

『データサイエンスのための統計学入門』（オライリー・ジャパン）の翻訳の誤り

♡ 3



Yutaka MOTEKI
2023年1月20日 13:35



『データサイエンスのための統計学入門 第2版』は、副題が「予測、分類、統計モデリング、統計的機械学習とR/Pythonプログラミング」となっている。原著が「Practical Statistics for Data Scientists」であることからわかるように、内容が、データを扱うサイエンティストにとってたいへんに実用的なものであり、統計分析の有用な手引き書になっている。しかし、よく読んでみると、意味の通らないところなどが各所に見られる。

▼ 目次

「作業割り当ての判断」とコレスポネンス分析

図示することの意味

表(table)と2×2行列との違い

[補足]

「作業割り当ての判断」とコレスポネンス分析

以下のような文がある。コレスポネンス分析が取りあげられている項の中である。

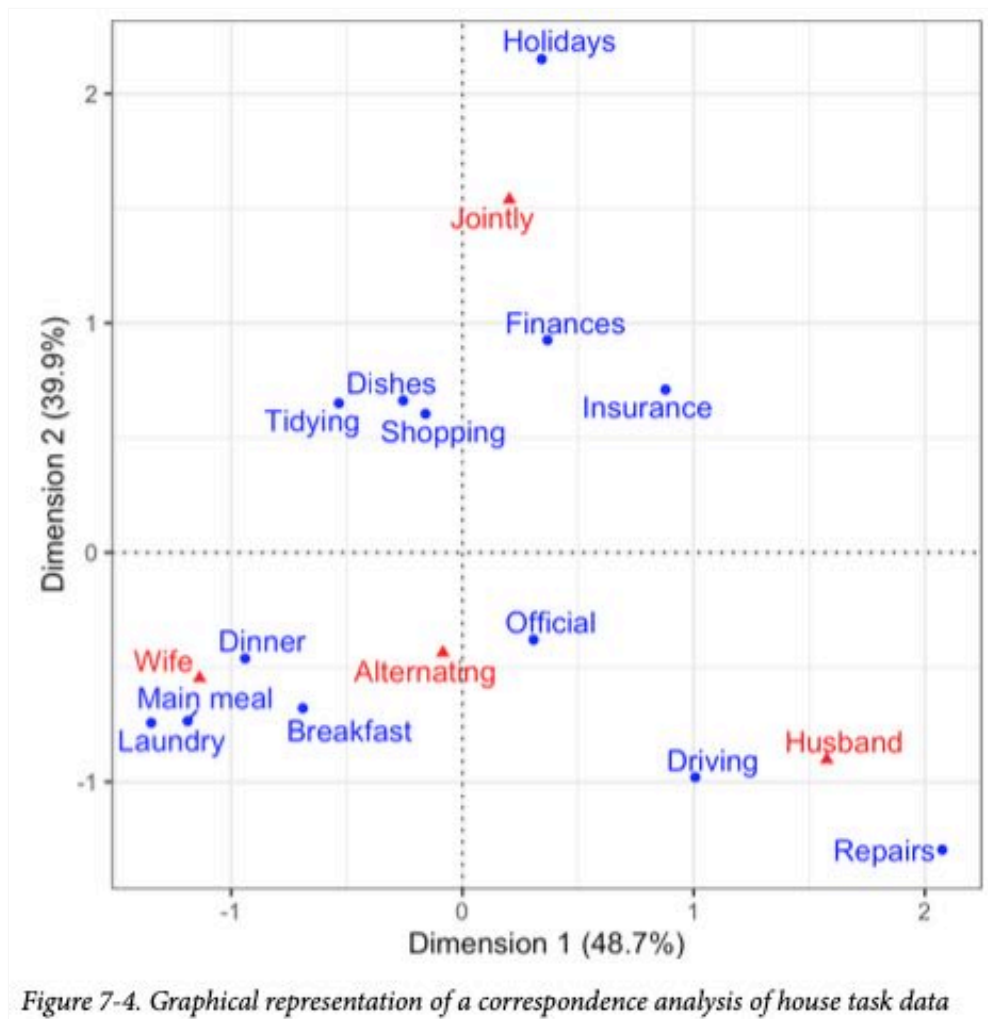
「Correspondence analysis is many decades old, as is the spirit of this example, judging by the assignment of tasks.」

(Practical Statistics for Data Scientists, 2nd Edition / by Peter Bruce, Andrew Bruce, Peter Gedeck. p.292)

これが次のように訳されている。

「コレスポネンス分析には数十年の歴史があり、この例に示すように作業割り当ての判断に用いられる。」(訳書308ページ)

図は以下の通り。



原書に掲載されているグラフ

翻訳の方を読んでいて、どうも意味が不明で納得がいかないので考えてみた。

「この例に示されている**spirit**」 (the spirit of this example) というものを原著者は「古い」あるいは「古くさい」と思っているということなのではないか。「作業割り当ての判断に用いられる」という訳は、コレスポネンス分析の利用場面を述べているように受け取れるので、間違っているように思われる。「judging by the assignment of tasks」の部分の誤訳であろう。

図示することの意味

図の説明の部分が以下のように訳されている。「図では、.....を示す」という

表現がわかりにくい。「図では、.....を示す」のではなく、家事13種類が「図に配置されている」ということではないか。横軸と縦軸の解釈はそれに基づいてなされるものであろう。

「例えば、図7-4では、家事を夫婦が共同で行うか単独で行うかを縦軸に、主たる責任を負うのは妻か夫かを横軸に示す。」（訳書308ページ）

英文は以下の通り。

「See for example, Figure 7-4, in which household tasks **are arrayed** according to whether they are done jointly or solo (vertical axis), and whether wife or husband has primary responsibility (horizontal axis).」

夫婦が家事をどう行うのかという変数と家事の種類の変数とが元のデータである。横軸の解釈として「夫か妻か」ということであり、縦軸の解釈として「共同か単独か」ということになる。横軸も縦軸も、家事の種類ではなく夫婦が家事をどう行うのかという変数に即して説明できるとみなされているようだ。これについては異論はない。しかし、図が、「家事を夫婦が共同で行うか単独で行うかを縦軸」に示したり、「主たる責任を負うのは妻か夫かを横軸」に示したりしていると判断できるのは、図上の2変数の配置を眺めてみることによってであろう。

表(table)と2×2行列との違い

おかしいと思うところは誤訳のようだ。次のような文が同じページにある。

「コレスポネンス分析の入力は、行が1変数、列が他の変数を表す 2×2 行列で、各セルがレコード数を表す。」

「 2×2 行列」というのは変だと思って英文を見ると、次のようになっている。

「The input can be seen as a table, with rows representing one variable and columns another, and the cells representing record counts.」

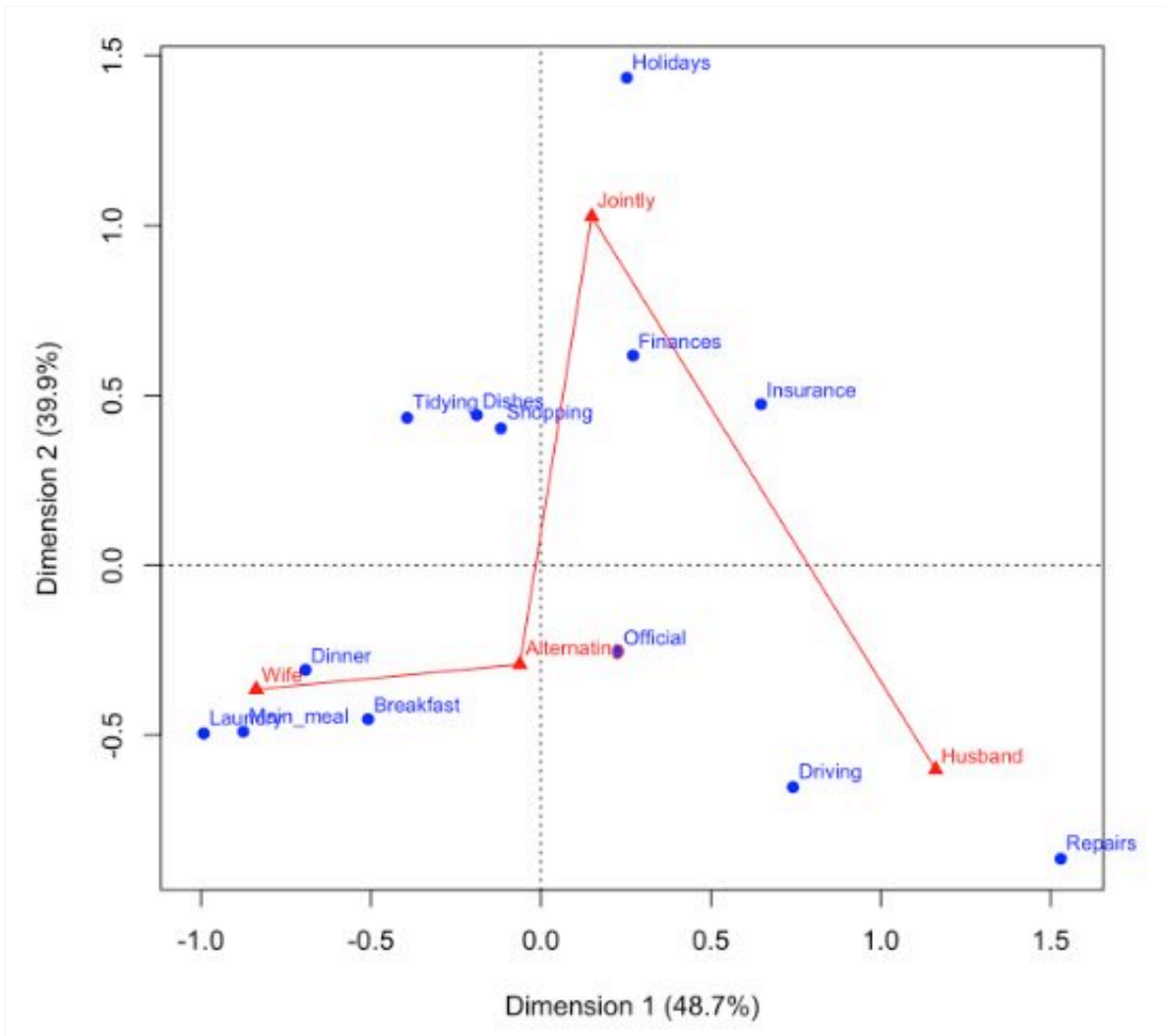
「 2×2 行列」は、tableのここのようだ。tableをそのように訳すのは間違いである。何かを勘違いしているのであろう。図から見てもわかるが、 4×13 の表 (**table**)が元のデータであろう。

翻訳書のたったの1ページにこんなにおかしなところがある。訳者の問題なのだろうが、「技術監修者」、「査読協力」者として巻末に出てくる人びとは本当に翻訳をチェックしたのだろうか。

[補足]

caパッケージのplot.ca を利用して以下のように線を引いてみた。

```
ca_analysis <- ca(housetasks)
plot(ca_analysis, lines=c("FALSE","TRUE"))
```



plot.caを利用

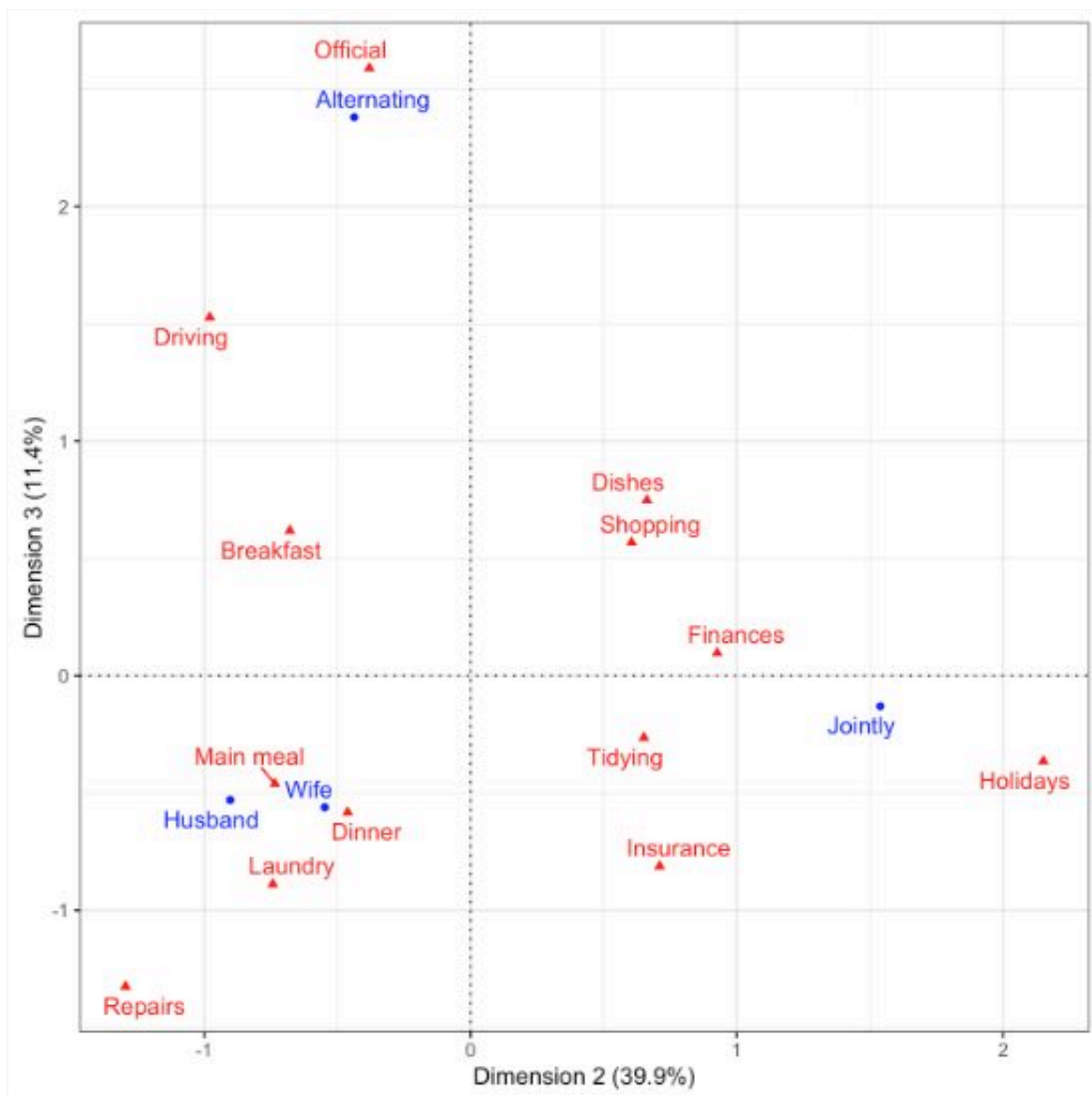
元のデータは以下のような表である。

```
knitr::kable(housetasks)
```

	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

これは、クロス表であるが、 2×2 ではない。

第2次元と第3次元とをそれぞれ図の横軸と縦軸に表示してみた。「交代で担当かどうか」ということが第3次元の意味なのであろう。



第1次元を外して図示

以下がコード

[gedeck/practical-statistics-for-data-scientists](https://github.com/johndcook/practical-statistics-for-data-scientists)

```
ggplot(coords, aes(x=Dim2, y=Dim3, color=type, label=rownames(coords),
  shape=type)) +
  geom_hline(yintercept=0, linetype='dotted', color='#444444') +
  geom_vline(xintercept = 0, linetype='dotted', color='#444444') +
  geom_point() +
  geom_text_repel() +
  xlab(sprintf('Dimension 2 (0.1f%%)', 100 * contrib[2])) +
  ylab(sprintf('Dimension 3 (0.1f%%)', 100 * contrib[3])) +
```



```
scale_color_manual(values = c('blue', 'red')) +  
theme_bw() +  
theme(legend.position = "none")
```

しかし、caパッケージのデフォルトの描画とスケールを同じにするためには以下のスクリプトを付け加える必要がある。

```
coords[,2] <- coords[,2]*ca_analysis$sv[2]  
coords[,3] <- coords[,3]*ca_analysis$sv[3]
```

