HW Chapter 8

8.6. Data on x_1 = sales and x_2 = profits for the 10 largest U.S. industrial corporations were listed in Exercise 1.4 of Chapter 1.

From Example 4.12

$$\bar{x} = \begin{bmatrix} 62309 \\ 2927 \end{bmatrix}, \quad S = \begin{bmatrix} 1000520000 & 25576000 \\ 25576000 & 1430000 \end{bmatrix}$$

a) Determine the sample principal components and their variances for these data.

$\hat{\lambda}_1 = 1.0012x10^9$, $\hat{e}_1 = \begin{bmatrix} 0.9997 \\ 0.0256 \end{bmatrix}$ or $\hat{e}_1 = \begin{bmatrix} -0.9997 \\ -0.0256 \end{bmatrix}$

$\hat{\lambda}_2 = 7.7570x10^5$, $\hat{e}_2 = \begin{bmatrix} -0.0256 \\ 0.9997 \end{bmatrix}$ or $\hat{e}_2 = \begin{bmatrix} 0.0256 \\ -0.9997 \end{bmatrix}$

$\hat{y}_i = \hat{e}_i' x = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2$

The sample principal components:

$\hat{y}_1 = 0.9997 x_1 + 0.0256 x_2$

$\hat{y}_2 = -0.0256 x_1 + 0.9997 x_2$

Notice here that the variable x_1, with coefficient 0.9997, receives the greatest weight in the component $\hat{y}_1$. It also has the largest correlation (in absolute value) with $\hat{y}_1$ [see Part d]. That x_1 contributes more to the determination of $\hat{y}_1$ than does x_2

Their variances:

Sample variance $(\hat{y}_1) = \hat{e}_1' S \hat{e}_1 = \hat{\lambda}_1 = 1.0012x10^9$

Sample variance $(\hat{y}_2) = \hat{e}_2' S \hat{e}_2 = \hat{\lambda}_2 = 7.7570x10^5$

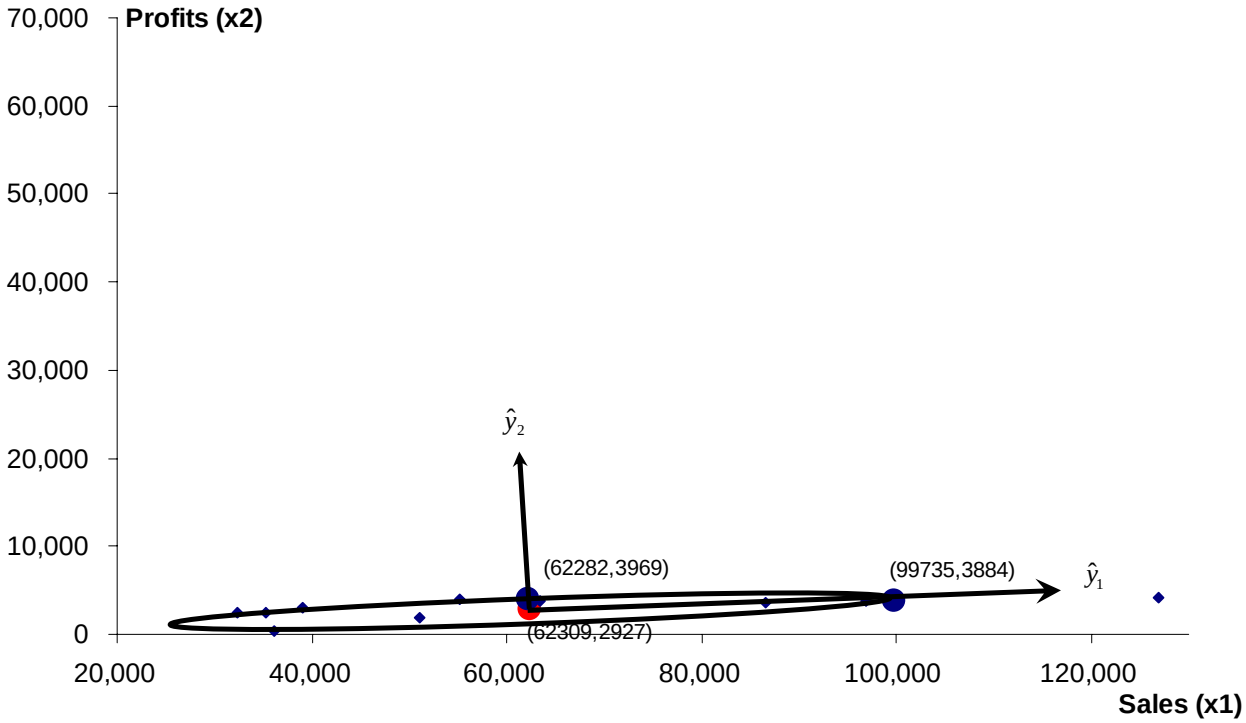
Sample covariance $(\hat{y}_1, \hat{y}_2) = \hat{e}_1' S \hat{e}_2 = 0$

Notice because of its large variance, x_1 completely dominates the first sample principal component. Moreover, this first sample principal component explains completely. [see Part b]

b) Find the proportion of the total sample variance explained by $\hat{y}_1$

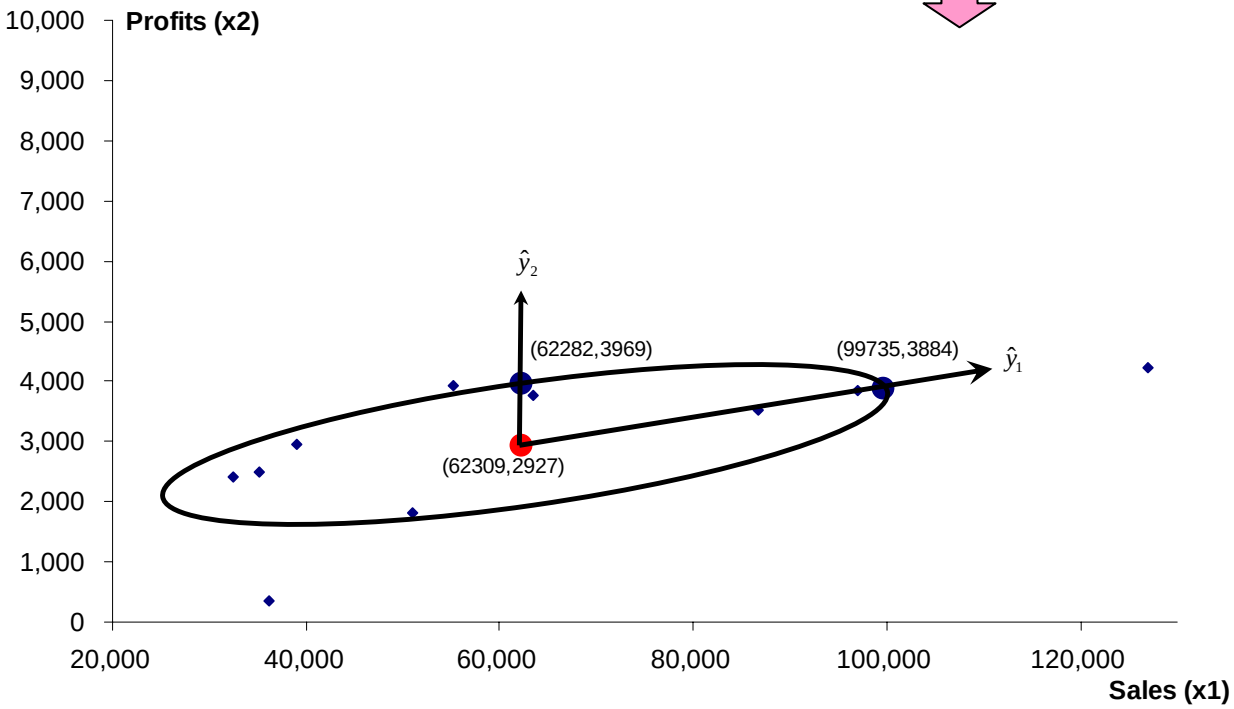
$$\begin{pmatrix} \text{the proportion of} \\ \text{the total sample variance} \\ \text{explained by } \hat{y}_1 \end{pmatrix} = \frac{\hat{\lambda}_1}{\hat{\lambda}_1 + \hat{\lambda}_2} = 0.9992 \text{ or } 99.92\%$$

c) Sketch the constant density ellipse $(x-\bar{x})'S^{-1}(x-\bar{x})=1.4$, and indicate the principal components $\hat{y}_1$ and $\hat{y}_2$ on your graph.



รูปที่ 1 **Sketch the constant density ellipse** $(x-\bar{x})'S^{-1}(x-\bar{x})=1.4$

กำหนดให้แกนนอน และแกนตั้งมีช่วงห่างที่เท่ากัน



รูปที่ 2 **Sketch the constant density ellipse** $(x-\bar{x})'S^{-1}(x-\bar{x})=1.4$

กำหนดให้แกนนอน และแกนตั้งมีช่วงห่างไม่เท่ากัน

Note1: รูปที่ 2 เนื่องจาก แกนนอนและแกนตั้งมี Scale ต่างกันจึงทำให้แกนของวงรี(Sample principal components $\hat{y}_1$ **and** $\hat{y}_2$) วาดไม่เป็นเส้นตั้งฉากกัน ซึ่งถ้ากำหนดให้แกนนอน และแกนตั้งมีช่วงห่างที่เท่ากันจะทำให้พิจารณารูปร่างวงรีได้ยากดังรูปที่ 1 ดังนั้นการวาดรูปวงรีที่ขจัด Scale ของแกนตั้งและแกนนอนออกไปให้เป็นแกนที่ไร้หน่วยนั้นจะทำให้ แกนของวงรี(Sample principal components $\hat{y}_1$ **and** $\hat{y}_2$) วาดเป็นเส้นตั้งฉากกัน และพิจารณารูปได้ง่ายกว่า สามารถจัดทำได้โดยการ Standardizing the sample principal components ดังรูปที่ 3 [see 8.7]

Note2: การหา $\sqrt{\hat{\lambda}_i}\sqrt{c^2}\hat{e}_i \qquad i=1,2$

$\longrightarrow \hat{y}_1$ หาจาก $\hat{\lambda}_i$ ที่มีค่ามากที่สุด $\sqrt{1.0012x10^9}\sqrt{1.4}\begin{bmatrix}0.9997\\0.0256\end{bmatrix}=\begin{bmatrix}37426\\957\end{bmatrix}$

ดังนั้นคู่อันดับที่จะนำไป Plot ในกราฟคือ $\begin{bmatrix}37426\\957\end{bmatrix}+\begin{bmatrix}62309\\2927\end{bmatrix}=\begin{bmatrix}99735\\3884\end{bmatrix}$

$\longrightarrow \hat{y}_2$ หาจาก $\hat{\lambda}_i$ ที่มีค่าถัดมา $\sqrt{7.7570x10^5}\sqrt{1.4}\begin{bmatrix}-0.0256\\0.9997\end{bmatrix}=\begin{bmatrix}-26.7\\1041.8\end{bmatrix}$

ดังนั้นคู่อันดับที่จะนำไป Plot ในกราฟคือ $\begin{bmatrix}-26.7\\1041.8\end{bmatrix}+\begin{bmatrix}62309\\2927\end{bmatrix}=\begin{bmatrix}62282.3\\3968.8\end{bmatrix}$

d) Compute the correlation coefficients $r_{\hat{y}_1,x_k}, k=1,2.$ What interpretation, if any, can you give to the first principal component?

เมื่อ eigenvectors คือ $\hat{e}_1 = \begin{bmatrix} 0.9997 \\ 0.0256 \end{bmatrix}$, $\hat{e}_2 = \begin{bmatrix} -0.0256 \\ 0.9997 \end{bmatrix}$

$$r_{\hat{y}_1,x_1} = \frac{\hat{e}_{11}\sqrt{\hat{\lambda}_1}}{\sqrt{s_{11}}} = \frac{0.9997\sqrt{1.0012x10^9}}{\sqrt{1000520000}} = 1$$

$$r_{\hat{y}_1,x_2} = \frac{\hat{e}_{12}\sqrt{\hat{\lambda}_1}}{\sqrt{s_{22}}} = \frac{0.0256\sqrt{1.0012x10^9}}{\sqrt{1430000}} = 0.6767$$

Interpretation

The variable x_1, with coefficient 0.9997 [see part a], receives the greatest weight in the component $\hat{y}_1$. It also has the largest correlation (in absolute value) with $\hat{y}_1$, ($r_{\hat{y}_1,x_1} = 1$). The correlation of x_1 with $\hat{y}_1$ = 1 is the largest of correlation with $\hat{y}_1$. That x_1 contributes more to the determination of $\hat{y}_1$ than does x_2. However, that x_2 has coefficient 0.0256 and the correlation = 0.6767 with $\hat{y}_1$, in this case, both variables aid in the interpretation of $\hat{y}_1$.

Note: อย่างไรก็ตามถึงแม้ว่า x_2 จะสามารถอธิบาย $\hat{y}_1$ ได้ เพราะมีค่าสัมประสิทธิ์ 0.0256 ในสมการของ $\hat{y}_1$ แต่ค่านี้เมื่อเทียบกับ $x_1 = 0.9997$ แล้วมีค่าน้อยมากๆ จนอาจจะกล่าวได้ว่า $\hat{y}_1$ นั้นถูกอธิบายได้จากตัว x_1 อย่างมาก โดยจะอธิบายอีกครั้งในข้อ 8.7

Can you give to the first principal component?

Yes, first sample principal component explains a proportion 0.9992 of the total population variance. The second sample principal component is unimportant.

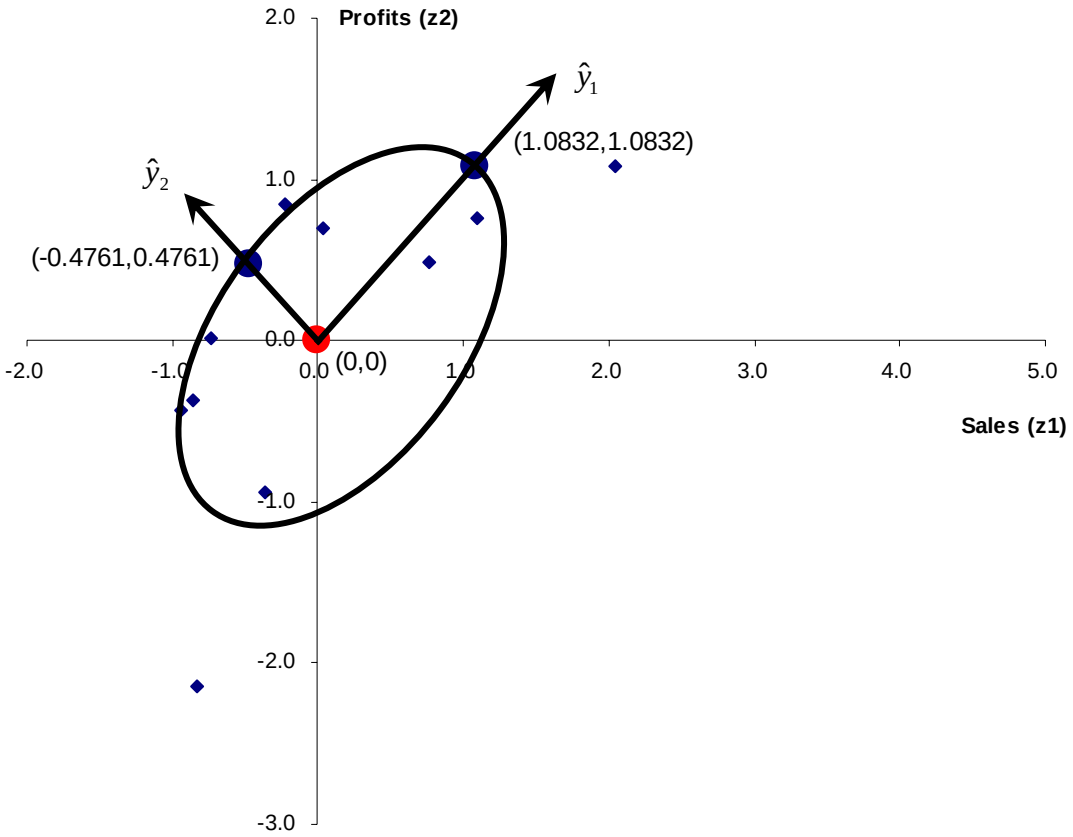
Note : เนื่องจาก $\hat{e}_{ik}$ = สามารถกลับเครื่องหมายได้

เมื่อ eigenvectors คือ $\hat{e}_1 = \begin{bmatrix} -0.9997 \\ -0.0256 \end{bmatrix}$, $\hat{e}_2 = \begin{bmatrix} 0.0256 \\ -0.9997 \end{bmatrix}$

$$r_{\hat{y}_1,x_1} = \frac{\hat{e}_{11}\sqrt{\hat{\lambda}_1}}{\sqrt{s_{11}}} = \frac{-0.9997\sqrt{1.0012x10^9}}{\sqrt{1000520000}} = -1$$

$$r_{\hat{y}_1,x_2} = \frac{\hat{e}_{12}\sqrt{\hat{\lambda}_1}}{\sqrt{s_{22}}} = \frac{-0.0256\sqrt{1.0012x10^9}}{\sqrt{1430000}} = -0.6767$$

8.7. Convert the covariance matrix S in Exercise 8.6 to a sample correlation



รูปที่ 3 Sketch the constant density ellipse of standardizing the sample principal components

a) Find the sample principal of the total sample variance explained by $\hat{y}_1$, $\hat{y}_2$ and their variances.

matrix R.

$$\left(V^{1/2}\right)^{-1} S \left(V^{1/2}\right)^{-1} = R = \begin{bmatrix} 1 & 0.6762 \\ 0.6762 & 1 \end{bmatrix}$$

$$\hat{\lambda}_1 = 1.6762 \quad , \hat{e}_1 = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix} \quad \text{or} \quad \hat{e}_1 = \begin{bmatrix} -0.7071 \\ -0.7071 \end{bmatrix}$$

$$\hat{\lambda}_2 = 0.3238 \quad , \hat{e}_2 = \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix} \quad \text{or} \quad \hat{e}_2 = \begin{bmatrix} 0.7071 \\ -0.7071 \end{bmatrix}$$

$$\hat{y}_i = \hat{e}_i' z = \hat{e}_{i1} z_1 + \hat{e}_{i2} z_2$$

The sample principal components:

$$\hat{y}_1 = 0.7071 z_1 + 0.7071 z_2$$

$$\hat{y}_2 = -0.7071 z_1 + 0.7071 z_2$$

Their variances:

Sample variance $(\hat{y}_1) = \hat{e}_1' R \hat{e}_1 = \hat{\lambda}_1 = 1.6762$

Sample variance $(\hat{y}_2) = \hat{e}_2' R \hat{e}_2 = \hat{\lambda}_2 = 0.3238$

Sample covariance $(\hat{y}_1, \hat{y}_2) = \hat{e}_1' R \hat{e}_2 = 0$

b) Compute the proportion of the total sample variance explained by $\hat{y}_1$.

$$\begin{pmatrix} \text{the proportion of} \\ \text{the total sample variance} \\ \text{explained by } \hat{y}_1 \end{pmatrix} = \frac{\hat{\lambda}_1}{p} = \frac{\hat{\lambda}_1}{tr(R)} = \frac{\hat{\lambda}_1}{\hat{\lambda}_1 + \hat{\lambda}_2} = 0.8381 \qquad \text{or } 83.81\%$$

c) Compare the correlation coefficients $r_{\hat{y}_1, z_k}, k = 1, 2.$ Interpret $\hat{y}_1$.

เมื่อ eigenvectors คือ $\hat{e}_1 = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}$, $\hat{e}_2 = \begin{bmatrix} -0.7071 \\ 0.7071 \end{bmatrix}$

$$r_{\hat{y}_1, z_1} = \hat{e}_{11}\sqrt{\hat{\lambda}_1} = 0.7071\sqrt{1.6762} = 0.9155$$

$$r_{\hat{y}_1, z_2} = \hat{e}_{12}\sqrt{\hat{\lambda}_2} = 0.7071\sqrt{1.6762} = 0.9155$$

Note : เนื่องจาก $\hat{e}_{ik}$ = สามารถกลับเครื่องหมายได้ ดังนั้นถ้ากลับเครื่องหมายเป็น

เมื่อ eigenvectors คือ $\hat{e}_1 = \begin{bmatrix} -0.7071 \\ -0.7071 \end{bmatrix}$, $\hat{e}_2 = \begin{bmatrix} 0.7071 \\ -0.7071 \end{bmatrix}$

$$r_{\hat{y}_1, z_1} = \hat{e}_{11}\sqrt{\hat{\lambda}_1} = -0.7071\sqrt{1.6762} = -0.9155$$

$$r_{\hat{y}_1, z_2} = \hat{e}_{12}\sqrt{\hat{\lambda}_2} = -0.7071\sqrt{1.6762} = -0.9155$$

Interpretation

The variable z_1 and z_2, with same coefficient 0.7071, receive great weight in the component $\hat{y}_1$. They also have large correlation (in absolute value) with $\hat{y}_1$, ($r_{\hat{y}_1, x_1} = 0.9155, r_{\hat{y}_1, x_2} = 0.9155$). The correlation of z_1 is as large as that for z_2, indicating that the variables are about equally important to the first sample principal component. Further, in this case, both coefficients are reasonably large and they have same sign, we would argue that both variables aid in the interpretation of $\hat{y}_1$.

d) Compare the components obtained in Part a with those obtained in Exercise 8.6(a). Given the original data displayed in Exercise 1.4, do you feel that it is better to determine principal components from the sample covariance matrix or sample correlation matrix? Explain.

The sample principal components of 8.6:

$\hat{y}_1 = 0.9997x_1 + 0.0256x_2$

$\hat{y}_2 = -0.0256x_1 + 0.9997x_2$

The sample principal components of 8.7:

$\hat{y}_1 = 0.7071z_1 + 0.7071z_2$

$\hat{y}_2 = -0.7071z_1 + 0.7071z_2$

การหา The sample principal components จาก S ดังในข้อ 8.6 ถึงแม้ว่า x_2 จะสามารถอธิบาย $\hat{y}_1$ ได้ เพราะมีค่าสัมประสิทธิ์ 0.0256 ในสมการของ $\hat{y}_1$ แต่ค่านี้เมื่อเทียบกับ $x_1 = 0.9997$ แล้วมีค่าน้อยมากๆ จนอาจจะกล่าวได้ว่า $\hat{y}_1$ นั้นถูกอธิบายได้จากตัว x_1 อย่างมากแต่เพียงตัวแปรเดียว แต่ในการหา The sample principal components จาก R ดังในข้อ 8.7 นั้นแตกต่างกันเพราะทั้งตัวแปร z_1 และ z_2 สามารถอธิบายตัวแปร $\hat{y}_1$ ได้ดีเท่าเทียมกัน ไม่สามารถตัดตัวแปรใดทิ้งไปได้(พิจารณาจาก correlation coefficients ทั้งคู่เท่ากับ 0.9155) ซึ่งการตัดสินใจเลือกใช้แบบใดนั้นขึ้นอยู่กับเหตุผลดังนี้ (ประกอบการให้เหตุผลจาก Richard A. Johnson and Dean W. Wichern, Applied multivariate statistics analysis, fifth edition, page 435.)

"Variables should probably be standardized if they are measured on scales with widely differing ranges or if the units of measurement are not commensurate. For example, if x_1 represents annual sales in the the $10,000 to $350,000 range and x_2 is the ratio (net annual income)/(total assets) that falls in the .01 to .60 range, then the total variation will be due almost exclusively to dollar sales. In this case, we would expect a single (important) principal component with a heavy weighting of x_1. Alternatively, if both variables are standardized, their subsequent magnitudes will be of the same order, and x_2 (or z_2) will play a larger role in the construction of the principal components."

ดังนั้นเมื่อพิจารณาข้อมูล Exercise 1.4 พบว่า ตัวแปรทั้ง 2 **<u>มีช่วงของข้อมูลแตกต่างกันอย่างมาก</u>** การพิจารณาหาค่าด้วย The sample principal components จาก Sample correlation matrix จึงมีความเหมาะสมกว่า ดังนั้นอธิบาย The first sample principal component of $\hat{y}_1$ ด้วย $\hat{y}_1 = 0.7071z_1 + 0.7071z_2$ โดยจะพบว่าค่าของช่วงข้อมูลที่มากกว่าในตัวแปร x_1 ที่มีค่าความแปรปรวนมาก จะถูกนำมาลดค่าสัมประสิทธิ์ของ The sample principal components มากกว่าตัวแปร x_2 เมื่อแปลง $z \Rightarrow x$ ดังนี้

$\hat{y}_1 = 0.7071z_1 + 0.7071z_2$

$\hat{y}_1 = \frac{0.7071}{\sqrt{1000520000}}(x_1 - \bar{x}_1) + \frac{0.7071}{\sqrt{143000000}}(x_2 - \bar{x}_2)$

$\hat{y}_1 = 0.0000223547(x_1 - \bar{x}_1) + 0.000591(x_2 - \bar{x}_2)$