Лабораторна робота №2
з дисципліни "Чисельні методи у статистиці"

Горбунов, 5 курс, "Прикладна та теоретична статистика"

12 травня 2022 р.

Вступ.

У даній роботі проведено аналіз даних про вибірки трилобітів, перевірено, чи є значущими зміни між вибірками. Буде показано, що залежність справді наявна, і побудовано алгоритми визначення групової належності. Підхід, що продемонстровано у цій роботі, дещо нагадує ті кроки, що були використані у роботі [1].

Хід роботи.

Постановка задачі.

Треба конвертувати дані про три вибірки трилобітів (файли glaber.dat, linnars1.dat та linnars2.dat) у формат R і проаналізувати методами дискримінантного аналізу, чи є між ними відмінності за чотирма дослідженими ознаками, методами дискримінантного аналізу. Якщо відмінності є, треба побудувати алгоритми визначення групової належності.

Огляд початкових даних.

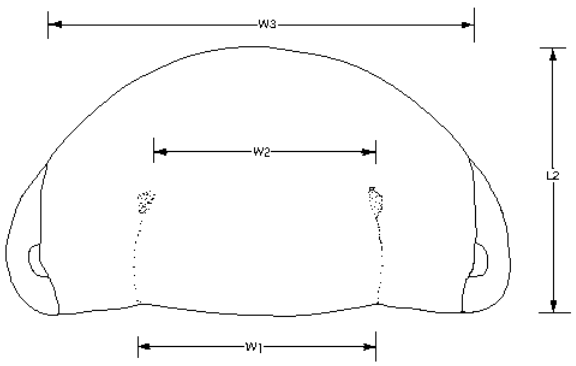

Рис. 1: Графічне пояснення спостережуваних характеристик у кожній з вибірок.

Маємо три вибірки з текстових файлів glaber.dat, linnars1.dat та linnars2.dat відповідно. Кількість елементів у кожній з вибірок становить $n_1 = 43$, $n_2 = 10$, $n_3 = 7$. Кожному спостереженню відповідають по чотири виміри, зроблених на кранідіях трилобітів, що зображені на рисунку вище. Характеристики L2, W1, W2 та W3 були виміряні на Stenopareia glaber з Норвегії (glaber.dat) разом з S. linnarssoni з Норвегії (linnars1.dat) та Швеції (linnars2.dat).

Виявлення відмінностей між видами трилобітів.

Здається, використання тестів типу MANOVA є недоцільним на такій вибірці незначного обсягу (чому – далі пояснимо). Тому здебільшого висновки будуть робитися або на око з діаграм розсіювання між парами характеристик.

Спочатку почнемо з використання MANOVA "в лоб". Його використання здається не зовсім правильним принаймні з точки зору обмеженості у вибірці: її обсяг зовсім малий, що не дозволить вірити у результати тестів на нормальність за цією вибіркою. Хоча можна було б побудувати QQ-діаграми та порівняти емпіричні з теоретичними квантилями нормального розподілу:

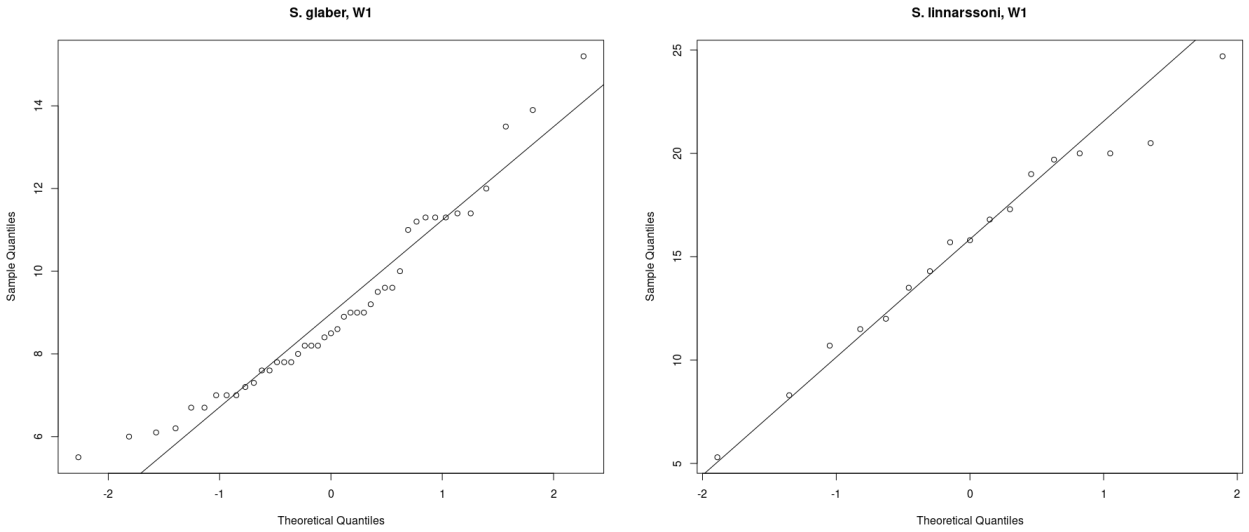

Рис. 2: QQ-діаграми для W1 в залежності від виду трилобітів.

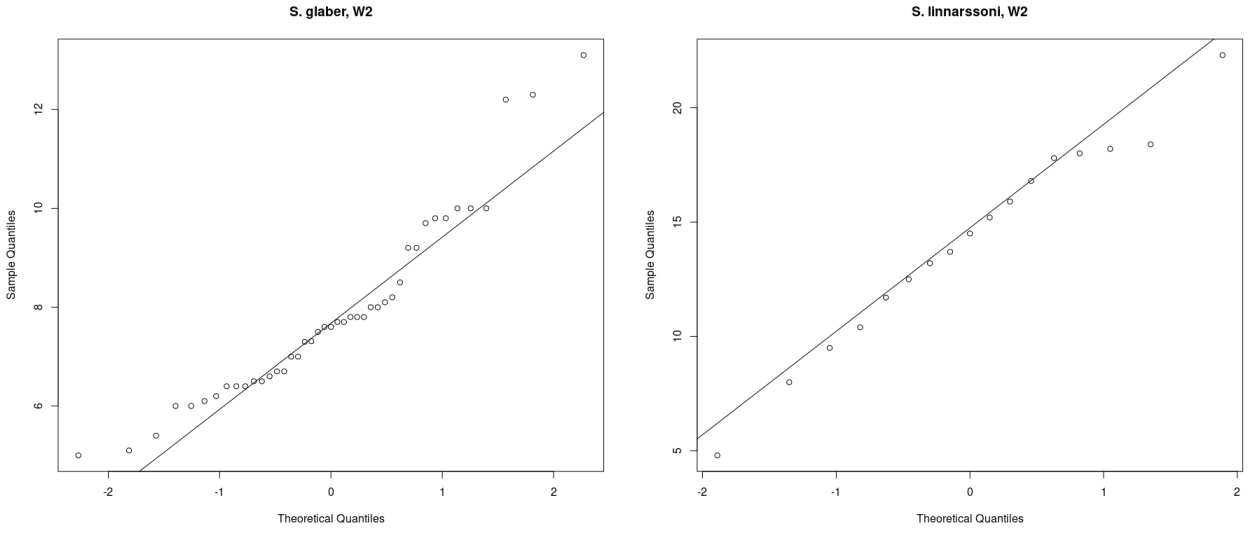

Рис. 3: QQ-діаграми для W2 в залежності від виду трилобітів.

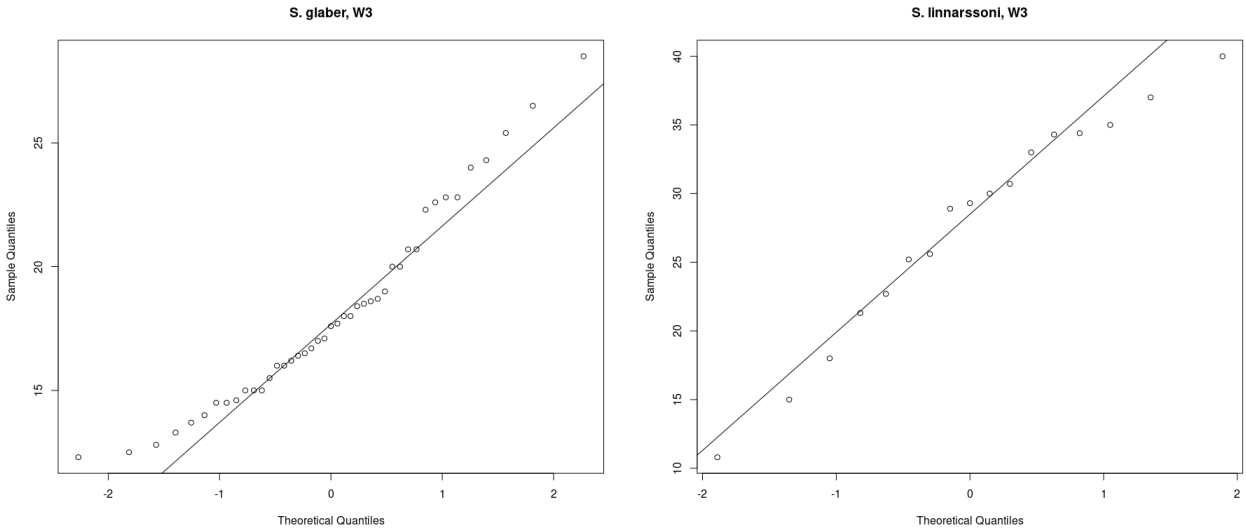

Рис. 4: QQ-діаграми для W3 в залежності від виду трилобітів.

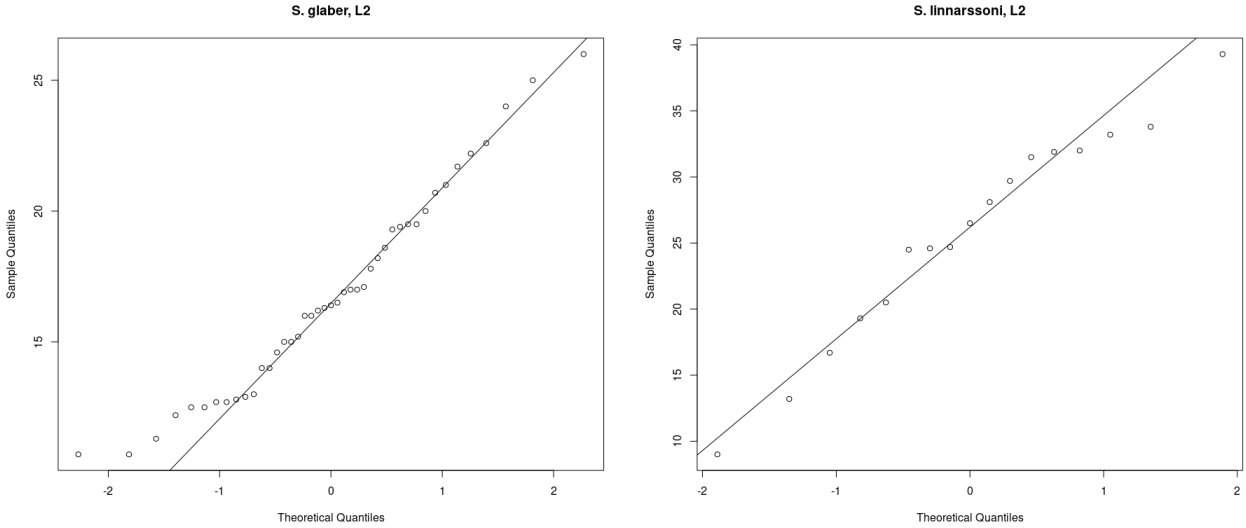

Рис. 5: QQ-діаграми для L2 в залежності від виду трилобітів.

Для деяких змінних та видів відхилення, звісно, помітні від гауссової моделі, однак можемо на хвилину "припустити", що умова нормальності виконана. Якщо б обсяги на кожній з вибірок (тобто мова іде про n_j) були б достатньо великими, то на певні відхилення від нормальності можна було б закрити очі (однак, наприклад, 17 не можна вважати "достатньо великим"). Реалізація MANOVA в R може бути наступною:

```
# Чи є залежність від виду трилобітів? Наївне застосування MANOVA
tri.manova <- manova(cbind(tri$W1, tri$W2, tri$W3, tri$L2) ~ tri$Code)
```

Досягнутий рівень значущості виходить занадто малим, що відіграє на користь гіпотези (формулюється як альтернатива) про значущу відмінність розподілу хакрактеристик в залежності від виду трилобітів:

```
          Df  Pillai approx F num Df den Df    Pr(>F)
tri$Code   1 0.67083   28.022      4     55 1.042e-12 ***
Residuals 58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Цікаво було б перевірити тест на те, чи є значущою відмінність між трилобітами одного виду, але різного походження (локації):

```
tri.s <- tri[tri$Code == "L",]
tri.manova.local <- manova(cbind(tri.s$W1,tri.s$W2,tri.s$W3,tri.s$L2) ~ tri.s$K)
```

От в даному разі виходить, що вибір локації не відіграє особливу роль у впливі на розподіл досліджуваних характеристик:

```
          Df  Pillai approx F num Df den Df  Pr(>F)
tri.s$K    1 0.48512   2.8266      4     12 0.07286 .
Residuals 15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Якісь числа ми отримали за MANOVA, однак на початку мова йшла про те, що висновки ще будуть зроблено з графічних ”тестів”. Зобразимо діаграми з вусами для спостережуваних характеристик в залежності від виду трилобіта:

```
boxplot(W1 ~ Code, data = tri)
# ... і аналогічно для інших трьох змінних
```

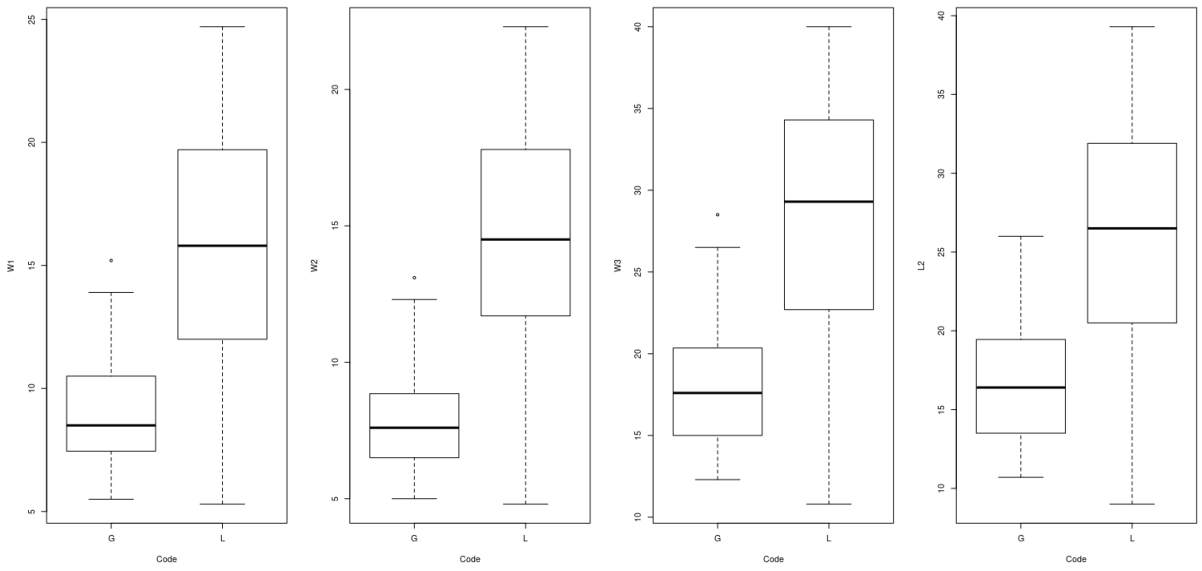

Рис. 6: Скриньки з вусами для кожного виміру, в залежності від спостережуваного виду.

З діаграми скриньок з вусами легко бачити відмінність розподілу довжини та ширини частин "тіла" трилобіта в залежності від виду. Можна припустити, що за формою трилобіти S. linnarssoni виходять і ширшими, і довшими у порівнянні з трилобітами Stenopareia glaber. Досить схожі висновки можна було б зробити з діаграми розсіювання пар для вимірів: Також

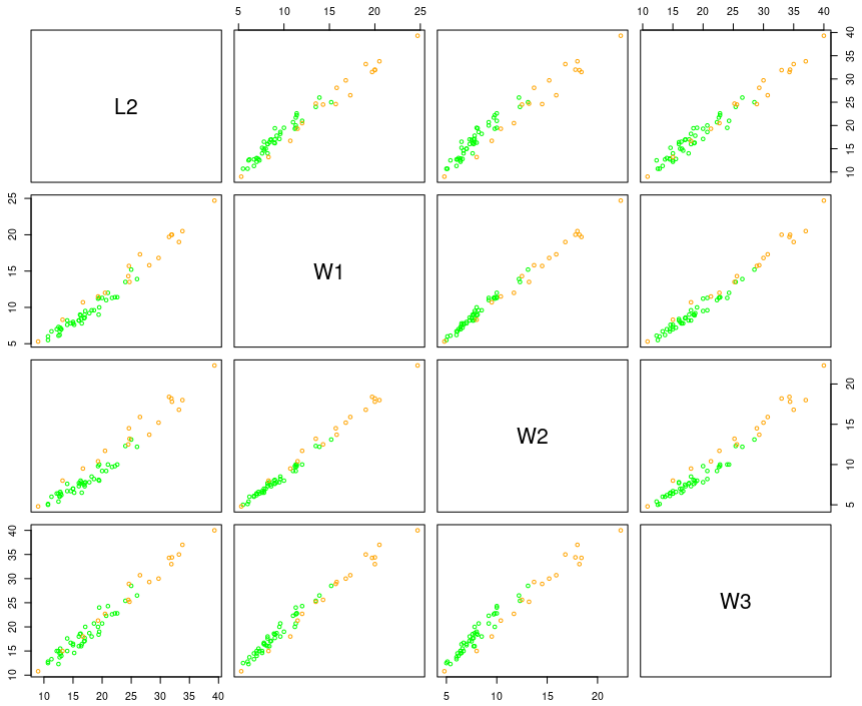

Рис. 7: Діагарми розсіювання пар змінних. Зелені точки відповідають вимірам за S. glaber, а помаранчеві – S. linnarssoni.

можна звернути увагу на те, що відстежується лінійна залежність між шириною та довжиною частин трилобіта. Однак слід зауважити, що ця лінійна залежність може відрізнятися в залежності від досліджуваного виду. Щоб підтвердити висунуте припущення, можна скористатися тестом Чоу. Наприклад, продемонструємо його використання на моделі вигляду:

$$\mathrm{L2} \approx (b_0^g + b_1^g \cdot \mathrm{W1}) \cdot \mathbb{1}\{\mathrm{Type} = \mathrm{Glaber}\} + (b_0^l + b_1^l \cdot \mathrm{W1}) \cdot \mathbb{1}\{\mathrm{Type} = \mathrm{Linnarssoni}\}$$

```
N <- nrow(tri); alpha <- 0.05
lm.h0 <- lm(L2  ~ W1, data = tri)
resid.h0 <- residuals(lm.h0)
lm.h1.subset1 <- lm(L2  ~ W1, data = subset(tri, Code == "G"))
lm.h1.subset2 <- lm(L2  ~ W1, data = subset(tri, Code == "L"))
resid.h1 <- c(residuals(lm.h1.subset1), residuals(lm.h1.subset2))
Sh0 <- sum(resid.h0^2); Sh1 <- sum(resid.h1^2)
# Статистика тесту
F.emp <- ((1 / 2) * (Sh0 - Sh1)) / ((1 / (N - 4)) * Sh1)
# Поріг тесту
F.theor <- qf(1 - alpha, 2, N - 4)
```

$F_{emp} \approx 7.946289 > 3.161861 \approx F_{theor}$, тому є підстави прийняти гіпотезу про наявність розшарування за видом трилобітів.

Метод головних компонент на трилобітах.

Спробуємо застосувати метод головних компонент для того, щоб відстежити, чи можна обрати такі головні напрямки так, щоб на проекції початкових даних на них мали краще (лінійне) розділення? Оскільки одиниці вимірювання для спостережуваних характеристик двожини та ширини однакові, то можна використати коваріаційну матрицю даних для цього підходу:

```
# Застосування методу головних компонент з використанням коваріаційної матриці
tri.pca <- princomp(tri[,1:4], cor = F)
# Виводимо базові результати
summary(tri.pca)
```

Виклик summary(tri.pca) повертає таблицю, яку грубо можна тракутвати як "впливовість" компонент:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	11.1509988	0.957257750	0.756804423	0.2355550889
Proportion of Variance	0.9877306	0.007278951	0.004549653	0.0004407537
Cumulative Proportion	0.9877306	0.995009593	0.999559246	1.0000000000

Табл. 1: Сингулярні числа, частки збереженої дисперсії та їхні накопичення відповідно.

З таблиці вище зрозуміло, що на першу компоненту припадає досить значна частка ($\approx 98.78\%$) інформації з вихідних даних. Зобразимо діаграму власних чисел, де для осі ординат візьмемо як звичайне, так і логарифмічне шкалування:

```
plot(tri.pca, main = "Діаграма власних чисел")
plot(tri.pca, main = "Діаграма власних чисел", log = "y")
```

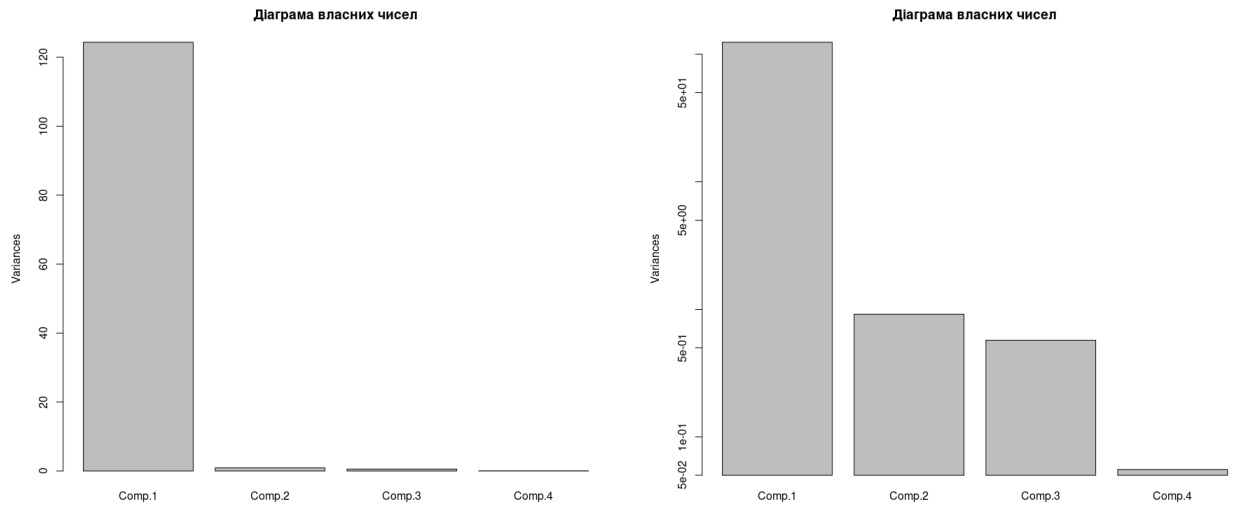

Рис. 8: Діаграма власних чисел. Зліва – звичайне шкалування, справа - логарифмічне.

З діаграми видно, що після першої компоненти відбувається злам, і з першого погляду здається, що наступні компоненти майже не відрізняються. Ця думка змінюється при переході до іншої (логарифмічної) шкали для осі ординат (або якщо пильно дивитися у вищенаведену таблицю), де чітко видно, що на наступні дві компоненти у певному сенсі не відрізняються та є впливовішими за останню.

Виведемо отримані навантаження за кожною компонентою:

```
# Виводимо матрицю навантажень
tri.pca$loadings
```

	Comp.1	Comp.2	Comp.3	Comp.4
L2	0.5925455	0.7748457	0.2074046	0.07407634
W1	0.3883039	-0.3133595	0.3452192	-0.79488971
W2	0.3507804	-0.4717825	0.5477796	0.59524100
W3	0.6124239	-0.2807869	-0.7333106	0.09138487

Табл. 2: Матриця навантажень на початкові змінні.

Можна було б робити певні висновки з цієї таблиці, однак на око краще сприймаються відповідні візуалізації. Тому для кожної компоненти зобразимо діаграму навантажень:

```
barplot(t(tri.pca$loadings[,1]), main = "Навантаження першої компоненти")
# І так для інших трьох
```

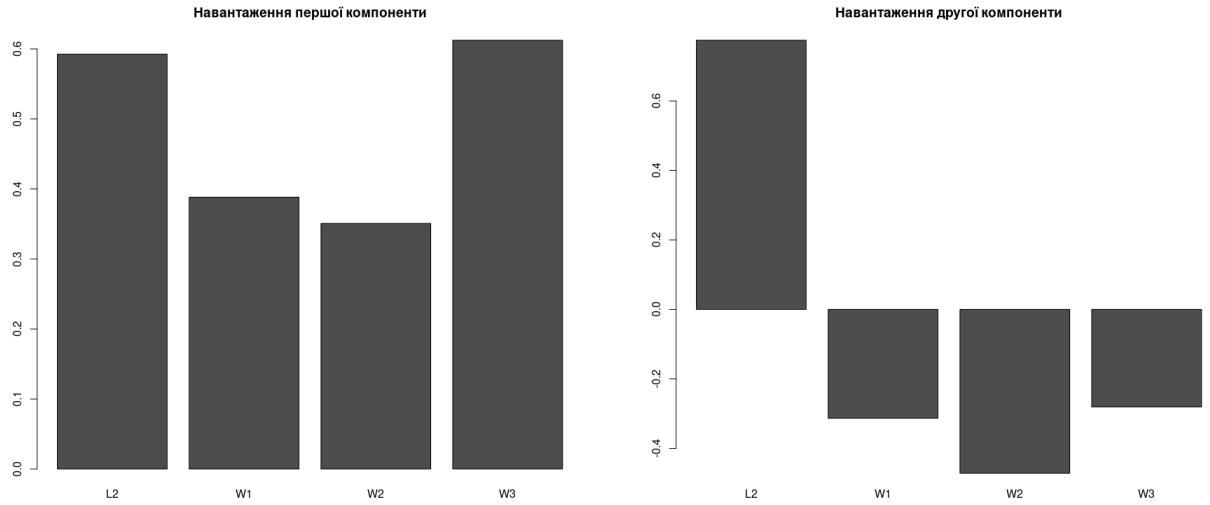

Рис. 9: Діаграми навантажень для перших двох компонент.

Спробуємо дати інтерпретацію для навантажень перших двох компонент. На діаграмі навантажень першої компоненти видно, що навантаження на змінні є додатними та відрізняються не катастрофічно. Тобто, якщо рухатися вздовдж першого головного напрямку, то початкові значення зростатимуть більш-менш однаково. Тобто вісь за першим головним напрямком можна вважати за ту, що відповідає за розміри трилобіта. Щодо навантажень на другу компоненту, то тут ситуація дещо інша: на довжину припадає додатнє навантаження, а на інші виміри (ширина різних частин трилобіта) – від'ємні. Якщо рухатися вздовж другого головного напрямку, то довжина збільшується, а ширини зменшуються. Вісь за цим напрямком відіграє роль зміни довжини до ширини частин трилобіта.

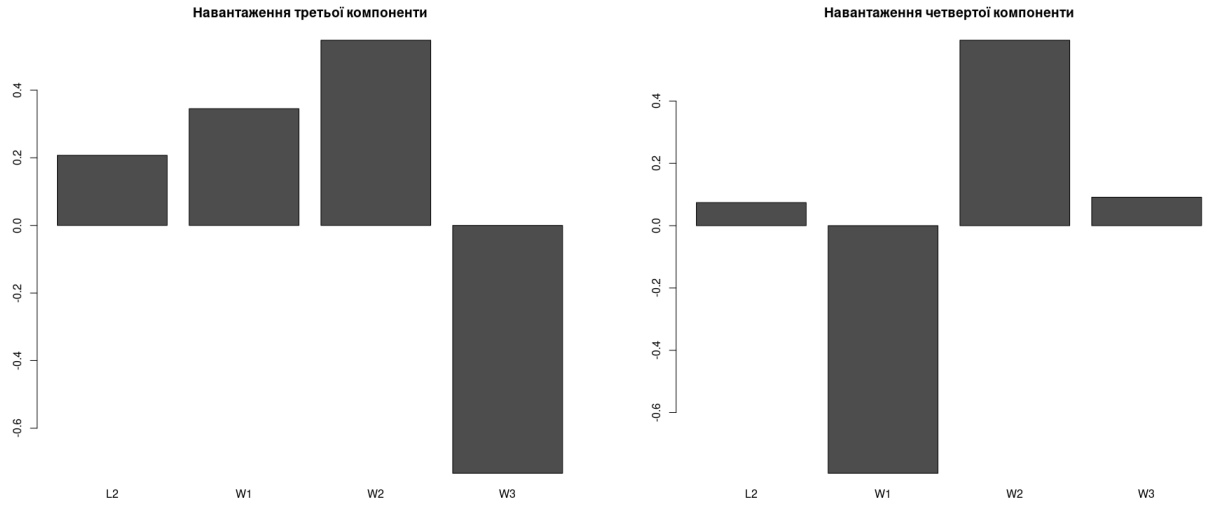

Рис. 10: Діаграми навантажень для останніх двох компонент.

Для останніх двох компонент характер навантажень досить схожий: на одну з ширин навантаження від'ємне, а на всі інші виміри навантаження виходять додатними. Це вже дещо важче інтерпретувати. Можливо, це якось пов'язано із вимірами ширин відповідних частин трилобіта.

Спробуємо виявити розмежування в залежності від виду трилобіта. Спроєктуємо початкові дані на перші дві компоненти та дослідимо форми розкаду на діаграмі розсіювання:

```
plot(tri.pca$scores[,1], tri.pca$scores[,2],
     xlab = "Comp.1", ylab = "Comp.2", col = RGB[tri$K], cex= 0.75, lwd = 2)
grid(); legend("topright", legend = CODENAMES, col = RGB, pch = 1)
```

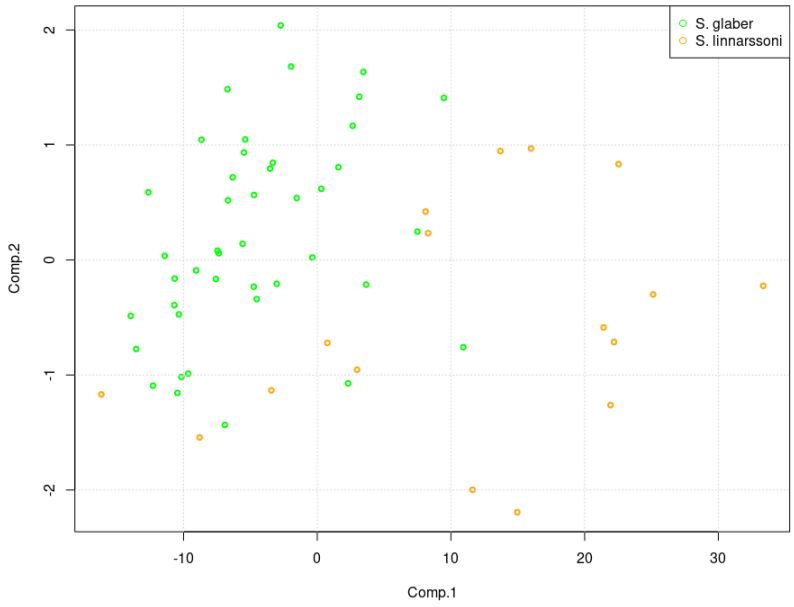

Рис. 11: Діаграма розсіювання на перші дві головні компоненти.

Якщо зробити проекцію даних на перші два головні напрями, то маємо непогане розділення між хмаринами даних в залежності від виду трилобітів. Стик цих двох хмарин можна побачити в області слабкого змішування точок з кожного класу. У порівнянні з тим, що спостерігалося на діаграмах розсіювання для не трансформованих даних, то дане розділення є у першому наближенні чітким. Цим результатом можна скористатися для того, щоб побудувати "рішуче правило", яке б відносило спостереження до одного з класів.

Лінійний дискримінантний аналіз на головні компоненти.

Використовуючи попередні міркування та результати, можемо побудувати модель лінійного дискримінантного аналізу на головні компоненти:

```
# Проектуємо початкові дані на перші два головні напрямки
tri.proj <- data.frame(cbind(tri.pca$scores[,c(1,2)], tri$Code))
colnames(tri.proj)[3] <- "Code"
# Робимо підгонку у моделі лін. диским. аналізу
tri.pca.lda <- lda(Code ~ ., data = tri.proj)
tri.pca.lda
```

Маємо такий звіт з підгонки моделі:

```
Call:
lda(Code ~ ., data = tri.proj)

Prior probabilities of groups:
        1         2
0.7166667 0.2833333

Group means:
    Comp.1     Comp.2
1 -4.52495  0.2184322
2 11.44546 -0.5525051

Coefficients of linear discriminants:
              LD1
Comp.1  0.1143385
Comp.2 -0.7489711
```

Апостеріорний розподіл виявився зсунутим на користь трилобітів S. glaber внаслідок їхньої переважної кількості у розширеній вибірці. Можна ще скористатися технікою крос-валідації у даній моделі для визначення апостеріорного розподілу виду трилобіта:

```
tri.pca.lda.cv <- lda(Code ~ ., data = tri.proj, CV = T)
tri.pca.lda.cv$posterior
```

Нижче наводимо перші десять варіантів апостеріорної ймовірності за відповідними спостереженнями:

	$P(\text{Code} = \text{Glaber} \mid X)$	$P(\text{Code} = \text{Linnarssoni} \mid X)$
1	0.9987617	0.001238289
2	0.9410635	0.058936516
3	0.9915115	0.008488482
4	0.9938321	0.006167881
5	0.9883142	0.011685772
6	0.9951182	0.004881766
7	0.9871923	0.012807735
8	0.9967474	0.003252628
9	0.9883875	0.011612529
10	0.3563010	0.643699045
⋮	⋮	⋮

Табл. 3: Апостеріорний розподіл виду трилобіта в залежності від спостереження.

За цим апостеріорним розподілом робиться прогноз виду, до якого може належати трилобіт. Покажемо таблицю спряженості для прогнозу та справжньому значенню виду:

```
table(forecast = tri.pca.lda.cv$class, real = tri$Code)
```

Прогноз \ Відгук	Glaber	Linnarssoni
Glaber	41	7
Linnarssoni	2	10

Табл. 4: Таблиці спряженості для прогнозу і відгука.

Помилка класифікації становить $(7+2)/(7+2+41+10) = 0.15$, що не вийшла зовсім великою. Похибку спровокували спірні спостереження, які створювали ефекти змішування на стику хмарин двох класів.

На завершення скористаємося класифікуючими функціями, що базуються на моделі лінійного дискримінантного аналізу. Застосуємо таку реалізацію, взяту з [?]:

```
ty.lda <- function(x, groups){
  x.lda <- lda(groups ~ ., as.data.frame(x))
  gr <- length(unique(groups))   ## groups might be factors or numeric
  v <- ncol(x) ## variables
  m <- x.lda$means ## group means
  w <- array(NA, dim = c(v, v, gr))
  for(i in 1:gr){
    tmp <- scale(subset(x, groups == unique(groups)[i]), scale = FALSE)
    w[,,i] <- t(tmp) %*% tmp
  }
  W <- w[,,1]
  for(i in 2:gr)
    W <- W + w[,,i]
  V <- W/(nrow(x) - gr)
  iV <- solve(V)
  ...
```

```
  ...
  class.funs <- matrix(NA, nrow = v + 1, ncol = gr)
  colnames(class.funs) <- paste("group", 1:gr, sep=".")
  rownames(class.funs) <- c("constant", paste("var", 1:v, sep = "."))
  for(i in 1:gr) {
    class.funs[1, i] <- -0.5 * t(m[i,]) %*% iV %*% (m[i,])
    class.funs[2:(v+1) ,i] <- iV %*% (m[i,])
  }
  x.lda$class.funs <- class.funs
  return(x.lda)
}
```

Залишається викликати цю функцію коректно:

```
> ty.lda.fit <- ty.lda(tri.proj[,c(1,2)], tri$Code)
> ty.lda.fit
Call:
lda(groups ~ ., data = as.data.frame(x))

Prior probabilities of groups:
        G         L
0.7166667 0.2833333

Group means:
    Comp.1     Comp.2
G -4.52495  0.2184322
L 11.44546 -0.5525051

Coefficients of linear discriminants:
              LD1
Comp.1  0.1143385
Comp.2 -0.7489711
> ty.lda.fit$class.funs
             group.1    group.2
constant -0.23186370 -1.4834463
var.1    -0.07786168  0.1969442
var.2     0.51003087 -1.2900781
```

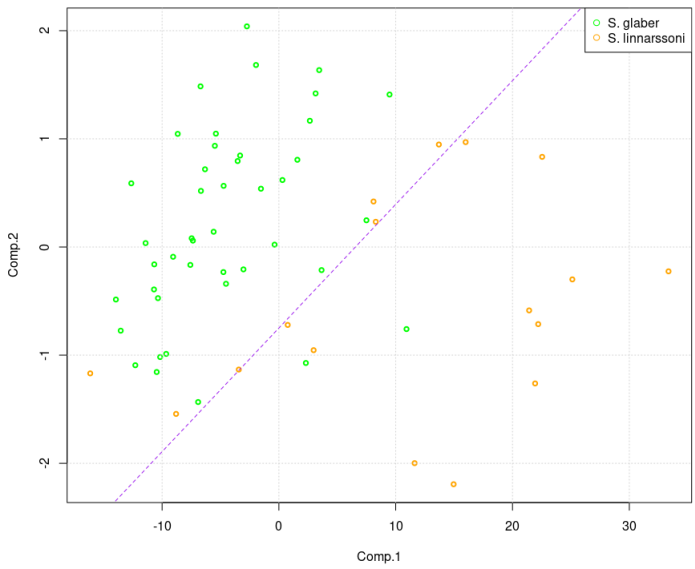

Рис. 12: Діаграма розсіювання на перші дві головні компоненти. Фіолетовим пунктиром відмічена дискримінантна (розділяюча) пряма.

Висновки.

Незважаючи на критично малий обсяг вибірки з трилобітів, зробити адекватні висновки про залежність спостережуваних змінних від виду можна. Тим не менш, побудувати класифікатор теж вийшло без особливих труднощів, використавши проектування на головні компоненти. Хоча будемо чесні, використання деяких тестів (на прикладі MANOVA) базувалися з дуже наївних припущень. Хотілося б мати більш змістовну вибірку (хоча б в тому сенсі, щоб за кожним видом було багато спостережень), однак наразі немає ніяких додаткових можливостей аби це зробити.

Література

[1] Bruton, D. L. & Owen, A. W.: The Norwegian Upper Ordovician illaenid trilobites. Norsk Geologisk Tidsskrift, Vol. 68, pp. 241-258. Oslo 1988. ISSN 0029-196X.

[2] Legendre and Legendre's Numerical Ecology (1998), page 625.