Анализ данных с R (III).

© С. В. Петров*, Е. М. Балдин†, В. Е. Лявшук‡

*p2004r@gmail.com
†E.M.Baldin@inp.nsk.su
‡lve@tut.by

Эмблема **R** взята с официального сайта проекта `http://developer.r-project.org/Logo/`

Оглавление

Глава 11

Поиск зависимостей

R можно и нужно применять в реальных исследованиях. Грамотный исследователь с помощью своего интеллекта, знаний статистики и возможностей, предоставляемых **R**, с лёгкостью может ответить на массу загадок мироздания. Если есть данные, естественно.

11.1. Кто оценит преподавателя?

Студенты знают, что за ними постоянно следят и их ценят, а каждый экзамен становится праздником на которых их оценивают. С этим всё в порядке. Но кто оценит преподавателей? Говорят анкетирование поможет. Проверим это.

Предыстория Давным давно в одной «далёкой галактике» провели анкетирование на тему «Преподаватель глазами студентов» и благополучно об этом забыли. После того как с анкет была сдута пыль веков, выяснилось, что данные опроса представлены таблицами. Каждая строка опросной таблицы состоит из средней по группе студентов оценки по соответствующей графе анкеты. Строка также характеризуется курсом, предметом, преподавателем и кафедрой. Формат записей CSV, то есть что-то вроде:

```
Кафедра истории и философии;Галактическая история;Галактион Иванович
Планетный;5-й курс;4,6;4,7;4,3;4,8;4,6;3.6;2,7;...
```

Поиск зависимостей Импортируем и объединяем таблицы:

```
> data <- rbind(read.csv2("анкета1.csv"),
+               read.csv2("анкета2.csv"),
```

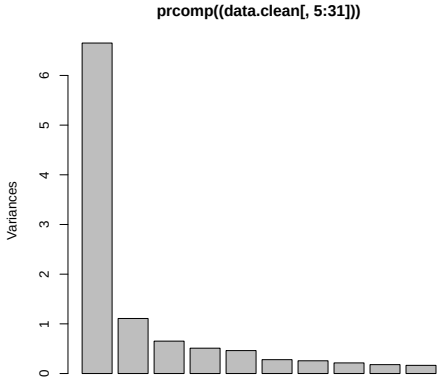

Рис. 11.1. График нагрузок основных компонент

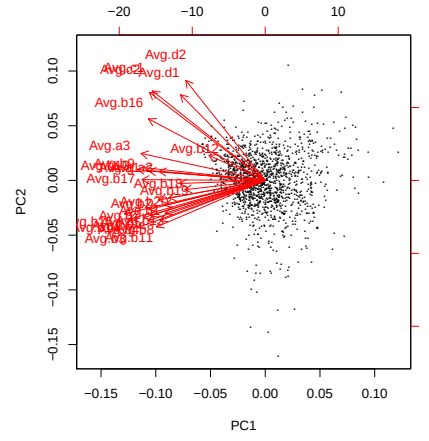

Рис. 11.2. Двойная диаграмма (biplot)

```
+                 ...
+                 read.csv2("анкетаN.csv"))
```

Увы, в реальной жизни никуда не деться от пропущенных значений, поэтому очищаем данные опросов от них:

```
> data.clean <- na.exclude(data)
```

Воспользуемся методом основных компонент, чтобы уменьшить размерность анализируемых данных. Строим график нагрузок основных компонент с помощью команды prcomp. В расчёте участвуют только ответы на вопросы анкеты:

```
> plot(prcomp((data.clean[,5:31])))
```

Присутствует простая структура из двух факторов. Довольно странно, что ответ на анкету из 26 вопросов у студента происходит, исходя всего из *двух* взаимонезависимых факторов. Иными словами студенты отвечали на все многочисленные вопросы анкеты фактически реально исходя всего лишь из двух причин.

Посмотрим на получившуюся простую структуру с целью идентификации двух выделенных в анализе факторов:

```
> biplot(prcomp(data.clean[,5:31]),
+        xlabs= rep( ".", length(data.clean[,3])))
```

Выделенные две взаимно независимые компоненты, путем сопоставления преподавателей участвующих в оценке, с местом которое они заняли в пространстве факторов можно идентифицировать. Первая компонента отражает ось «Было

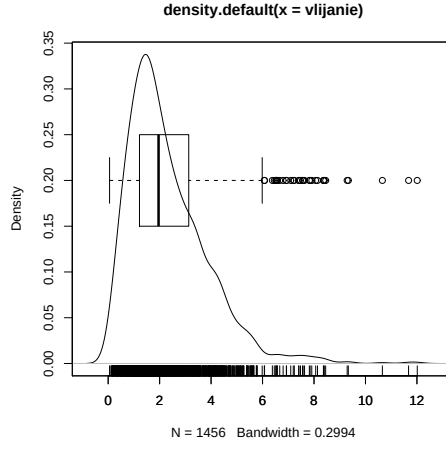

Рис. 11.3. Анализ анкеты

сложно на предмете — Было просто». Вторая компонента отражает ось «Добрый преподаватель — Строгий преподаватель». По соображениям этикета подробности этого процесса мы вынужденно пропускаем.

У составителя анкеты, да и у оцениваемого преподавателя, естественно возникает философский вопрос: «Что лучше? Быть добрым или злым? И что делать если тебе досталась сложная для студентов учебная программа?». Дело в том, что любое простое суммирование баллов набранных преподавателями, с целью получить интегральную оценку, будет эквивалентно проведению под определенным углом через построенное нами факторное пространство оси измерения отражающего данную интегральную оценку.

На самом деле мера определяющая качество взаимодействия в системе «Студент — Преподаватель» имеется. Задумаемся над вопросом: «Кто из преподавателей получил оценки вблизи начала координат факторной плоскости?». Кто всегда получает среднюю оценку? Очевидный ответ: тот, о котором ничего особенного *не помнят*!

Если преподаватель взаимодействовал со студентом достаточно сильно в процессе обучения, то он оставил в его сознании чёткий след. Этот след может отражать и доброту, и строгость, и сложность, и простоту ..., а также их любое сочетание! Иными словами сила взаимодействия в системе «Преподаватель — Студент» — это длина вектора проведенного из начала координат к точке на факторной плоскости соответствующей анкете оцениваемого преподавателя.

Нам остается воспользоваться теоремой великого Пифагора

```
> result.pca <-prcomp((data.clean[,6:32]))$x[,1:2]
> vlijanie <- (result.pca[,1]^2 + result.pca[,2]^2)^0.5
```

Посмотрим каково распределение полученной величины.

```
> plot(density(vlijanie))  # рисуем плотность распределения данных
> rug(jitter(vlijanie))    # добавляем по оси x штрихи значений данных
> boxplot(vlijanie,
+         add= TRUE,         # добавляем на тот же график
+         horizontal=TRUE,  # располагаем boxplot по горизонтали
+         at= 0.2,           # позиция срединной линии boxplot
+         boxwex = 0.2)      # масштаб ширины boxplot
```

Оцененная сила воздействия преподавателя на студентов неоднозначная величина. Безусловно плохо когда преподавателя студенты «не помнят», но что с противоположным случаем? Может ли быть воздействие слишком сильным? Очевидно по аналогии с другими рецепторами организма любое слишком сильное воздействие может вести к повреждению и уж точно ощущается воспринимающим как очевидный дискомфорт.

«Золотая середина» получилась действительно в середине. И это породило определённые трудности. Выделить «лучших», как обычно стремятся организаторы таких опросов принципиально не возможно, так как они плавно «перетекают» в «не лучших».

Дело в том, что вопросы задаваемые в анкете не позволяют измерить именно то, о чём они напрямую спрашивают. Это крайне наивный, как показывает наш анализ, подход. Анкета позволяет только измерить некоторые объективные характеристики исходя из которых отвечает заполняющий анкету.

Выводы Анализируемая анкета (да и любая другая на эту тему) позволит всего лишь узнать кто из преподавателей вообще остался не замечен студентами и его надо подогнать. Ну и присмотреться к тем преподавателям, у кого сила взаимодействия с студентом необычайно сильна.

11.2. Кадровая политика ордена иезуитов

Есть те, кто считают, что история — это не наука. Она, как правило, не говорит на языке математики и в какой-то мере даёт простор для спекуляций, то есть надуманных субъективных оценок. Объективность безусловно всячески приветствуется, но где её взять? Основную опасность для исторического исследования представляет искушение описать ушедшую реальность при помощи классификаций и понятий современности. Но такие понятия как «государство», «экономика», «собственность» содержат внутри себя клише, сложившиеся в контексте современной культуры. Описание прошлого на основе примерки к нему подобных клише несет в себе риск модернизации и ставит под вопрос валидность исторического исследования. Парадокс в том, что одни и те же данные, хранящиеся

в архивных документах, позволяют выдвигать и доказывать диаметрально противоположные гипотезы. В этом случае лучше послушать, что говорят данные сами, без нашей подсказки.

Предыстория Не будем как в случае анкет сдувать архивную пыль. Это вредно для здоровья. Возьмём в руки скучный документ. Его оригинал находится в Риме в Главном архиве ордена иезуитов Archivum Romanum Societatis Iesu (ARSI). Пятый фонд в этом архиве называется Germania и касается жизни Германской Ассистенции Товарищества, куда входили также Польская и Литовская провинция ордена.

Том номер 130 из фонда Germania фактически представляет собой канцелярскую книгу, в которую для сведения генерала заносились предложения провинциалов по кандидатурам на должность настоятеля определенного дома ордена и следующего провинциала. Таких записей с кандидатурами в период с 1684 — 1705 г. в книге зарегистрировано 412, т. е. ведение книги начато при правлении генерала Шарля де Нойэля[1] (1682-86) и окончено со смертью генерала Тирса Гонсалеса [2] (1687–1705). Генералу, как правило, представлялись три кандидатуры (по латыни terno), из которых он обычно выбирал первую по счету (отступлений от правил в книге зарегистрировано только 9). Если же ни одна из трёх предложенных кандидатур не устраивала генерала, то тогда его канцелярия запрашивала у провинциала новое представление. Так произошло в 53 случаях из 412 (доля отвергнутых представлений равна 13%), большинство из которых пришлось на конец правления де Ноэля и первую половину правления Гонсалеса, что свидетельствует об активном воздействии генералов на кадровую политику провинций.

Данные в Germ.130 представлены в виде последовательных записей вида:

[1]NOYELLE Charles de, род. 18 VII 1615 в Брюсселе, вступил в орден 29 IX 1630 в Бельгийской провинции, ум. 12 XII 1686 в Риме. С 1661 г. ассистент Германской ассистенции Товарищества Иисуса. Избран генералом ордена 5 VII 1682 на XII Генеральной Конгрегации. Управлял орденом почти 4 года. В это время в лоне французской церкви усилились галликанские течения, стремящиеся к независимости местной церкви от Рима. В 1683 г. под Веной христианские польско-австрийско-германские войска под командованием Яна III Собеского разбили армию Османской империи. Иезуиты принимали участие в военной кампании в качестве походных капелланов. В 1684 г. генерал инициировал открытие миссии чешских иезуитов в Москве, в которой также приняли участие иезуиты Литовской провинции. Поддерживал деятельность о. Маврикия Вотта ОИ (VOTA (Votta Carlo Maurizio) при дворе Яна III Собеского.

[2]Thyrsus Gonzalez de Santalla, род. 18 I 1624 в Арганза (Испания), вступил в орден 3 III 1643, ум. 27 X 1705 в Риме. Профессор философии и теологии в Саламанке в 1655-65 и 1676-87 гг. Миссионер. Избран генералом ордена 6 VII 1687 на XIII Генеральной Конгрегации и исполнял обязанности до смерти в 1705 г. В 1688 г. направил в Польшу визитатора Игнатия Дертинса (Diertins). Оказал влияние на обращение в католическую веру саксонского электора Фридриха Августа (будущего короля Речи Посполитой Августа II) и поддерживал вовлеченность в политику о. Маврикия Вотта ОИ. Активно боролся против пробабилизма в нравственном богословии (от лат. probabilis — приемлемый, возможный, вероятный — взгляд, согласно которому знание является только вероятным, т. к. истина недостижима), которому приписывал падение нравов. Выступал в печати, в том числе против янсенизма.

	Pro Domo Vilnensi	Pro Domo Nesvisiensis	Pro Domo Varsaviensi
Zawistowski Franciscus	0	2	0
Dzieniszewski Albertus	0	0	0
Rymgayło Josephus	0	1	0
Kołakowski Martinus	0	1	2
Narmunth Nicolaus	4	0	2

Таблица 11.1. Кросстаблица (небольшой кусочек)

```
Pro rectoratu Vilnensi, 1684, Kitnowski Petrus, Krasnodebski Adamus,
Wyrwicz Andreas
```

То есть запись содержит название должности, год, и три кандидатуры, предложенные провинциалом. Иногда (крайне редко) провинциал предлагал менее, чем три кандидатуры за раз, и, видимо, отдельно обосновывал своё предложение. За двадцать лет так случилось только 8 раз (2 раза запись содержит две кандидатуры и 6 раз только одну кандидатуру). Иногда, наоборот, провинциал предлагал больше, чем три кандидатуры на выбор. В период с 1684 г. по 1695 г. так произошло 20 раз (16 раз по 4 кандидатуры, 3 раза по 5 кандидатур и 1 раз 7 кандидатур). В последующие 10 лет ведения книги подобных случаев не зафиксировано. Примечательно, что в несколько чаще нарушение правила выбора из трех кандидатур происходило в случае домов Польской провинции. Так на должность провинциала Польской провинции предлагалось в 1687 г. 4 кандидата, в 1691 г. — 7 кандидатов, в 1695 — 5 кандидатов.

Поиск зависимостей Используя утилиты обработки текстовой информации **awk** и **sed** удалось получили список пар «претендент — должность». Причём должность фактически означает географическую привязку. Данный список был преобразован (свёрнут) в таблицу (в статье приведён лишь небольшой её кусочек), строки которой означали географически локализованные ректорские должности, а столбцы — претендентов:

```
> library("xtable")
> iesu_table <- table(read.table("data_iesu.txt", sep=",")
+                       )[c(137, 27, 112, 55, 92),2:4]
> colnames(iesu_table)<-c("Pro~Domo␣Vilnensi",
+                         "Pro~Domo␣Nesvisiensis",
+                         "Pro~Domo␣Varsaviensi")
> xtable(iesu_table,
> align="lp{1.6cm}p{1.6cm}p{1.6cm}")
```

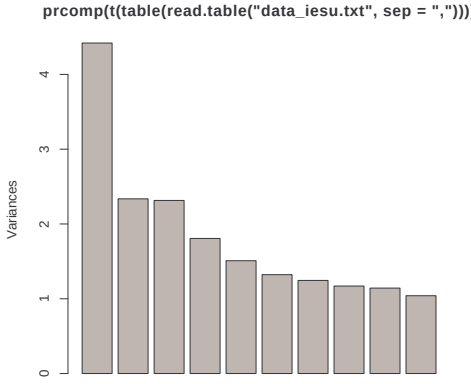

Рис. 11.4. Руководители Ордена иезуитов при принятии кадровых решений руководствовались лишь тремя факторами

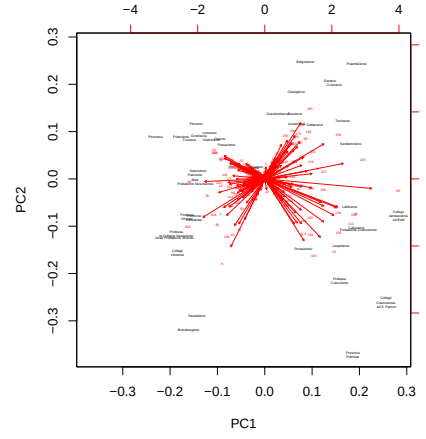

Рис. 11.5. Двойная диаграмма первых двух факторов

На пересечение строки и столбца помещалось количество выдвижений данного кандидата на данную должность за весь период наблюдений. Таблица была обработана методом основных компонент.

```
> plot(prcomp(t(table(read.table("data_iesu.txt",
+                                sep=","")))))
```

Вскрытая анализом факторная структура состоит из трёх факторов. Следующий шаг: посмотрим на реальные данные в пространстве первых двух факторов.

```
> labelxs <- gsub("_",
+                 "␣",
+                 rownames(t(table(read.table("data_iesu.txt",
+                                             sep=","))))))
> labelxs <- sub("(Pro␣rectoratu)␣", "", labelxs)
> labelxs <- sub("(Pro␣Domo)␣", "", labelxs)
> labelxs <- sub("(tiae)␣", "tiae\n␣", labelxs)
> labelxs <- sub("(Pro␣regenda)␣", "", labelxs)
> labelxs <- sub("(Provincia␣)", "Provincia\n", labelxs)
> labelxs <- sub("(Professa␣)", "Professa\n", labelxs)
> labelxs <- sub("(Collegii␣)", "Collegii\n", labelxs)
> labelxs <- sub("(␣ad␣)", "\n␣ad␣", labelxs)
```

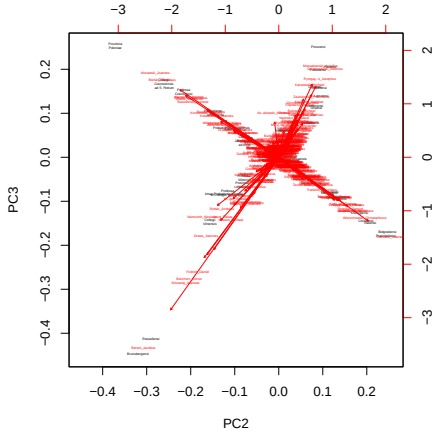

Рис. 11.6. Двойная диаграмма второго и третьего фактора

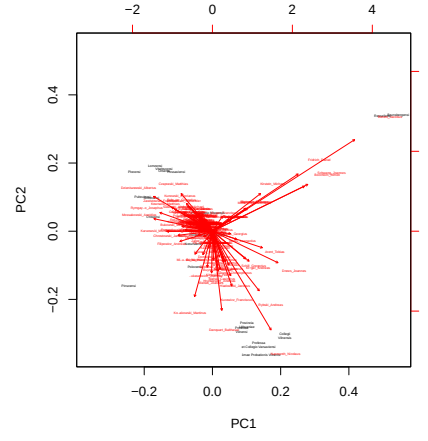

Рис. 11.7. Третий фактор в Литве

```
> biplot(prcomp(t(table(read.table("data_iesu.txt",
+                                  sep=",")))),
+        cex=c(0.25,0.25),
+        arrow.len = 0.025,
+        ylabs = 1:length(t(table(read.table("data_iesu.txt",
+                                  sep=",")))[1,]),
+        xlabs = labelxs,
+        )
```

Красным нанесены переменные, в которых оценивались должности. На основе вклада переменных и были выведены оси–факторы. Горизонтальная ось (1-й фактор) отражает наличие естественной группировки, следовательно это проявление ложного коэффициента корреляции. Этот корреляционный вклад вносят две несвязанные группы. Какие? По названием коллегиумов, заглянув в энциклопедию (`http://www.jezuici.krakow.pl/bibl/enc.htm`) обнаружим, что они расположены в разных провинциях: польской и литовской. Обе провинции находятся на территории одного федеративного государства Речи Посполитой, в котором было несколько административных центров: Краков, Варшава, Вильно. Вертикальная ось отражает иерархию существующую внутри провинций. Об этом свидетельствует концентрация в нижней части графика руководящих должностей в столичных городах: Варшаве, Вильно и Кракове. Таким образом выделенную ось мы можем с полным правом назвать «Столичность — Глубинка».

```
> biplot(prcomp(t(table(read.table("data_iesu.txt",
```

```
+                                  sep=","))))  ,
+       choices= c(2,3),
+       xlabs = labelxs,
+       arrow.len = 0.025,
+       cex=0.25)
```

Третий фактор показывает стоящую отдельно группу должностей Пруссии, входившей формально в Литовскую провинцию. Это хороший повод отдельно рассмотреть факторную структуру Литовской провинции (подмножество из файла с данными `data_iesu.txt`).

```
> labelxs <- gsub("_",
+               "␣",
+               rownames(t(table(read.table("data_litwa.csv",
+                                           sep=","))))
> labelxs <- sub("(Pro␣rectoratu)␣", "", labelxs)
> labelxs <- sub("(Pro␣Domo)␣", "", labelxs)
> labelxs <- sub("(tiae)␣", "tiae\n␣", labelxs)
> labelxs <- sub("(Pro␣regenda)␣", "", labelxs)
> labelxs <- sub("(Provincia␣)", "Provincia\n", labelxs)
> labelxs <- sub("(Professa␣)", "Professa\n", labelxs)
> labelxs <- sub("(Collegii␣)", "Collegii\n", labelxs)
> labelxs <- sub("(␣ad␣)", "\n␣ad␣", labelxs)


> biplot(prcomp(t(table(read.table("data_litwa.csv",
+                                  sep=",")))),
+       arrow.len = 0.025,
+       xlabs = labelxs,
+       cex=0.25)
```

Мы видим слева Пинск, а справа прусские Брунсберга и Решель. Что их отличает эти должности друг от друга? Национальная принадлежность кандидатов. Посмотрев на имена вы сразу же заметите явную концентрацию славянских фамилий (оканчание на -ски) слева, а иностранных (немецко звучащих) — справа.

Первый фактор явно отражает национальные различия как в пространстве признаков, так и в пространстве значений. Полюса оси — это «Прусскость — Белорусскость».

Второй фактор знакомая нам «Столичность — Провинциальность», отражающая иерархию в организационной структуре провинции. Данный переход подтверждает спектр фамилий руководящих кадров ордена. Слева видны польские фамилии, а справа не польские (чаще немецкие).

Эта же картина переносится на должности. В провинциальных городах немецкая часть спектра фамилий отсутствует вообще. столицах же присутствуют и немцы, и поляки, причём они расположены согласно оси первого фактора.

И что с того? Что мы имеем в сухом остатке. Что вскрыла такая обработка исторических сведений? Что собой представляет пространство факторов? Оно представляет собой пространство принятия кадровых решений (фрагмент соответствующей карты мира) в сознании провинциалов ордена и двух генералов ордена иезуитов в период с 1684 по 1705 год.

Что получили в результате историки? Фактическую информацию для проверки своих гипотез. Мы увидели иерархию о которой никто никогда явно не говорил, но все учитывали когда принимали решения. Естественно эту зависимость можно обнаружить и «методом пристального вглядывания». Статистика всего лишь инструмент, но в ряде случаев инструмент удобный.