１に貢献する要因は８つあります。

```
d3_x %>% filter(str_detect(Dim, "Dim.1")) %>%
  ggplot(aes(x = reorder(Variable, -Correlation), y = Correlation)) +
  geom_line(aes(group = 1), linetype = "dotted") +
  geom_point(aes(shape = C, fill = C), size = 3) +
  geom_hline(yintercept= 0, linetype = "dashed") +
  scale_x_discrete("") +
  scale_y_continuous("Dimension 1", limits = c(-1,1)) +
  scale_shape_manual(values = c(19, 21)) +
  scale_fill_manual(values = c("black", "white")) +
  theme(legend.background=element_blank(),
        legend.position=c(1,1),
        legend.justification=c(1,1),
        legend.title=element_blank())
```

```
cap = "主成分２に対すして，相関の高い順でならべた要因です。
黒丸は貢献度が 100/14 を超えた要因です。
主成分２に貢献する要因は５つあります。"
d3_x %>% filter(str_detect(Dim, "Dim.2")) %>%
  ggplot(aes(x = reorder(Variable, -Correlation), y = Correlation)) +
  geom_line(aes(group = 1), linetype = "dotted") +
  geom_point(aes(shape = C, fill = C), size = 3) +
  geom_hline(yintercept= 0, linetype = "dashed") +
    scale_x_discrete("") +
  scale_y_continuous("Dimension 2", limits = c(-1,1)) +
  scale_shape_manual(values = c(19, 21)) +
  scale_fill_manual(values = c("black", "white")) +
  theme(legend.background=element_blank(),
        legend.position=c(1,1),
        legend.justification=c(1,1),
        legend.title=element_blank())
```

```
cap = "主成分３に対すして，相関の高い順でならべた要因です。
黒丸は貢献度が 100/14 を超えた要因です。
```

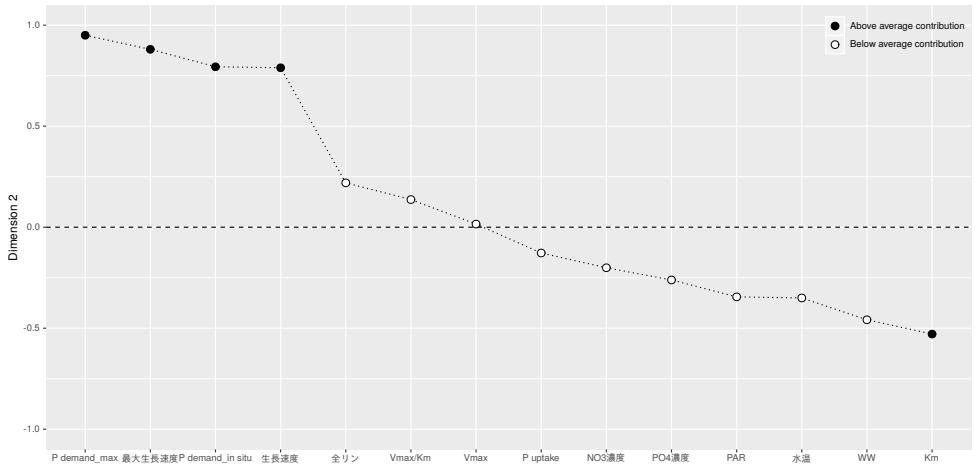

Fig. 2　主成分２に対すして，相関の高い順でならべた要因です。黒丸は貢献度が 100/14 を超えた要因です。主成分２に貢献する要因は５つあります。

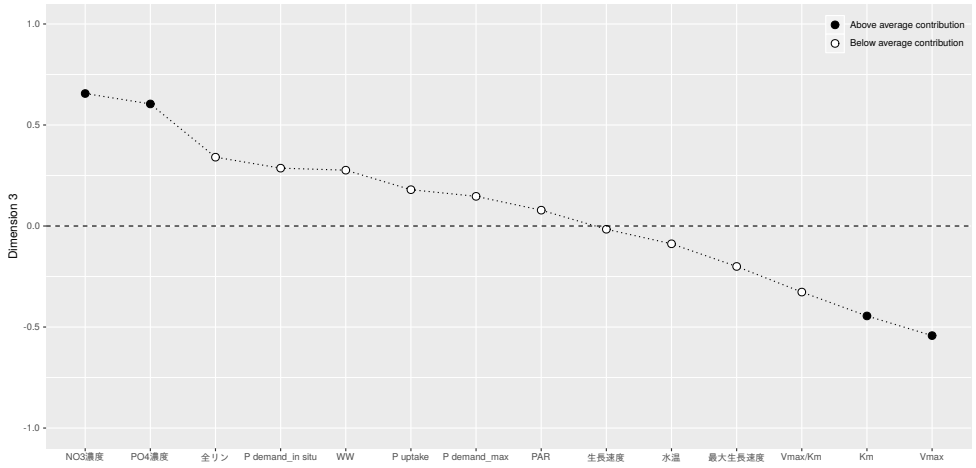

Fig. 3　主成分３に対すして，相関の高い順でならべた要因です。黒丸は貢献度が 100/14 を超えた要因です。主成分３に貢献する要因は３つあります。

```
主成分３に貢献する要因は３つあります。"
d3_x %>% filter(str_detect(Dim, "Dim.3")) %>%
  ggplot(aes(x = reorder(Variable, -Correlation), y = Correlation)) +
  geom_line(aes(group = 1), linetype = "dotted") +
  geom_point(aes(shape = C, fill = C), size = 3) +
  geom_hline(yintercept= 0, linetype = "dashed") +
  scale_x_discrete("") +
  scale_y_continuous("Dimension 3", limits = c(-1,1)) +
  scale_shape_manual(values = c(19, 21)) +
  scale_fill_manual(values = c("black", "white")) +
  theme(legend.background=element_blank(),
        legend.position=c(1,1),
        legend.justification=c(1,1),
        legend.title=element_blank())
```

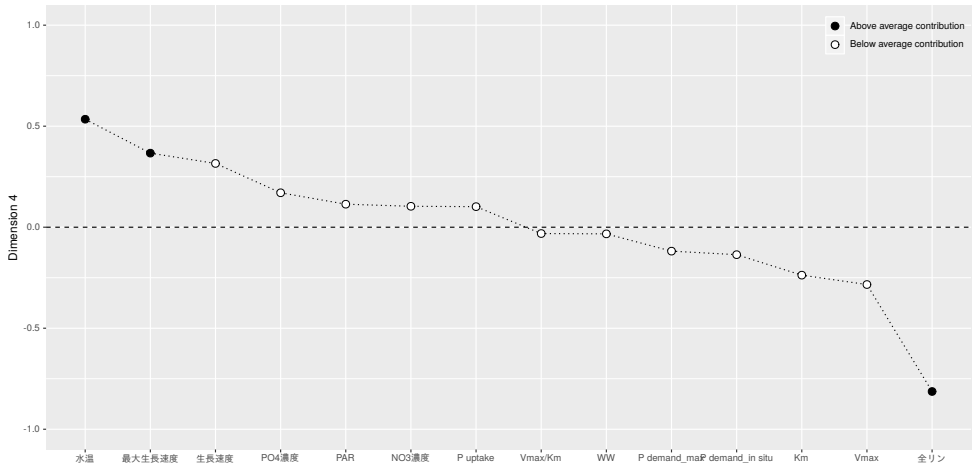

Fig. 4　主成分４に対すして，相関の高い順でならべた要因です。黒丸は貢献度が 100/14 を超えた要因です。主成分４に貢献する要因は３つあります。

```
cap = "主成分４に対すして，相関の高い順でならべた要因です。
黒丸は貢献度が 100/14 を超えた要因です。
主成分４に貢献する要因は３つあります。"
d3_x %>% filter(str_detect(Dim, "Dim.4")) %>%
  ggplot(aes(x = reorder(Variable, -Correlation), y = Correlation)) +
  geom_line(aes(group = 1), linetype = "dotted") +
  geom_point(aes(shape = C, fill = C), size = 3) +
  geom_hline(yintercept= 0, linetype = "dashed") +
  scale_x_discrete("") +
  scale_y_continuous("Dimension 4", limits = c(-1,1)) +
  scale_shape_manual(values = c(19, 21)) +
  scale_fill_manual(values = c("black", "white")) +
  theme(legend.background=element_blank(),
        legend.position=c(1,1),
        legend.justification=c(1,1),
        legend.title=element_blank())
```

3　主成分分析の図

主成分分析の図に，correlation circle を乗せるための関数です。

```
circleFun <- function(center = c(0,0),diameter = 1, npoints = 100){
  r = diameter / 2
  tt <- seq(0,2*pi,length.out = npoints)
  xx <- center[1] + r * cos(tt)
  yy <- center[2] + r * sin(tt)
  return(tibble(x = xx, y = yy))
}
```

```
cap = "The first principal component plane, ここで主成分１と２を座標軸にしています。
矢印が円に近いほど，その要因がこの面に強く貢献している。つまり，矢印の長さが１のとき，
円と重なり，その要因は他の主成分に貢献していない。
矢印と矢印の間の角度の cos() をとれば，そのペアの相関係数を求められます。"
```

```
xlab = str_glue("Dim. 1 ({round(eval[1,2], 1)}%)")
ylab = str_glue("Dim. 2 ({round(eval[2,2], 1)}%)")
ggplot(pcacoord) +
  geom_vline(xintercept=0, linetype = "dashed") +
  geom_hline(yintercept=0, linetype = "dashed") +
  geom_segment(aes(x = 0, y = 0,
                   xend = Dim.1, yend = Dim.2),
               arrow=arrow(angle = 20, length = unit(5, "mm"), type="closed")) +
  geom_text(aes(x = d1, y = d2, label = Variable,
                angle = 360*adj1/(2*pi)),
            hjust = 1,
            data = pcacoord %>% filter(abs(pcacoord$a1) > pi/2)) +
  geom_text(aes(x = d1, y = d2, label = Variable,
                angle = 360*adj1/(2*pi)),
            hjust = 0,
            data = pcacoord %>% filter(abs(pcacoord$a1) < pi/2)) +
  geom_path(aes(x = x, y = y),
            data = circleFun(diameter = 2)) +
  scale_x_continuous(xlab, limits = c(-1.2, 1.2)) +
  scale_y_continuous(ylab, limits = c(-1.2, 1.2)) +
  coord_equal()
```

```
cap = "The third principal component plane, ここで主成分３と４を座標軸にしています。
この面の場合，前リンだけ強く貢献しています。"

xlab2 = str_glue("Dim. 3 ({round(eval[3,2], 1)}%)")
ylab2 = str_glue("Dim. 4 ({round(eval[4,2], 1)}%)")
ggplot(pcacoord) +
  geom_vline(xintercept=0, linetype = "dashed") +
  geom_hline(yintercept=0, linetype = "dashed") +
  geom_segment(aes(x = 0, y = 0,
                   xend = Dim.3, yend = Dim.4),
               arrow=arrow(angle = 20, length = unit(5, "mm"), type="closed")) +
  geom_text(aes(x = d3, y = d4, label = Variable,
                angle = 360*adj2/(2*pi)),
            hjust = 1,
            data = pcacoord %>% filter(abs(pcacoord$a2) > pi/2)) +
  geom_text(aes(x = d3, y = d4, label = Variable,
                angle = 360*adj2/(2*pi)),
            hjust = 0,
            data = pcacoord %>% filter(abs(pcacoord$a2) < pi/2)) +
  geom_path(aes(x = x, y = y),
            data = circleFun(diameter = 2)) +
  scale_x_continuous(xlab2, limits = c(-1.2, 1.2)) +
  scale_y_continuous(ylab2, limits = c(-1.2, 1.2)) +
  coord_equal()
```

```
cap = "主成分に対する各要因の修正分負荷量 (loading)。最初の５つの主成分だけ示しています。"
get_pca(d3out) %>% pluck("coord") %>% as_tibble(rownames = "Variable") %>%
  arrange(desc(Dim.1)) %>%
  knitr::kable(format = "latex",
               booktabs = TRUE, digits=4,
```

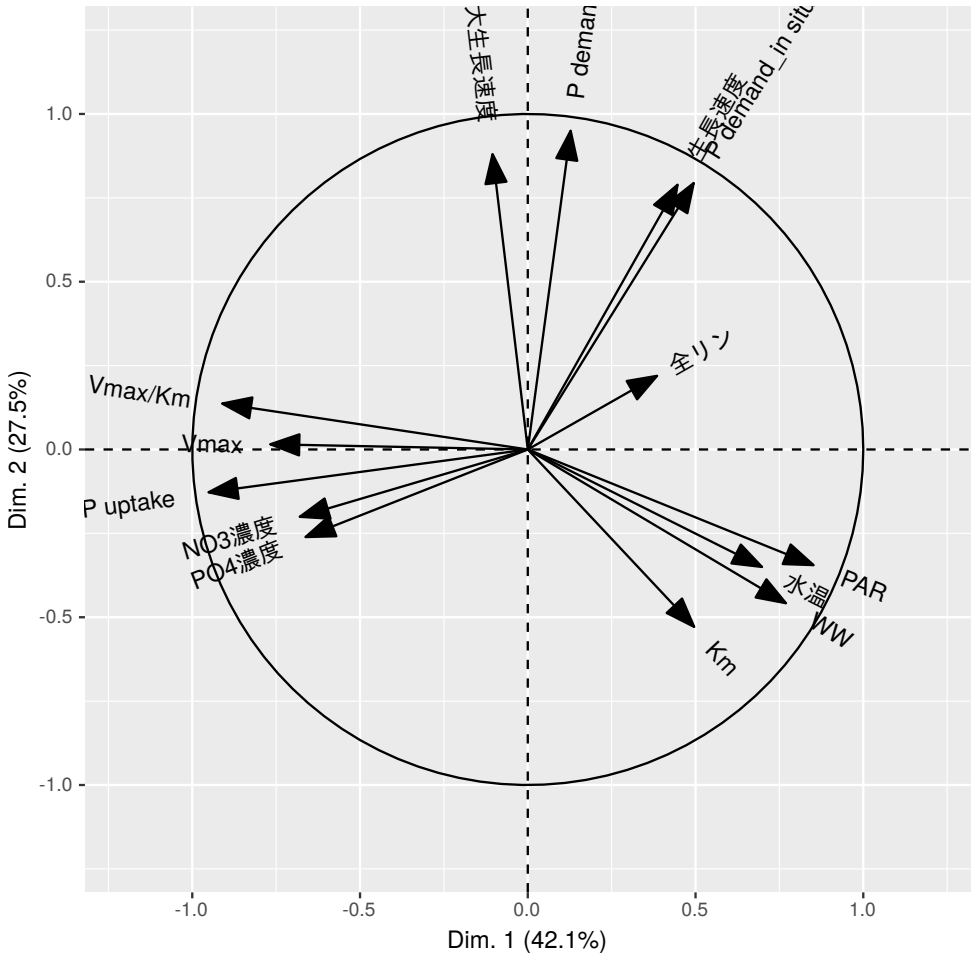

Fig. 5　The first principal component plane, ここで主成分１と２を座標軸にしています。矢印が円に近いほど，その要因がこの面に強く貢献している。つまり，矢印の長さが１のとき，円と重なり，その要因は他の主成分に貢献していない。矢印と矢印の間の角度の cos() をとれば，そのペアの相関係数を求められます。

```
                linesep = "", caption=cap)
```

```
cap = "主成分に対する各要因の相関係数（correlation)。最初の５つの主成分だけ示しています。"
get_pca(d3out) %>% pluck("cor") %>% as_tibble(rownames = "Variable") %>%
  arrange(desc(Dim.1)) %>%
  knitr::kable(format = "latex",
               booktabs = TRUE, digits=4,
               linesep = "", caption=cap)
```

```
cap = "主成分に対する各要因の貢献度（contribution)。最初の５つの主成分だけ示しています。列の和は１００です。要
get_pca(d3out) %>% pluck("contrib") %>% as_tibble(rownames = "Variable") %>%
  arrange(desc(Dim.1)) %>%
  knitr::kable(format = "latex",
               booktabs = TRUE, digits=4,
               linesep = "", caption=cap)
```

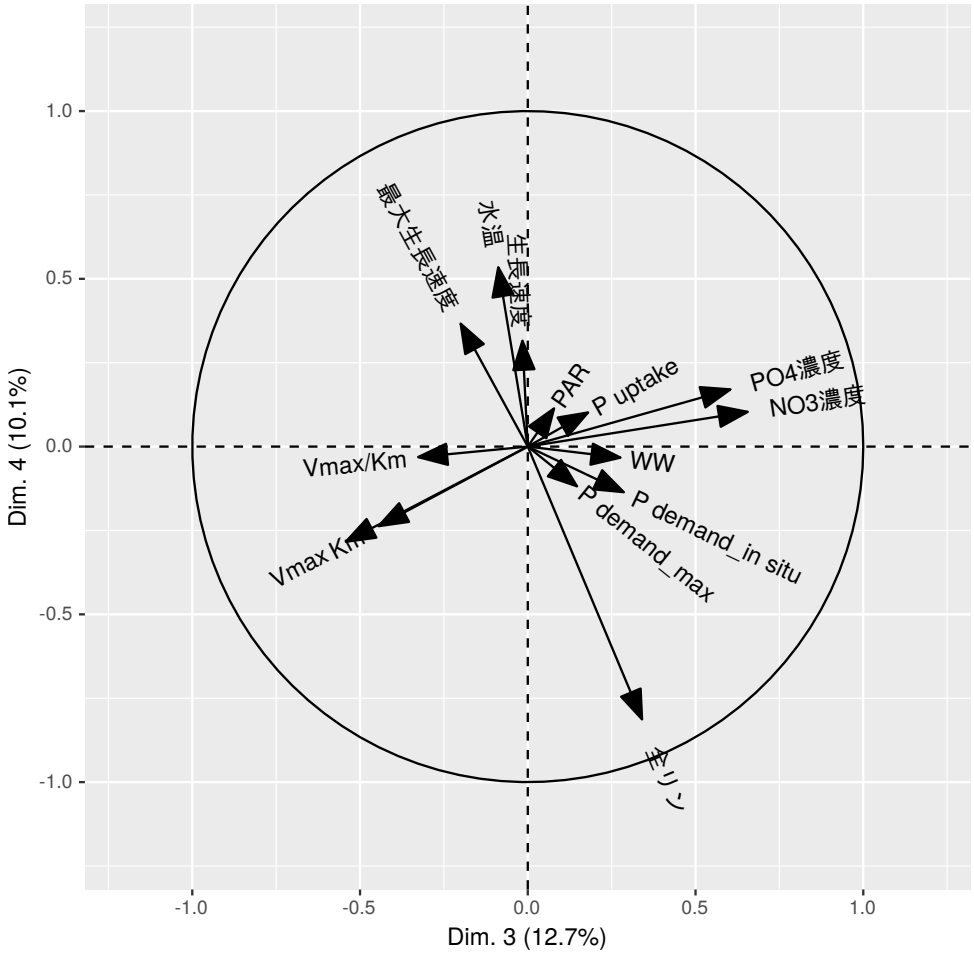

Fig. 6　The third principal component plane, ここで主成分３と４を座標軸にしています。この面の場合，前リンだけ強く貢献しています。

Table 1　主成分に対する各要因の修正分負荷量 (loading)。最初の５つの主成分だけ示しています。

Variable	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PAR	0.8519	-0.3447	0.0784	0.1139	0.2154
WW	0.7696	-0.4582	0.2764	-0.0329	-0.1973
水温	0.6983	-0.3502	-0.0880	0.5345	-0.2500
Km	0.4961	-0.5288	-0.4451	-0.2373	0.4080
P demand_in situ	0.4947	0.7942	0.2864	-0.1359	0.0993
生長速度	0.4465	0.7891	-0.0160	0.3156	0.1998
全リン	0.3855	0.2195	0.3407	-0.8131	-0.1215
P demand_max	0.1278	0.9501	0.1469	-0.1185	-0.0264
最大生長速度	-0.1050	0.8804	-0.2002	0.3666	0.0181
PO4 濃度	-0.6626	-0.2610	0.6042	0.1704	0.1789
NO3 濃度	-0.6797	-0.2007	0.6558	0.1035	0.1435
Vmax	-0.7677	0.0155	-0.5425	-0.2837	0.0254
Vmax/Km	-0.9111	0.1368	-0.3269	-0.0313	-0.0990
P uptake	-0.9517	-0.1276	0.1798	0.1018	0.0003

Table 2　主成分に対する各要因の相関係数 (correlation)。最初の５つの主成分だけ示しています。

Variable	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PAR	0.8519	-0.3447	0.0784	0.1139	0.2154
WW	0.7696	-0.4582	0.2764	-0.0329	-0.1973
水温	0.6983	-0.3502	-0.0880	0.5345	-0.2500
Km	0.4961	-0.5288	-0.4451	-0.2373	0.4080
P demand_in situ	0.4947	0.7942	0.2864	-0.1359	0.0993
生長速度	0.4465	0.7891	-0.0160	0.3156	0.1998
全リン	0.3855	0.2195	0.3407	-0.8131	-0.1215
P demand_max	0.1278	0.9501	0.1469	-0.1185	-0.0264
最大生長速度	-0.1050	0.8804	-0.2002	0.3666	0.0181
PO4 濃度	-0.6626	-0.2610	0.6042	0.1704	0.1789
NO3 濃度	-0.6797	-0.2007	0.6558	0.1035	0.1435
Vmax	-0.7677	0.0155	-0.5425	-0.2837	0.0254
Vmax/Km	-0.9111	0.1368	-0.3269	-0.0313	-0.0990
P uptake	-0.9517	-0.1276	0.1798	0.1018	0.0003

Table 3　主成分に対する各要因の貢献度 (contribution)。最初の５つの主成分だけ示しています。列の和は１００です。要因が同等に貢献している場合，貢献度は 100 / 14 = 7.14 です。

Variable	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
P uptake	15.3561	0.4225	1.8196	0.7316	0.0000
Vmax/Km	14.0729	0.4856	6.0152	0.0694	2.2146
PAR	12.3061	3.0828	0.3461	0.9171	10.4724
WW	10.0412	5.4476	4.2985	0.0766	8.7894
Vmax	9.9927	0.0062	16.5607	5.6877	0.1456
水温	8.2670	3.1814	0.4356	20.1834	14.1119
NO3 濃度	7.8336	1.0450	24.2024	0.7576	4.6522
PO4 濃度	7.4445	1.7680	20.5448	2.0527	7.2246
Km	4.1722	7.2559	11.1473	3.9779	37.5873
P demand_in situ	4.1493	16.3645	4.6157	1.3054	2.2248
生長速度	3.3805	16.1575	0.0144	7.0379	9.0145
全リン	2.5201	1.2504	6.5317	46.7158	3.3318
P demand_max	0.2768	23.4223	1.2136	0.9914	0.1568
最大生長速度	0.1869	20.1103	2.2544	9.4955	0.0741