Метод главных компонент

Алексей Померанцев

Содержание

1 Введение

1.1 Введение

В этом пособии рассказывается о методе главных компонент (*Principal Component Analysis*, PCA) – базовом подходе, применяемом в хемометрике для решения разнообразных задач. Текст ориентирован, прежде всего, на специалистов в области анализа экспериментальных данных: химиков, физиков, биологов, и т.д. Он может служить пособием для исследователей, начинающих изучение этого вопроса. Продолжить изучение вопроса можно с помощью указанной в заключении литературы

В пособии интенсивно используются понятия и методы матричной алгебры – вектор, матрица, и т.п. Читателям, которые плохо знакомы с этим аппаратом, рекомендуется изучить, или, хотя бы просмотреть, пособие *Матрицы и векторы*.

Изложение иллюстрируется примерами, выполненными в рабочей книге Excel, People.xls, которая сопровождает этот документ. Эта книга может работать без использования *Chemometrics Add-In*.

Примеры, приведенные в пособии, имеют абстрактный, модельный характер, однако, по сути, они тесно связаны с задачами, встречающимися на практике. Предполагается, что читатель имеет базовые навыки работы в среде Excel, умеет проводить простейшие матричные вычисления с использованием функций листа, таких как `MMULT`, `TREND`.

2 Базовые сведения

2.1 Данные

Метод главных компонент применяется к данным, записанным в виде матрицы **X** – прямоугольной таблицы чисел размерностью I строк и J столбцов.

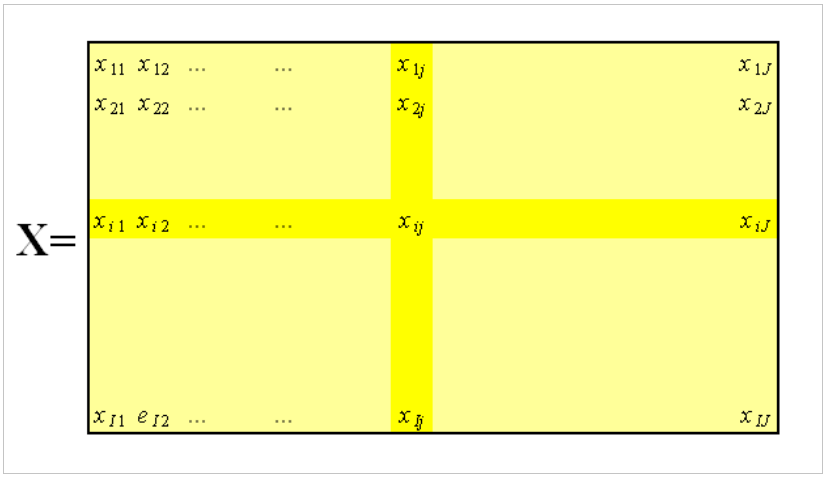


Рис. 2.1. Матрица данных

Традиционно строки этой матрицы называются *образцами*. Они нумеруются индексом i, меняющимся от 1 до I. Столбцы называются *переменными*, и они нумеруются индексом $j = 1, \ldots, J$.

Цель PCA – извлечение из этих данных нужной информации. Что является информацией, зависит от сути решаемой задачи. Данные могут содержать нужную нам информацию, они даже могут быть избыточными. Однако, в некоторых случаях, информации в данных может не быть совсем.

Размерность данных – число образцов и переменных – имеет большое значение для успешной добычи информации. Лишних данных не бывает – лучше, когда их много, чем мало. На практике это означает, что если получен спектр какого–то образца, то не нужно выбрасывать все точки, кроме нескольких характерных длин волн, а использовать их все, или, по крайней мере, значительный кусок.

Данные всегда (или почти всегда) содержат в себе нежелательную составляющую, называемую шумом. Природа этого шума может быть различной, но, во многих случаях, шум – это та часть данных, которая

не содержит искомой информации. Что считать шумом, а что – информацией, всегда решается с учетом поставленных целей и методов, используемых для ее достижения.

Шум и избыточность в данных обязательно проявляют себя через корреляционные связи между переменными. Погрешности в данных могут привести к появлению не систематических, а случайных связей между переменными. Понятие эффективного (химического) ранга и скрытых, латентных переменных, число которых равно этому рангу, является важнейшим понятием в PCA

2.2 Интуитивный подход

Постараемся передать суть метода главных компонент, используя интуитивно–понятную геометрическую интерпретацию. Начнем с простейшего случая, когда имеются только две переменные x_1 и x_2. Такие данные легко изобразить на плоскости (Рис. 2.2).

	X_1	X_2
1	0.407	0.353
2	0.475	0.355
3	-0.088	-0.045
4	0.394	0.325
5	0.274	0.202
6	0.131	0.258
7	-0.053	-0.031
8	-0.124	-0.128
9	-0.469	-0.344
10	0.088	0.171
11	-0.261	-0.162
12	0.401	0.341
13	-0.376	-0.143
14	-0.251	-0.255
15	-0.325	-0.316
16	0.464	0.248
17	-0.310	-0.207
18	0.307	0.247
19	-0.399	-0.303
20	-0.253	-0.253
21	-0.341	-0.291

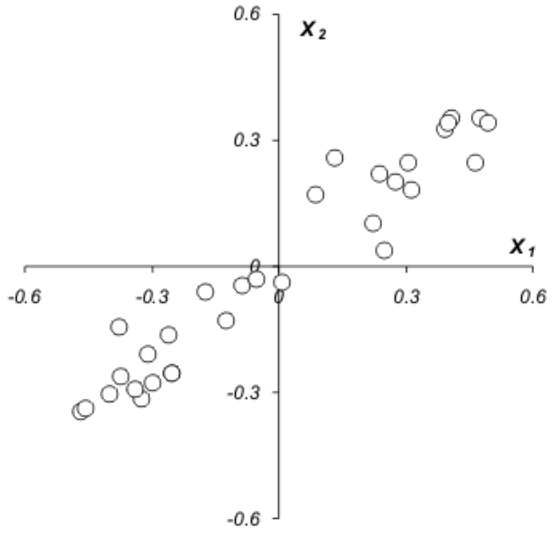


Рис. 2.2. Графическое представление двумерных данных

Каждой строке исходной таблицы (т.е. образцу) соответствует точка на плоскости с соответствующими координатами. Они обозначены пустыми кружками на Рис. 2.2. Проведем через них прямую, так, чтобы вдоль нее происходило максимальное изменение данных. На рисунке эта прямая выделена синим цветом; она называется первой главной компонентой – *PC1*. Затем спроецируем все исходные точки на эту ось. Получившиеся точки закрашены красным цветом. Теперь мы можем предположить, что на самом деле все наши экспериментальные точки и должны были лежать на этой новой оси. Просто какие–то неведомые силы отклонили их от правильного, идеального положения, а мы вернули их на место. Тогда все отклонения от новой оси можно считать шумом, т.е. ненужной нам информацией. Правда, мы должны быть в этом уверены. Проверить шум ли это, или все еще важная часть данных, можно поступив с этими остатками так же, как мы поступили с исходными данными – найти в них ось максимальных изменений.

Она называется второй главной компонентой (*PC2*). И так надо действовать, до тех пор, пока шум уже не станет действительно шумом, т.е. случайным хаотическим набором величин.

В общем, многомерном случае, процесс выделения главных компонент происходит так:

1. Ищется центр облака данных, и туда переносится новое начало координат – это нулевая главная компонента (PC0)
2. Выбирается направление максимального изменения данных – это первая главная компонента (PC1)
3. Если данные описаны не полностью (шум велик), то выбирается еще одно направление (PC2) – перпендикулярное к первому, так чтобы описать оставшееся изменение в данных и т.д.

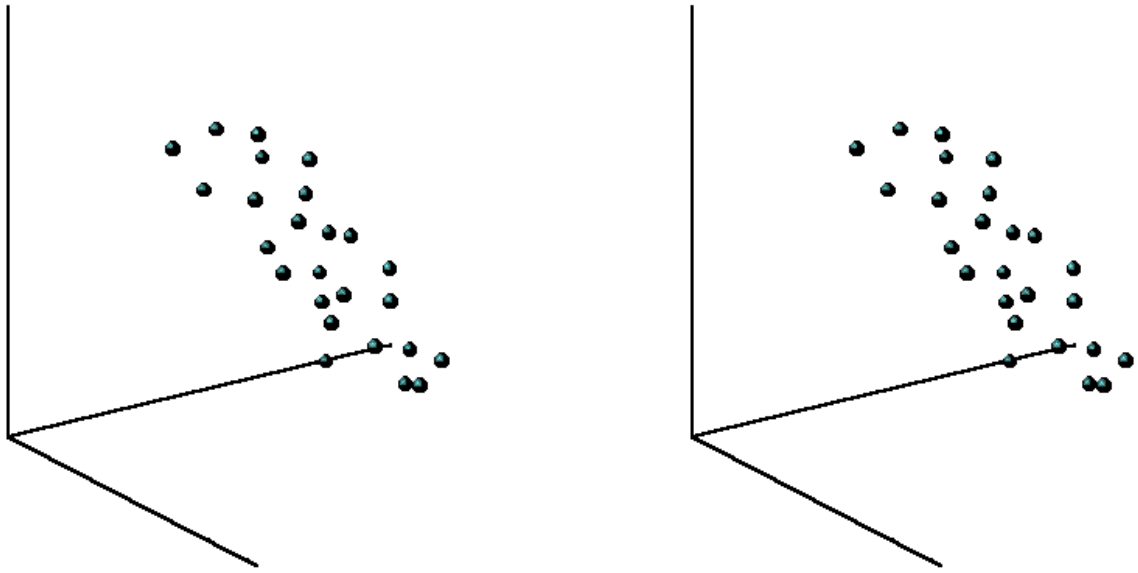


Рис. 2.3. Графическое представление метода главных компонент

В результате, мы переходим от большого количества переменных к новому представлению, размерность которого значительно меньше. Часто удается упростить данные на порядки: от 1000 переменных перейти всего к двум. При этом ничего не выбрасывается – все переменные учитываются. В то же время несущественная для сути дела часть данных отделяется, превращается в шум. Найденные главные компоненты и дают нам искомые скрытые переменные, управляющие устройством данных.

2.3 Понижение размерности

Суть метода главных компонент – это существенное понижение размерности данных. Исходная матрица **X** заменяется двумя новыми матрицами **T** и **P**, размерность которых, A, меньше, чем число переменных (столбцов) J у исходной матрицы **X**.

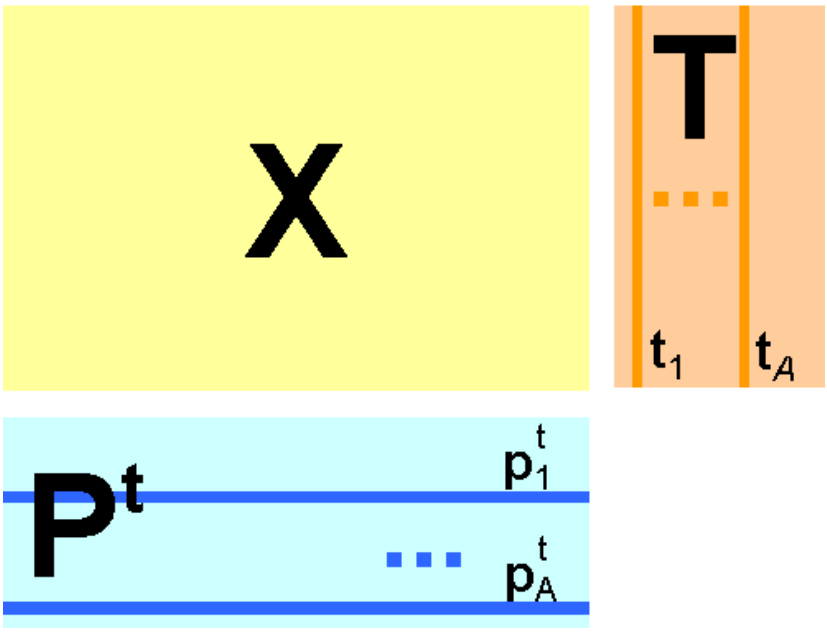


Рис. 2.4. Декомпозиция матрицы **X**

Вторая размерность – число образцов (строк) I сохраняется. Если декомпозиция выполнена правильно – размерность A выбрана верно, то матрица **T** несет в себе столько же информации, сколько ее было в начале, в матрице **X**. При этом матрица **T** меньше, и, стало быть, проще, чем **X**.

3 Метод главных компонент

3.1 Формальное описание

Пусть имеется матрица переменных **X** размерностью $(I \times J)$, где I – число образцов (строк), а J – это число независимых переменных (столбцов), которых, как правило, много ($J >> 1$). В методе главных компонент используются новые, формальные переменные $\mathbf{t}_a$ ($a = 1, \ldots, A$), являющиеся линейной комбинацией исходных переменных $\mathbf{x}_j$ ($j = 1, \ldots, J$):

$$\mathbf{t}_a = \mathbf{p}_{a1}\mathbf{x}_1 + \cdots + \mathbf{p}_{aJ}\mathbf{x}_J$$

С помощью этих новых переменных матрица X разлагается в произведение двух матриц **T** и **P**:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{t}} + \mathbf{E}$$

Матрица **T** называется *матрицей счетов* (scores). Ее размерность $(I \times A)$.

Матрица **P** называется *матрицей нагрузок* (loadings). Ее размерность $(J \times A)$.

E – это *матрица остатков*, размерностью $(I \times J)$.

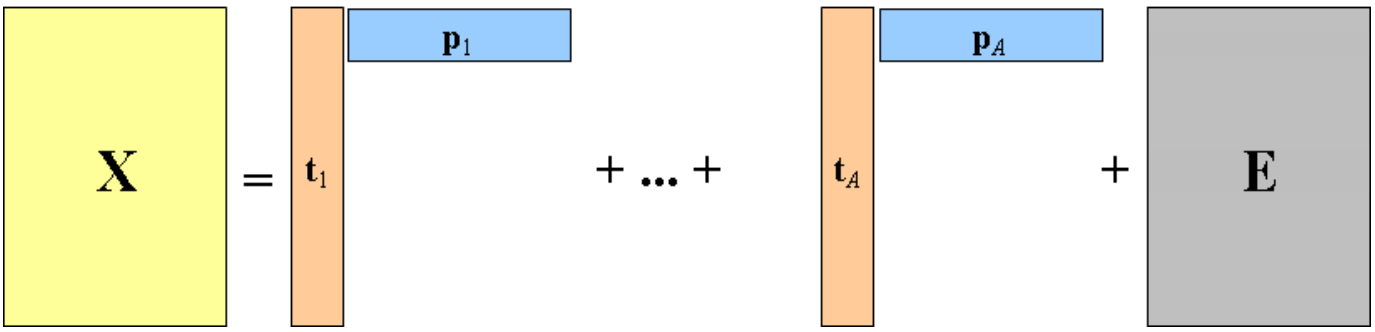


Рис. 3.1. Разложение по главным компонентам

Новые переменные $\mathbf{t}_a$ называются *главными компонентами* (Principal Components), поэтому и сам метод называется *методом главных компонент* (PCA). Число столбцов $\mathbf{t}_a$ в матрице **T**, и $\mathbf{p}_a$ в матрице **P**, равно A,

которое называется *числом главных компонент* (PC). Эта величина заведомо меньше числа переменных J и числа образцов I.

Важным свойством PCA является ортогональность (независимость) главных компонент. Поэтому матрица счетов $\mathbf{T}$ не перестраивается при увеличении числа компонент, а к ней просто прибавляется еще один столбец – соответствующий новому направлению. Тоже происходит и с матрицей нагрузок $\mathbf{P}$.

3.2 Алгоритм NIPALS

Чаще всего для построения PCA счетов и нагрузок, используется рекуррентный алгоритм *NIPALS*, который на каждом шагу вычисляет одну компоненту. Сначала исходная матрица $\mathbf{X}$ преобразуется (как минимум – центрируется) и превращается в матрицу $\mathbf{E}_0$, $a = 0$. Далее применяют следующий алгоритм.

1. Выбрать начальный вектор $\mathbf{t}$
2. Расчитать вектор нагрузок: $\mathbf{p}^{\mathrm{t}} = \mathbf{t}^{\mathrm{t}}\mathbf{E}_a / \mathbf{t}^{\mathrm{t}}\mathbf{t}$
3. Сделать нормировку: $\mathbf{p} = \mathbf{p}/(\mathbf{p}^{\mathrm{t}}\mathbf{p})^{\frac{1}{2}}$
4. Расчитать вектор счетов: $\mathbf{t} = \mathbf{E}_a\mathbf{p}/\mathbf{p}^{\mathrm{t}}\mathbf{p}$
5. Проверить сходимость, если нет, то идти на 2

После вычисления очередной (a-ой) компоненты, полагаем $\mathbf{t}_a = \mathbf{t}$ и $\mathbf{p}_a = \mathbf{p}$. Для получения следующей компоненты надо вычислить остатки $\mathbf{E}_{a+1} = \mathbf{E}_a - \mathbf{t}\mathbf{p}^{\mathrm{t}}$ и применить к ним тот же алгоритм, заменив индекс a на $a + 1$.

В этом пособии для построения PCA используется специальная надстройка для программы Excel (Add–In) Chemometrics.xla. Она дополняет список стандартных функций Excel и позволяет проводить PCA разложение на листах рабочей книги.

После того, как построено пространство из главных компонент, новые образцы $\mathbf{X}_{new}$ могут быть на него спроецированы, иными словами – определены матрицы их счетов $\mathbf{T}_{new}$. В методе PCA это делается очень просто

$$\mathbf{T}_{new} = \mathbf{X}_{new}\mathbf{P}$$

3.3 PCA и SVD

Метод главных компонент тесно связан с другим разложением – по сингулярным значениям, SVD. В последнем случае исходная матрица $\mathbf{X}$ разлагается в произведение трех матриц

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{t}}$$

Здесь $\mathbf{U}$ – матрица, образованная ортонормированными собственными векторами $\mathbf{u}_r$ матрицы $\mathbf{X}\mathbf{X}^{\mathrm{t}}$, соответствующим значениям λ_r:

$$\mathbf{X}\mathbf{X}^{\mathrm{t}}\mathbf{u}_r = \lambda_r \mathbf{u}_r$$

$\mathbf{V}$ – матрица, образованная ортонормированными собственными векторами $\mathbf{v}_r$ матрицы $\mathbf{X}^{\mathrm{t}}\mathbf{X}$:

$$\mathbf{X}^{\mathrm{t}}\mathbf{X}\mathbf{v}_r = \lambda_r \mathbf{v}_r$$

$\mathbf{S}$ – положительно определенная диагональная матрица, элементами которой являются сингулярные значения $\sigma_1 \geq \cdots \geq \sigma_R \geq 0$ равные квадратным корням из собственных значений λ_r: $\sigma_r = \sqrt{\lambda_r}$.

Связь между PCA и SVD определяется следующими простыми соотношениями

$$\mathbf{T} = \mathbf{U}\mathbf{S}$$

$$\mathbf{P} = \mathbf{V}$$

3.4 Счета

Матрица счетов $\mathbf{T}$ дает нам проекции исходных образцов (J–мерных векторов $\mathbf{x}_1, \ldots, \mathbf{x}_I$) на подпространство главных компонент (A-мерное). Строки $\mathbf{t}_1, \ldots, \mathbf{t}_I$ матрицы $\mathbf{T}$ – это координаты образцов в новой системе координат. Столбцы $\mathbf{t}_1, \ldots, \mathbf{t}_A$ матрицы $\mathbf{T}$ – ортогональны и представляют проекции всех образцов на одну новую координатную ось.

При исследовании данных методом PCA, особое внимание уделяется графикам счетов. Они несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов каждый образец изображается в координатах $(\mathbf{t}_i, \mathbf{t}_j)$, чаще всего – $(\mathbf{t}_1, \mathbf{t}_2)$, обозначаемых *PC1* и *PC2*. Близость двух точек означает их схожесть, т.е. положительную корреляцию. Точки, расположенные под прямым углом, являются некоррелированными, а расположенные диаметрально противоположно – имеют отрицательную корреляцию.

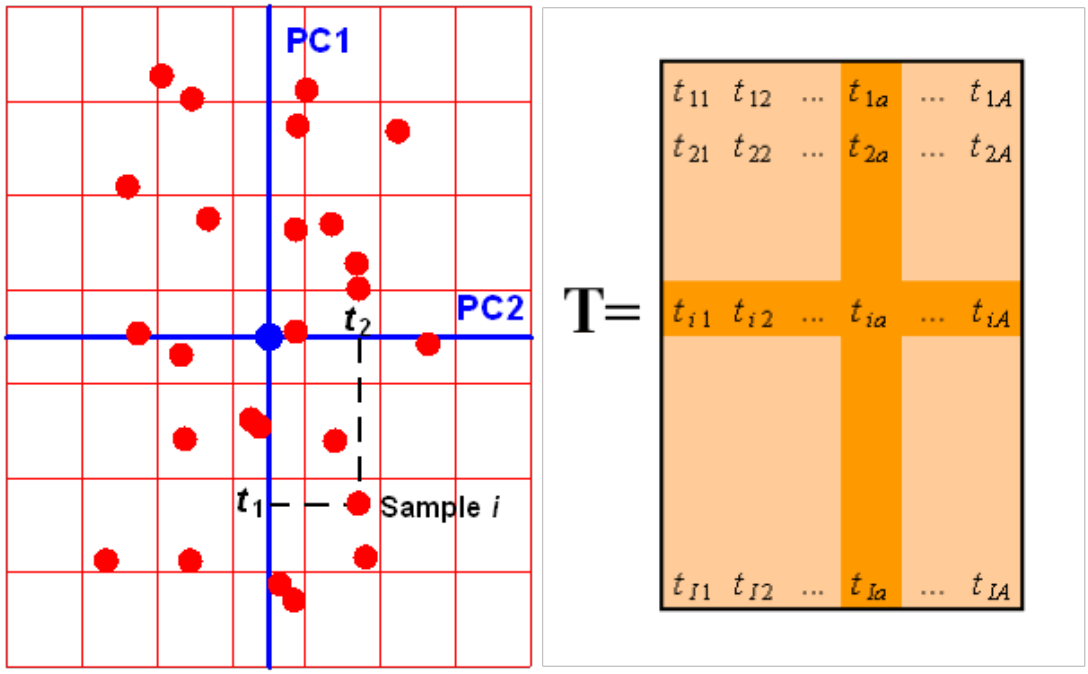


Рис. 3.2. График счетов

Подробнее о том, как из графиков счетов извлекается полезная информация, будет рассказано в примере.

Для матрицы счетов имеют место следующие соотношения:

$$\mathbf{T}^t\mathbf{T} = \mathbf{\Lambda} = \text{diag}\lambda_1, \ldots, \lambda_A$$

где величины $\lambda_1 \geq \cdots \geq \lambda_A \geq 0$ – это собственные значения. Они характеризуют важность каждой компоненты

Нулевое собственное значение λ_0 определяется как сумма всех собственных значений, т.е.

Для вычисления PCA-счетов в надстройке Chemometrics Add-In используется функция `ScoresPCA`, которая имеет следующий синтаксис:

```
ScoresPCA (X [, PC] [,CentWeightX] [, Xnew])
```

- `X` – множество значений **X** (калибровочный набор)
- `PC` – необязательный аргумент. Целая величина, определяющая число главных компонент (A), используемых в методе главных компонент.
- `CentWeightX` – необязательный аргумент. Целая величина, определяющая проводится ли центрирование и/или шкалирование X переменных.

- `Xnew` – необязательный аргумент. Множество новых значений $\mathbf{X}_{new}$ (проверочный набор) для которых вычисляются и выводятся значения счетов **T**.

Матрица новых значений `Xnew` должна иметь столько же столбцов (переменных), как и матрица `X`. Если аргумент `Xnew` опущен, то предполагается, что он совпадает с массивом `X`. Результат является массивом (матрицей), в которой число строк равно числу строк (образцов, I) в массиве `Xnew`, а число столбцов равно числу ГК (A).

Аналогичная стандартная функция листа: `ТЕНДЕНЦИЯ`.

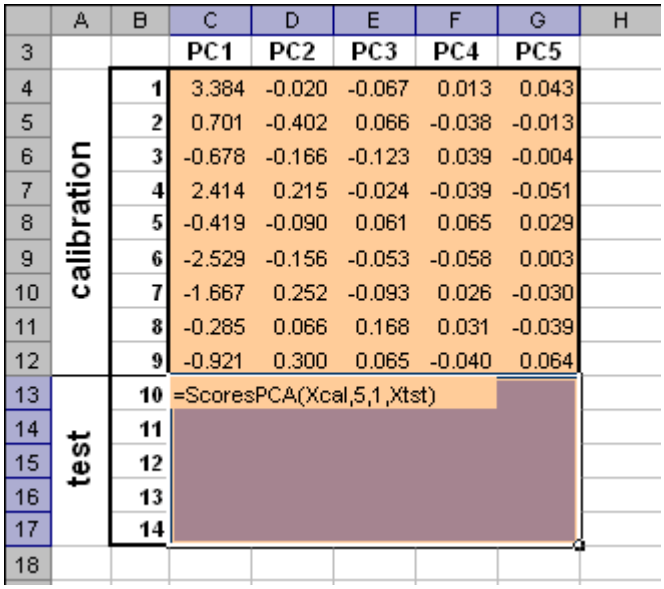


Рис. 3.3. Пример ввода функции ScoresPCA

Функция `ScoresPCA` является формулой массива и ее ввод должен завершаться нажатием комбинации `CTRL+SHIFT+ENTER`.

3.5 Нагрузки

Матрица нагрузок P – это матрица перехода из исходного пространства переменных $\mathbf{x}_1, \ldots, \mathbf{x}_J$ (J-мерного) в пространство главных компонент (A-мерное). Каждая строка матрицы **P** состоит из коэффициентов, связывающих переменные **t** и **x**. Например, a-я строка – это проекция всех переменных $\mathbf{x}_1, \ldots, \mathbf{x}_J$ на a-ю ось главных компонент. Каждый столбец **P** – это проекция соответствующей переменной $\mathbf{x}_j$ на новую систему координат.

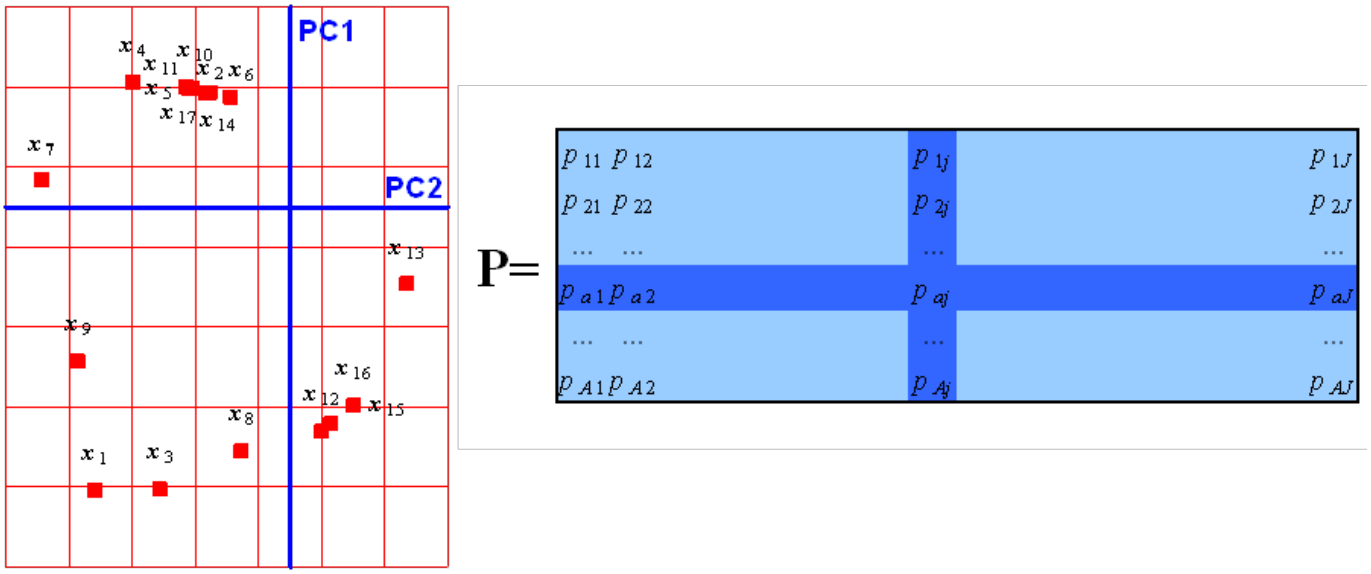


Рис. 3.4. График нагрузок

График нагрузок применяется для исследования роли переменных. На этом графике каждая переменная $\mathbf{x}_j$ отображается точкой в координатах $(\mathbf{p}_i, \mathbf{p}_j)$, например $(\mathbf{p}_1, \mathbf{p}_2)$. Анализируя его аналогично графику счетов, можно понять, какие переменные связаны, а какие независимы. Совместное исследование парных графиков счетов и нагрузок, также может дать много полезной информации о данных.

В методе главных компонент нагрузки – это ортогональные нормированные вектора, т.е.

$$\mathbf{P}^t\mathbf{P} = \mathbf{I}$$

Для вычисления PCA-нагрузок в надстройке Chemometrics Add-In используется функция `LoadingsPCA`:

```
LoadingsPCA (X [, PC] [, CentWeightX])
```

- `X` – множество значений `X` (калибровочный набор)
- `PC` – необязательный аргумент. Целая величина, определяющая число главных компонент (A), используемых в методе главных компонент.
- `CentWeightX` – необязательный аргумент. Целая величина, определяющая проводится ли центрирование и/или шкалирование `X` переменных.

Результат является массивом (матрицей), в которой число столбцов равно числу ГК (A), а число строк равно числу столбцов (переменных, J) в массиве `X`. Аналогичная стандартная функция листа: `МОБР`

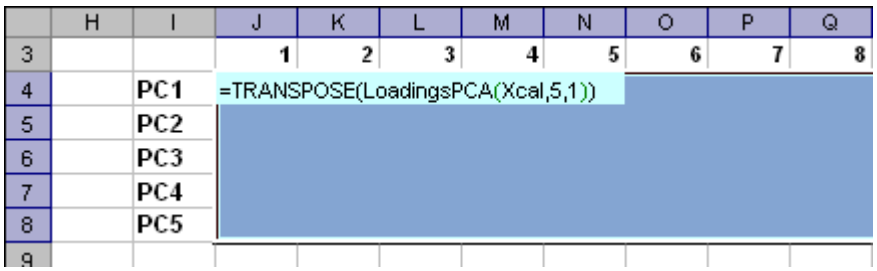


Рис. 3.5. Пример ввода функции LoadingsPCA

Функция `LoadingsPCA` является формулой массива и ее ввод должен завершаться нажатием комбинации `CTRL+SHIFT+ENTER`.

3.6 Данные специального вида

Результат моделирования методом главных компонент не зависит от порядка, в котором следуют образцы и/или переменные. Иными словами строки и столбцы в исходной матрице **X** можно переставить, но ничего принципиально не изменится. Однако, в некоторых случаях, сохранять и отслеживать этот порядок очень полезно – это позволяет лучше понять устройство моделируемых данных.

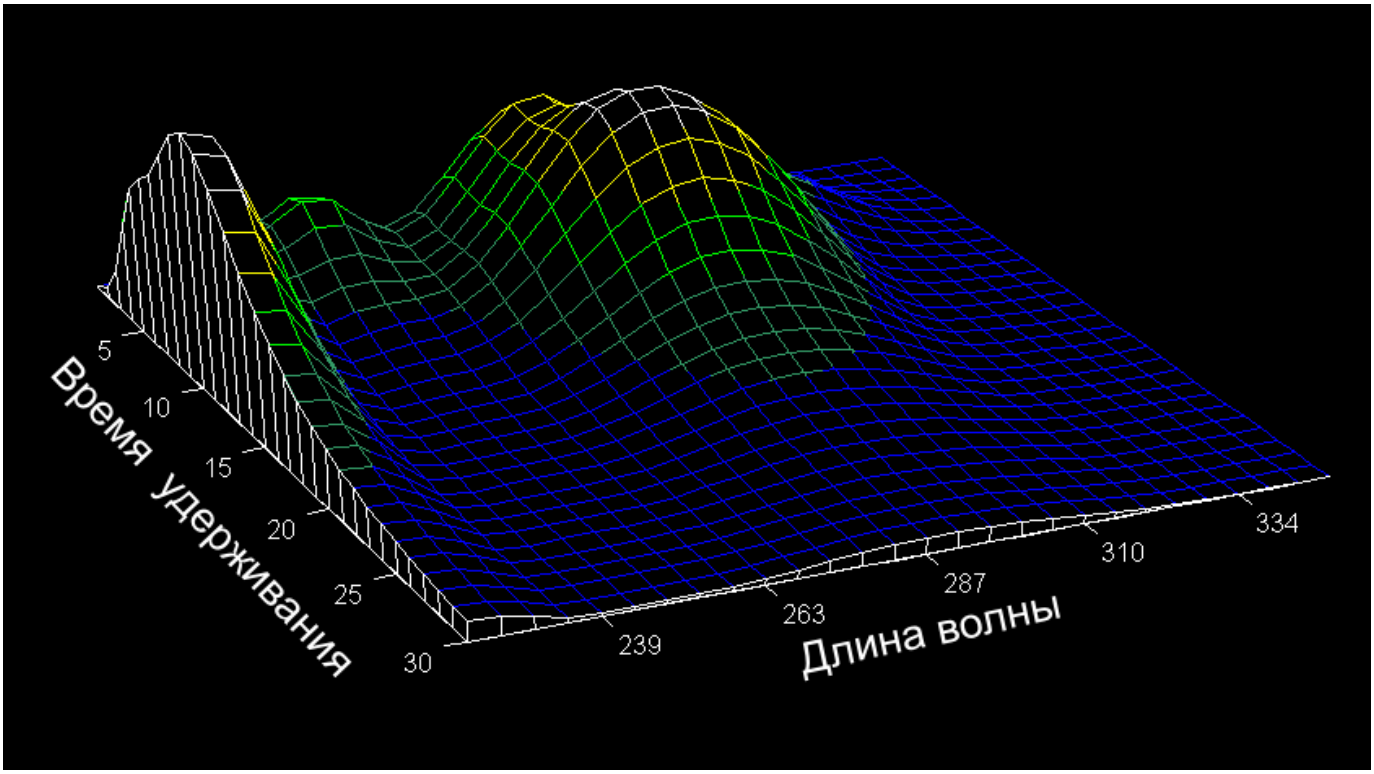


Рис. 3.6. Данные ВЭЖХ–ДДМ

Рассмотрим простой пример – моделирование данных, полученных методом высокоэффективной

жидкостной хроматографией с детектированием на диодной матрице (ВЭЖХ–ДДМ). Данные представляются матрицей, размерностью 30 образцов (I) на 28 переменных (J). Образцы соответствуют временам удерживания от 0 до 30 с, а переменные – длинам волн от 220 до 350 нм, на которых происходит детектирование. Данные ВЭЖХ–ДДМ представлены на Рис 3.6.

Эти данные хорошо моделируются методом PCA с двумя главными компонентами. Ясно, что в этом примере нам важен порядок, в котором идут образцы и переменные – он задается естественным ходом времени и спектральным диапазоном. Полученные счета и нагрузки полезно изобразить на графиках в зависимости от соответствующего параметра – счета от времени, а нагрузки от длины волны. (см. Рис 3.7)

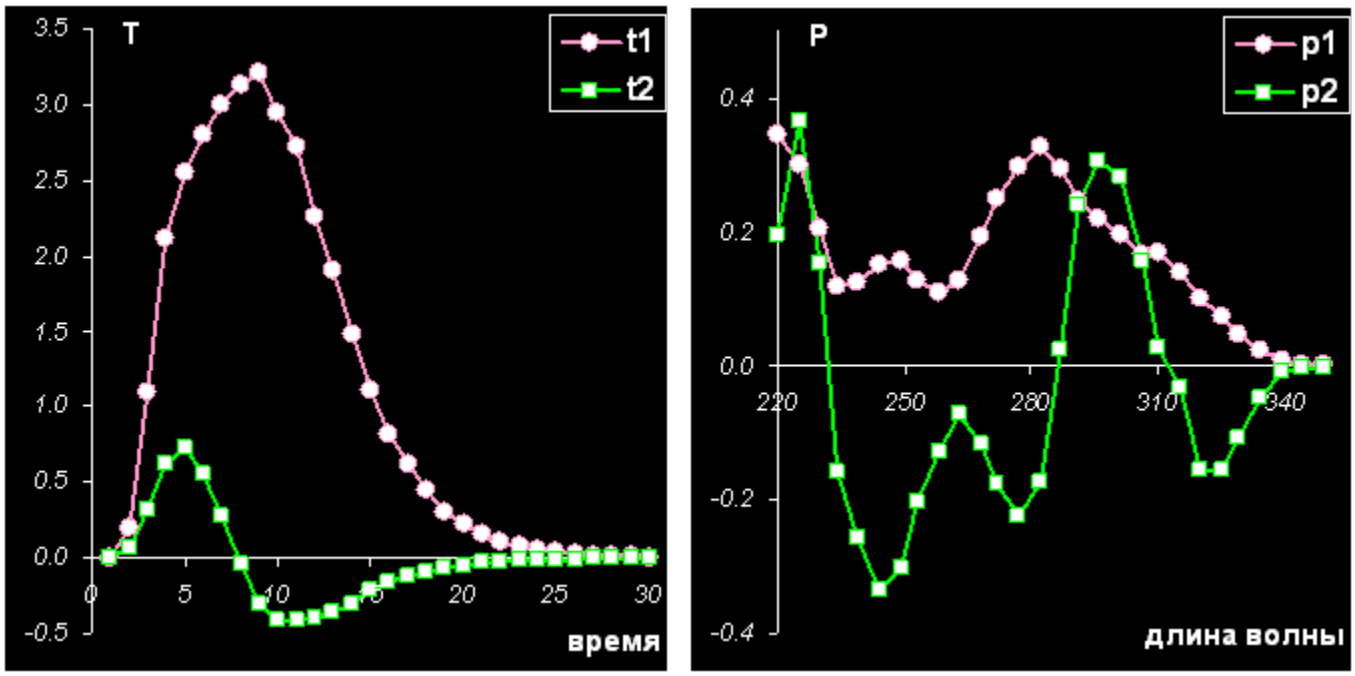


Рис. 3.7. Графики счетов и нагрузок для данных ВЭЖХ–ДДМ

Подробнее этот пример разобран в пособии Разрешение многомерных кривых.

3.7 Погрешности

PCA декомпозиция матрицы $\mathbf{X}$ является последовательным, итеративным процессом, который можно оборвать на любом шаге $a = A$. Получившаяся матрица

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^{t}$$

вообще говоря, отличается от матрицы $\mathbf{X}$. Разница между ними

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}}$$

называется *матрицей остатков.*

Рассмотрим геометрическую интерпретацию остатков. Каждый исходный образец $\mathbf{x}_i$ (строка в матрице $\mathbf{X}$) можно представить как вектор в J–мерном пространстве с координатами

$$\mathbf{x}_i = (x_{i1}, x_{i1}, \dots, x_{iJ})$$

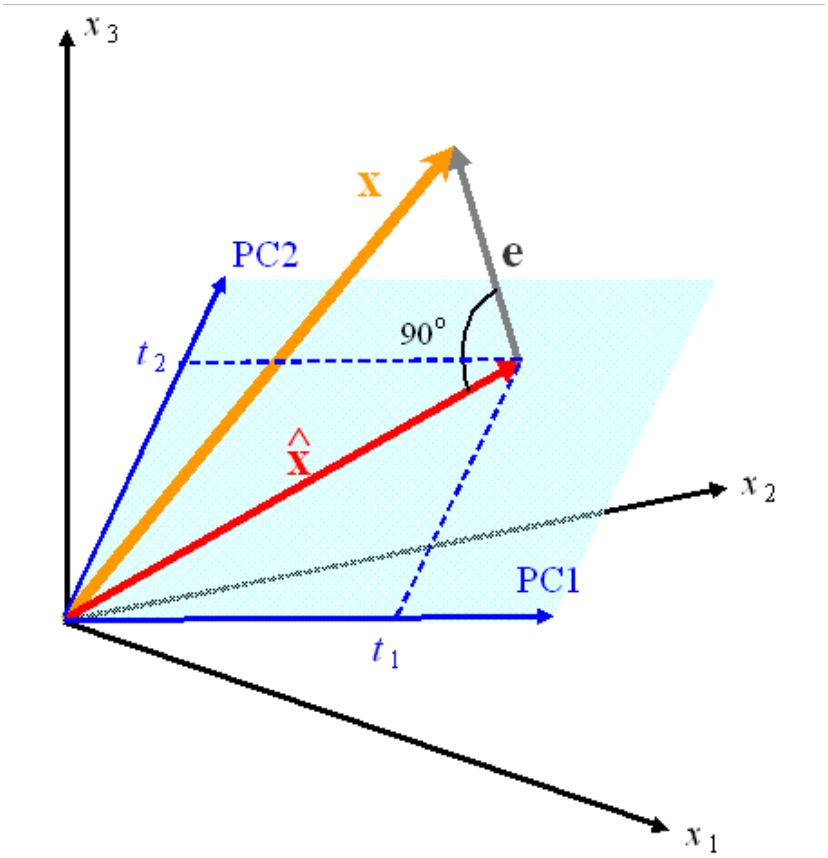


Рис. 3.8. Геометрия PCA

PCA проецирует его в вектор, лежащий в пространстве главных компонент, $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{iA})$ размерностью A. В исходном пространстве этот же вектор $\mathbf{t}_i$ имеет координаты

$$\hat{\mathbf{x}}_i = (\hat{x}_{i1}, \hat{x}_{i1}, \dots, \hat{x}_{iJ})$$

Разница между исходным вектором и его проекцией является вектором остатков

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}_i = (e_{i1}, e_{i1}, \dots, e_{iJ})$$

Он образует i–ю строку в матрице остатков **E**.

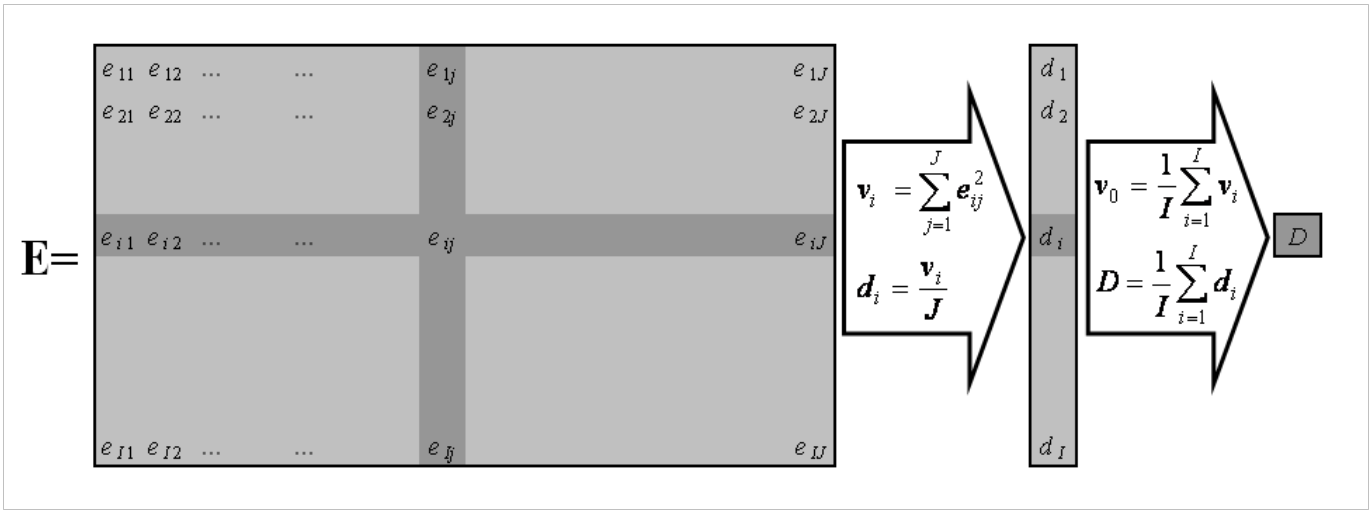


Рис. 3.9. Вычисление остатков

Исследуя остатки можно понять, как устроены данные и хорошо ли они описываются PCA моделью.

Величина –

$$v_i = \sum_{j=1}^{J} e_{ij}^2$$

определяет квадрат отклонения исходного вектора $\mathbf{x}_i$ от его проекции на пространство PC. Чем оно меньше, тем лучше приближается i–ый образец. Для вычисления отклонений можно использовать стандартные функции листа или специальную пользовательскую функцию.

Эта же величина, деленная на число переменных

$$d_i = \frac{1}{J}\sum_{j=1}^{J} e_{ij}^2$$

дает оценку дисперсии (вариации) i–го образца.

Среднее (для всех образцов) расстояние v_0 вычисляется как

$$v_0 = \frac{1}{I} \sum_{i=1}^{I} v_i$$

Оценка общая (для всех образцов) дисперсии вычисляется так –

$$D = \frac{1}{I} \sum_{i=1}^{I} d_i$$

3.8 Проверка

В случае, когда PCA модель предназначена для предсказания или для классификации, а не для простого исследования данных, такая модель нуждается в подтверждении (валидации). При проверке методом тест–валидации исходный массив данных состоит из двух независимо полученных наборов, каждый из которых является достаточно представительным. Первый набор, называемый обучающим, используется для моделирования. Второй набор, называемый проверочным, служит только для проверки модели. Построенная модель применяется к данным из проверочного набора, и полученные результаты сравниваются с проверочными значениями. Таким образом принимается решение о правильности, точности моделирования.

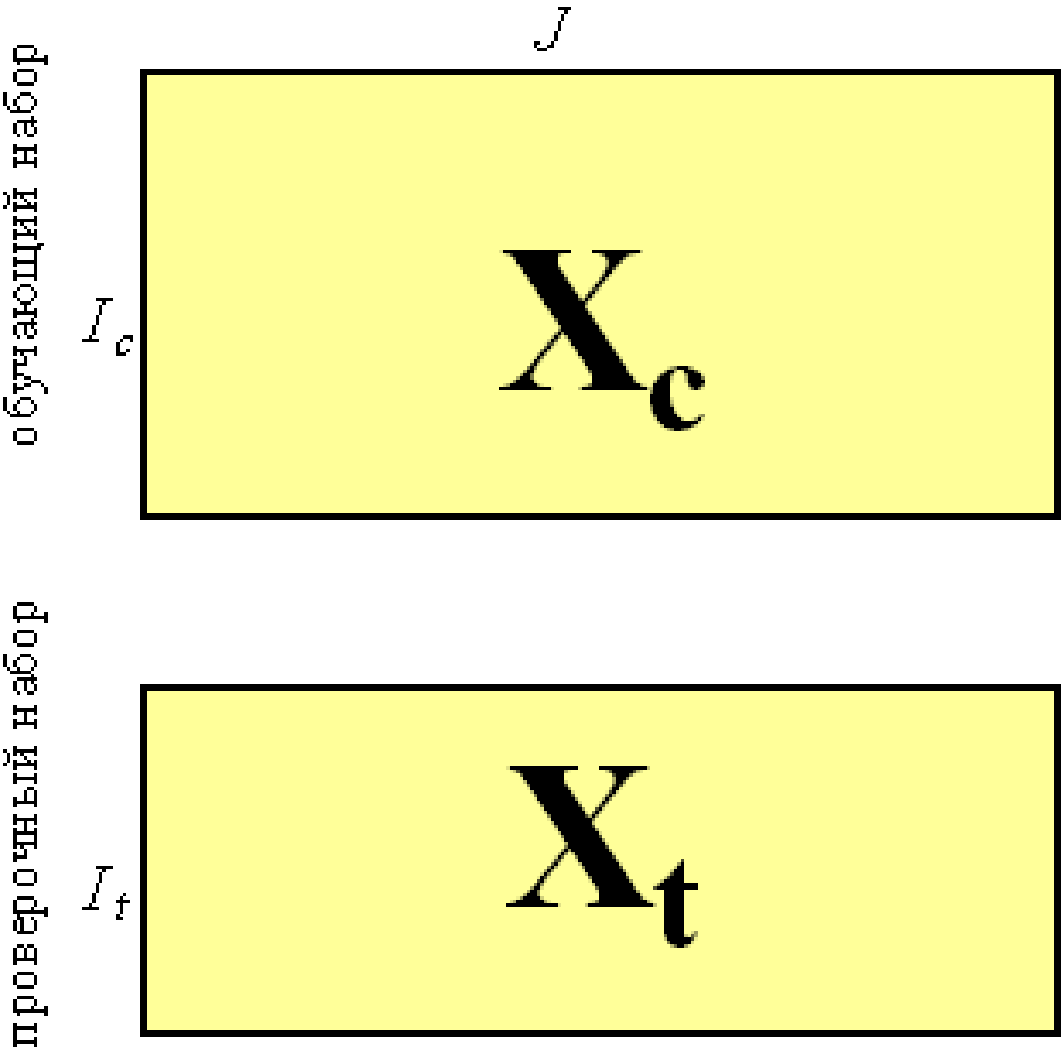


Рис. 3.10. Обучающий и проверочный наборы

В некоторых случаях объем данных слишком мал для такой проверки. Тогда применяют другой метод – перекрестной проверки (кросс–валидация). Используется также проверка методом коррекции размахом, суть которой предлагается изучить самостоятельно.

3.9 “Качество” декомпозиции

Результатом PCA моделирования являются величины $\hat{\mathbf{X}}_c$ – оценки, найденные по модели, построенной на обучающем наборе $\mathbf{X}_c$. Результатом проверки служат величины $\hat{\mathbf{X}}_t$ – оценки проверочных значений $\mathbf{X}_t$, вычисленные по той же модели, но как новые образцы. Отклонение оценки от проверочного значения вычисляют как матрицу остатков:

в обучении: $\mathbf{E}_c = \mathbf{X}_c - \hat{\mathbf{X}}_c$

и в проверке: $\mathbf{E}_t = \mathbf{X}_t - \hat{\mathbf{X}}_t$

Следующие величины характеризуют “качество” моделирования в среднем.

Полная дисперсия остатков в обучении ($TRVC$) и в проверке ($TRVP$) –

$$TRVC = \frac{1}{I_c J} \sum_{i=1}^{I_c} \sum_{j=1}^{J} e_{ij}^2 = D_c$$

$$TRVP = \frac{1}{I_t J} \sum_{i=1}^{I_t} \sum_{j=1}^{J} e_{ij}^2 = D_t$$

Полная дисперсия выражается в тех же единицах (точнее их квадратах), что и исходные величины X.

Объясненная дисперсия остатков в обучении ($ERVC$) и в проверке ($ERVP$)

$$ERVC = 1 - \sum_{i=1}^{I_c} \sum_{j=1}^{J} e_{ij}^2 \Big/ \sum_{i=1}^{I_c} \sum_{j=1}^{J} x_{ij}^2$$

$$ERVP = 1 - \sum_{i=1}^{I_t} \sum_{j=1}^{J} e_{ij}^2 \Big/ \sum_{i=1}^{I_t} \sum_{j=1}^{J} x_{ij}^2$$

Объясненная дисперсия – это относительная величина. При ее вычислении используется естественная нормировка – сумма квадратов всех исходных величин x_{ij}. Обычно она выражается в процентах или в долях единицы. Во всех этих формулах величины e_{ij} – это элементы матриц $\mathbf{E}_c$ или $\mathbf{E}_t$. Для характеристик, наименование которых оканчивается на C (например, $TRVC$), используется матрица $\mathbf{E}_c$ (обучение), а для тех, которые оканчиваются на P (например, $TRVP$), берется матрица $\mathbf{E}_t$ (проверка).

3.10 Выбор числа главных компонент

Как уже отмечалось выше, метод главных компонент – это итерационная процедура, в которой новые компоненты добавляются последовательно, одна за другой. Важно знать, когда остановить этот процесс, т.е. как определить правильное число главных компонент, A. Если это число слишком мало, то описание данных будет не полным. С другой стороны, избыточное число главных компонент приводит к переоценке, т.е. к ситуации, когда моделируется шум, а не содержательная информация.

Для выбора значения числа главных компонент обычно используется график, на котором объясненная дисперсия (ERV) изображается в зависимости от числа PC. Пример такого графика приведен на Рис. 3.11.

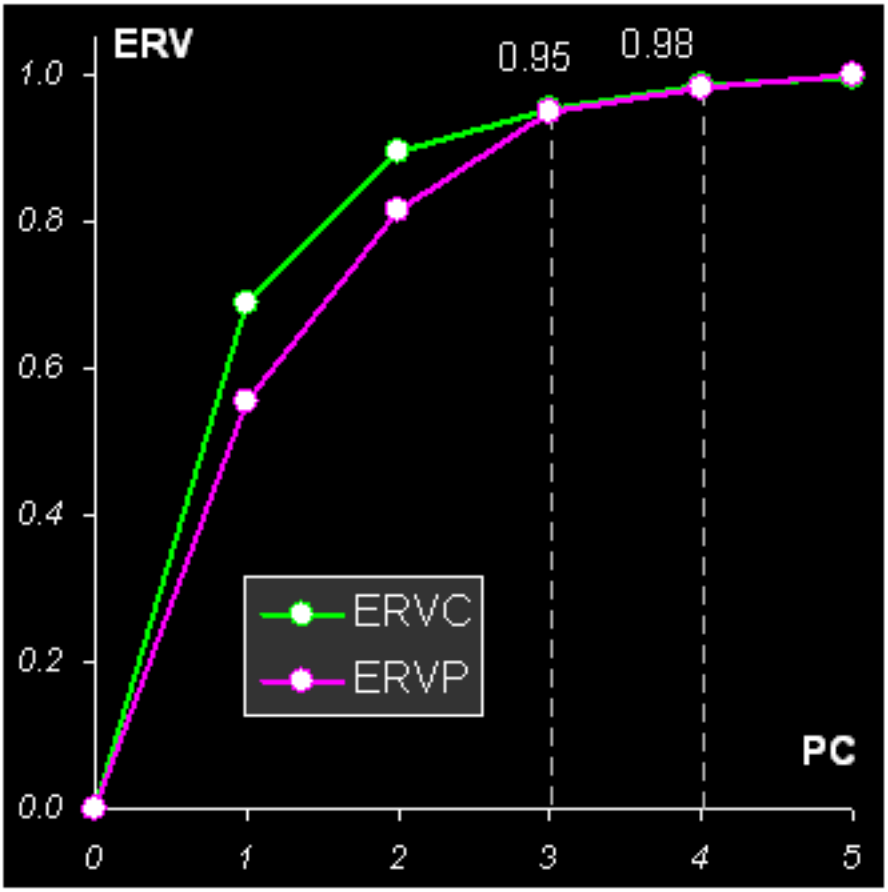


Рис. 3.11. Выбор числа PC

Из этого графика видно, что правильное число PC – это 3 или 4. Три компоненты объясняют 95%, а четыре 98% исходной вариации. Окончательное решение о величине A можно принять только после содержательного анализа данных.

Другим полезным инструментом является график, на котором изображаются собственные значения в зависимости от числа PC. Пример показан на Рис. 3.12.

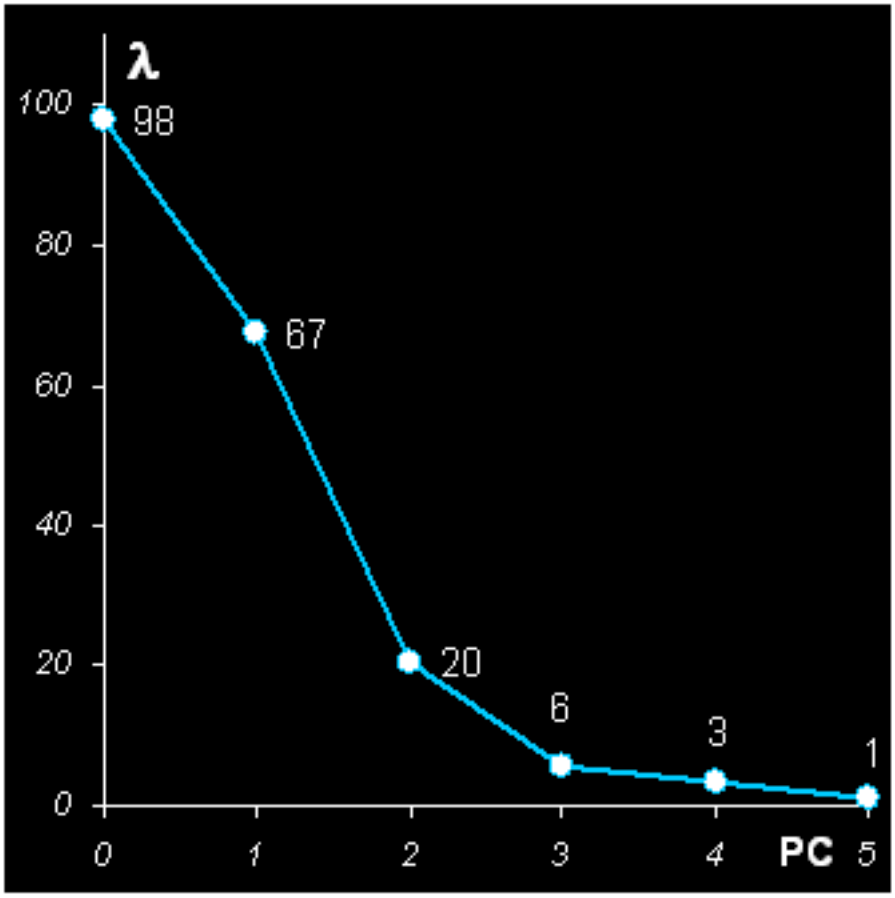


Рис. 3.12. График собственных значений

Из этого рисунка опять видно, что для $a = 3$ происходит резкое изменение формы графика – излом. Поэтому верное число PC – это три или четыре.

3.11 Неединственность PCA

Разложение по методу главных компонент

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E}$$

не является единственным. Вместо матриц $\mathbf{T}$ и $\mathbf{P}$ можно использовать другие матрицы $\tilde{\mathbf{T}}$ и $\tilde{\mathbf{P}}$, которые дадут аналогичную декомпозицию:

$$\mathbf{X} = \tilde{\mathbf{T}}\tilde{\mathbf{P}}^{\mathrm{t}} + \mathbf{E}$$

с той же матрицей ошибок **E**. Простейший пример – это одновременное изменение знаков у соответствующих компонент векторов $\mathbf{t}_a$ и $\mathbf{p}_a$, при котором произведение

$$\tilde{\mathbf{t}}_a\tilde{\mathbf{p}}_a^{\mathrm{t}} = \mathbf{t}_a\mathbf{p}_a^{\mathrm{t}}$$

остается неизменным. Алгоритм *NIPALS* дает именно такой результат – с точностью до знака, поэтому его реализация в разных программах может приводить к расхождениям в направлениях главных компонент.

Более сложный случай – это одновременное вращение матриц **T** и **P**. Пусть **R** – это ортогональная матрица вращения размерностью $A \times A$, т.е такая матрица, что $\mathbf{R}^{\mathrm{t}} = \mathbf{R}^{-1}$. Тогда

$$\mathbf{T}\mathbf{P}^{\mathrm{t}} = \mathbf{T}\mathbf{R}\mathbf{R}^{-1}\mathbf{P}^{\mathrm{t}} = (\mathbf{T}\mathbf{R})(\mathbf{P}\mathbf{R})^{\mathrm{t}} = \tilde{\mathbf{T}}\tilde{\mathbf{P}}^{\mathrm{t}}$$

Заметим, что новые матрицы счетов и нагрузок сохраняют все свойства старых,

$$\tilde{\mathbf{T}}^{\mathrm{t}}\tilde{\mathbf{T}} = \mathbf{T}^{\mathrm{t}}\mathbf{T} = \mathbf{\Lambda}$$

$$\tilde{\mathbf{P}}^{\mathrm{t}}\tilde{\mathbf{P}} = \mathbf{P}^{\mathrm{t}}\mathbf{P} = \mathbf{I}$$

Это свойство PCA называется вращательной неопределенностью. Оно интенсивно используется при решении задач разделения кривых, в частности методом прокрустова вращения. Если отказаться от условий ортогональности главных компонент, то декомпозиция матрицы станет еще более общей. Пусть теперь **R** – это произвольная невырожденная матрица размерностью $A \times A$. Тогда

$$\tilde{\mathbf{T}} = \mathbf{T}\mathbf{R}$$

$$\tilde{\mathbf{P}} = \mathbf{P}(\mathbf{R}^{\mathrm{t}})^{-1}$$

Эти матрицы счетов и нагрузок уже не удовлетворяют условию ортогональности и нормирования. Зато они могут состоять только из неотрицательных элементов, а также подчиняться другим требованиям, накладываемым при решении задач разделения сигналов.

3.12 Подготовка данных

Во многих случаях, перед применением PCA, исходные данные нужно предварительно подготовить: отцентрировать и/или отнормировать. Эти преобразования проводятся по столбцам – переменным.

Центрирование – это вычитание из каждого столбца $\mathbf{x}_j$ среднего (по столбцу) значения:

$$m_j = (x_{1j} + \cdots + x_{Ij})/I$$

Центрирование необходимо потому, что оригинальная PCA модель, $\mathbf{X} = \mathbf{TP}^{\mathrm{t}} + \mathbf{E}$, не содержит свободного члена.

Второе простейшее преобразование данных – это *нормирование*. Это преобразование выравнивает вклад разных переменных в PCA модель. При этом преобразовании каждый столбец $\mathbf{x}_j$ делится на свое стандартное отклонение.

$$s_j = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ij} - m_j)^2}$$

Комбинация центрирования и нормирования по столбцам называется *автошкалированием*.

$$\tilde{x}_{ij} = (x_{ij} - m_j)/s_j$$

Любое преобразование данных – центрирование, нормирование, и т.п. – всегда делается сначала на обучающем наборе. По этому набору вычисляются значения m_j и s_j, которые затем применяются и к обучающему, и к проверочному набору.

В надстройке *Chemometrics Add In* подготовка данных проводится автоматически. Если подготовку нужно провести вручную, то для нее можно использовать стандартные функции листа или специальную пользовательскую функцию.

В задачах, где структура исходных данных **X** априори предполагает однородность и гомоскедастичность, подготовка данных не только не нужна, но и вредна. Именно такой случай представляют ВЭЖХ–ДДМ данные

3.13 Размах и отклонение

При заданном числе главных компонент A, величина

$$h_i = \mathbf{t}_i^{\mathrm{t}}(\mathbf{T}_A^{\mathrm{t}}\mathbf{T}_A)^{-1}\mathbf{t}_i = \sum_{a=1}^{A}\frac{t_{ia}^2}{\lambda_a}$$

называется *размахом* (leverage). Эта величина равна квадрату расстояния Махаланобиса от центра модели до i–го образца в пространстве счетов, поэтому размах характеризует как далеко находится каждый образец в гиперплоскости главных компонент.

Для размаха имеет место соотношение

$$h_0 = \frac{1}{I}\sum_{i=1}^{I} h_i \equiv \frac{A}{I}$$

которое выполняется тождественно – по построению PCA.

Для вычисления размахов можно использовать стандартные функции листа или специальную пользовательскую функцию.

Другой важной характеристикой PCA модели является отклонение v_i, которое вычисляется как сумма квадратов остатков – квадрат эвклидова расстояния от плоскости главных компонент до объекта i.

Для вычисления отклонений можно использовать стандартные функции листа или специальную пользовательскую функцию.

Две эти величины: h_i и v_i определяют положение объекта (образца) относительно имеющейся PCA модели. Слишком большие значения размаха и/или отклонения свидетельствуют об особенности такого объекта, который может быть экстремальным или выпадающим образцом.

Анализ величин h_i и v_i составляет основу SIMCA – метода классификации с обучением.

4 Люди и страны

4.1 Пример

Метод главных компонент иллюстрируется примером, помещенным в файл People.

Этот файл включает в себя следующие листы:

- `Intro`: краткое введение
- `Layout`: схемы, объясняющая имена массивов, используемых в примере
- `Data`: данные, используемые в примере.
- `MVA`: PCA декомпозиция, выполненная с помощью надстройки Chemometrics.xla
- `PCA`: копия всех результатов PCA не привязанная к надстройке Chemometrics.xla
- `Scores1–2`: анализ младших счетов PC1–PC2
- `Scores3–4`: анализ старших счетов PC3–PC4
- `Loadings`: анализ нагрузок
- `Residuals`: анализ остатков

4.2 Данные

Анализ базируется на данных европейского демографического исследования, опубликованных в книге К. Эсбенсена.

По причинам дидактического характера используется лишь небольшой набор из 32 человек, из которых 16 представляют Северную Европу (Скандинавия) и столько же – Южную (Средиземноморье). Для баланса выбрано одинаковое количество мужчин и женщин – по 16 человек. Люди характеризуются двенадцатью переменными, перечисленными в слежующей таблице.

Переменная	Описание
Height	Рост: в сантиметрах
Weight	Вес: в килограммах
Hair	Волосы: короткие: −1, или длинные: +1
Shoes	Обувь: размер по европейскому стандарту
Age	Возраст: в годах

Переменная	Описание
Income	Доход: в тысячах евро в год
Beer	Пиво: потребление в литрах в год
Wine	Вино: потребление в литрах в год
Sex	Пол: мужской: −1, или женский: +1
Strength	Сила: индекс, основанный на проверке физических способностей
Region	Регион: север : −1, или юг: +1
IQ	Коэффициент интеллекта, измеряемый по стандартному тесту

Заметим, что такие переменные, как *Sex*, *Hair* и *Region* имеют дискретный характер с двумя возможными значениями: −1 или +1, тогда как остальные девять переменных могут принимать непрерывные числовые значения.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1			**Raw Data**											
2			**Height**	**Weight**	**Hair**	**Shoes**	**Age**	**Income**	**Beer**	**Wine**	**Sex**	**Strength**	**Region**	**IQ**
3	**1**	**MH**	198	92	-1	48	48	45	420	115	-1	98	-1	100
4	**2**	**MH**	184	84	-1	44	33	33	350	102	-1	92	-1	130
5	**3**	**MH**	183	83	-1	44	37	34	320	98	-1	91	-1	127
6	**4**	**MH**	182	80	-1	42	35	30	398	65	-1	85	-1	140
7	**5**	**MH**	180	80	-1	43	36	30	388	63	-1	84	-1	129
8	**6**	**MH**	183	81	-1	42	37	35	345	45	-1	90	-1	105
9	**7**	**MH**	180	82	-1	44	43	37	355	82	-1	88	-1	109
10	**8**	**MH**	180	81	-1	44	46	42	362	90	-1	86	-1	113
11	**9**	**MS**	185	82	-1	45	26	16	295	180	-1	92	1	109
12	**10**	**MS**	187	84	-1	46	27	16.5	299	178	-1	95	1	119
13	**11**	**MS**	177	65	-1	41	26	18	209	160	-1	86	1	120
14	**12**	**MS**	180	72	-1	43	33	19	236	175	-1	85	1	115
15	**13**	**MS**	181	75	-1	43	42	31	198	161	-1	83	1	105
16	**14**	**MS**	176	68	-1	42	50	36	195	177	-1	82	1	96
17	**15**	**MS**	175	67	1	42	55	38	185	187	-1	80	1	105
18	**16**	**MS**	178	75	-1	42	30	24	203	208	-1	81	1	118
19	**17**	**FH**	166	47	-1	36	32	28	270	78	1	75	-1	112
20	**18**	**FH**	170	60	1	38	23	20	312	99	1	81	-1	110
21	**19**	**FH**	172	64	1	39	24	22	308	91	1	82	-1	102
22	**20**	**FH**	169	51	1	36	24	23	250	89	1	78	-1	98
23	**21**	**FH**	168	52	1	37	27	23.5	260	86	1	78	-1	100
24	**22**	**FH**	157	47	1	36	32	32	235	92	1	70	-1	127
25	**23**	**FH**	164	50	1	38	41	34	255	134	1	76	-1	101
26	**24**	**FH**	162	49	1	37	40	34	265	124	1	75	-1	108
27	**25**	**FS**	168	50	1	37	49	34	170	162	1	76	1	135
28	**26**	**FS**	166	49	1	36	21	14	150	245	1	75	1	123
29	**27**	**FS**	158	46	1	34	30	18	120	120	1	70	1	119
30	**28**	**FS**	163	50	1	36	18	11	143	136	1	75	1	102
31	**29**	**FS**	162	50	1	36	20	11.5	133	146	1	74	1	132
32	**30**	**FS**	165	51	1	36	36	26	121	129	1	76	1	126
33	**31**	**FS**	161	48	1	35	41	31.5	116	196	1	75	1	120
34	**32**	**FS**	160	48	1	35	40	31	118	198	1	74	1	129
35	**mean**		173.1	64.5	0.0	39.9	34.4	27.4	249.5	131.6	0.0	81.5	0.0	115.1
36	**STD**		10.1	15.2	1.0	3.9	9.5	8.9	90.6	49.5	1.0	7.3	1.0	12.2
37														

Рис. 4.1. Исходные данные в примере People

4.3 Исследование данных

Прежде всего, любопытно посмотреть на графиках, как связаны между собой все эти переменные. Зависит ли рост (*Height*) от веса (*Weight*)? Отличаются ли женщины от мужчин в потреблении вина (*Wine*)? Связан ли доход (*Income*) с возрастом (*Age*)? Зависит ли вес (*Weight*) от потребления пива (*Beer*)?

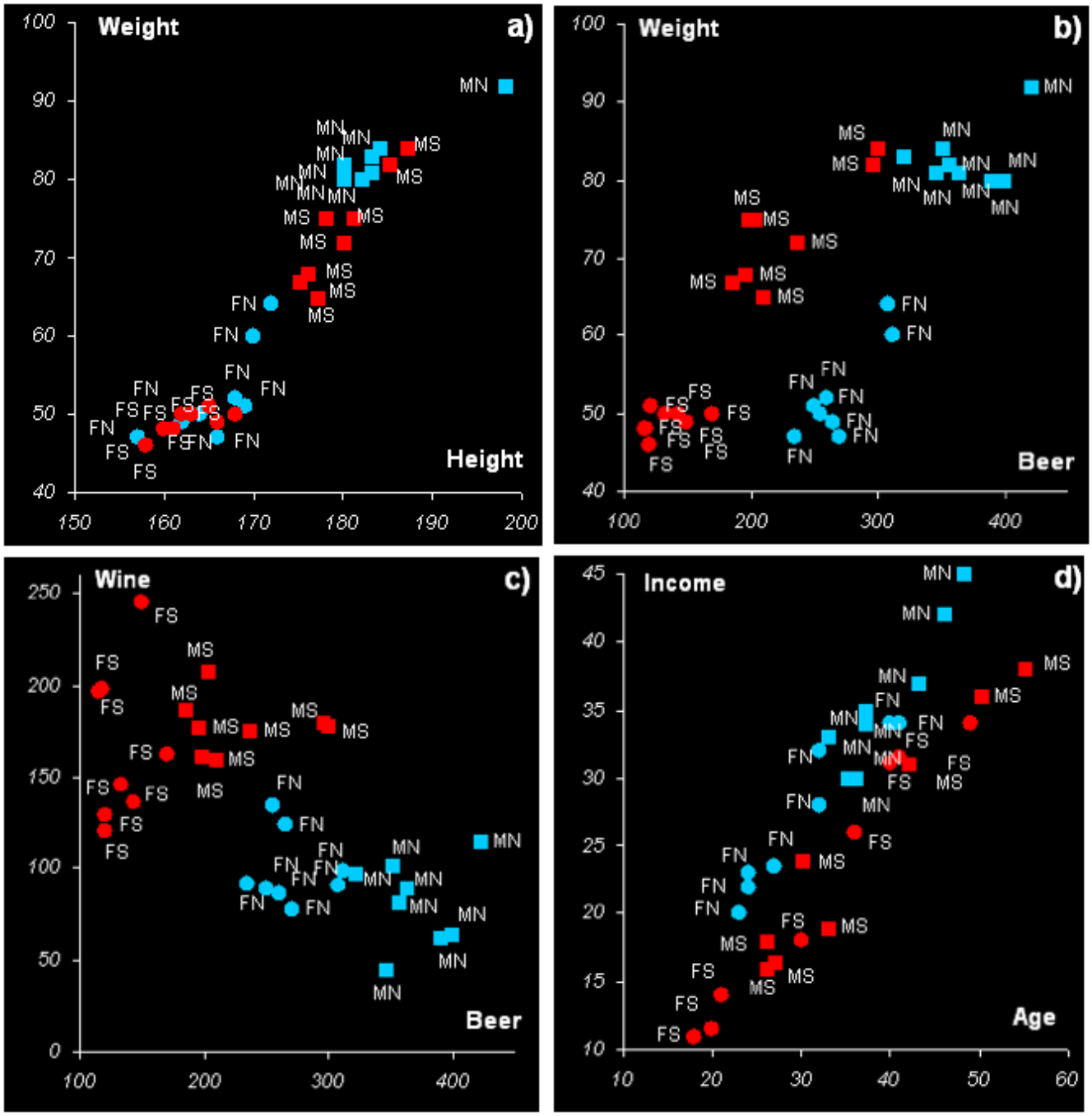


Рис. 4.2. Связи между переменными в примере People. Женщины (F) обозначены кружками ●, а мужчины (M) – квадратами ■. Север (N) представлен голубым, а юг (S) – красным цветом.

Некоторые из этих зависимостей показаны на Рис. 4.1. Для наглядности на всех графиках использованы одни и те же обозначения: женщины (F) показаны кружками, мужчины (M) – квадратами, север (N) представлен голубым, а юг (S) – красным цветом.

Связь между весом (*Weight*) и ростом (*Height*) показана на Рис. 4.2a. Очевидна, прямая (положительная) пропорциональность. Учитывая маркировку точек, можно заметить также, что мужчины (M) в большинстве своем тяжелее и выше женщин (F).

На Рис. 4.2b показана другая пара переменных: вес (*Weight*) и пиво (*Beer*). Здесь, помимо очевидных фактов, что большие люди пьют больше, а женщины – меньше, чем мужчины, можно заметить еще две отдельные группы – южан и северян. Первые пьют меньше пива при том же весе.

Эти же группы заметны и на Рис. 4.2c, где показана зависимость между потреблением вина (*Wine*) и пива (*Beer*). Из него видно, что связь между этими переменными отрицательна – чем больше потребляется пива, тем меньше вина. На юге пьют больше вина, а на севере – пива. Интересно, что в обеих группах женщины располагаются слева, но не ниже по отношению к мужчинам. Это означает, что, потребляя меньше пива, прекрасный пол не уступает в вине.

Последний график на Рис. 4.2d показывает, как связаны возраст (*Age*) и доход (*Income*). Легко видеть, что даже в этом сравнительно небольшом наборе данных есть переменные, как с положительной, так и с отрицательной корреляцией.

Можно ли построить графики для всех пар переменных выборки? Вряд ли. Проблема состоит в том, что для 12 переменных существует $12(12-1)/2 = 66$ таких комбинаций.

4.4 Подготовка данных

Перед тем, как подвергнуть данные анализу методом главных компонент, их надо подготовить. Простой статистический расчет показывает, что они нуждаются в автошкалировании (см. Рис. 4.3).

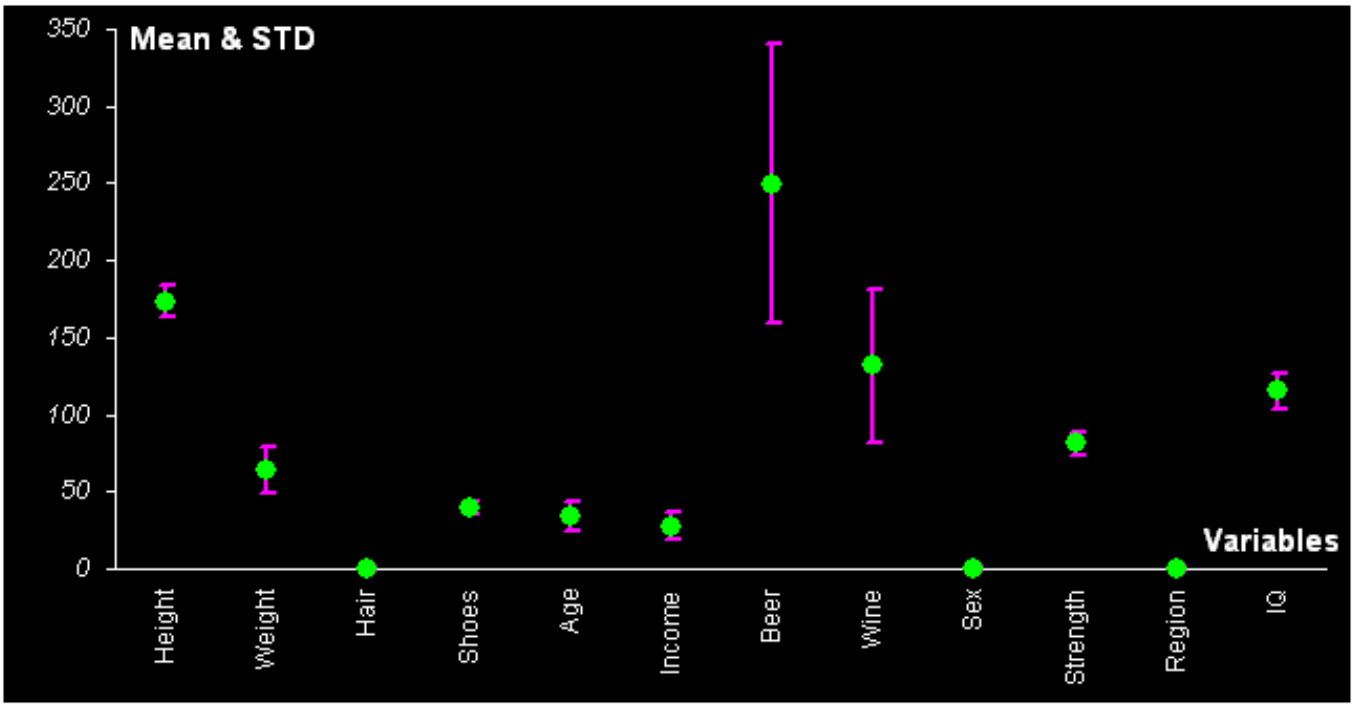


Рис. 4.3. Средние значения и СКО для переменных в примере People

Средние значения по многим переменным отличаются от нуля. Кроме того, среднеквадратичные отклонения сильно разнятся. После автошкалирования среднее значение всех переменных становится равно нулю, а отклонение – единица.

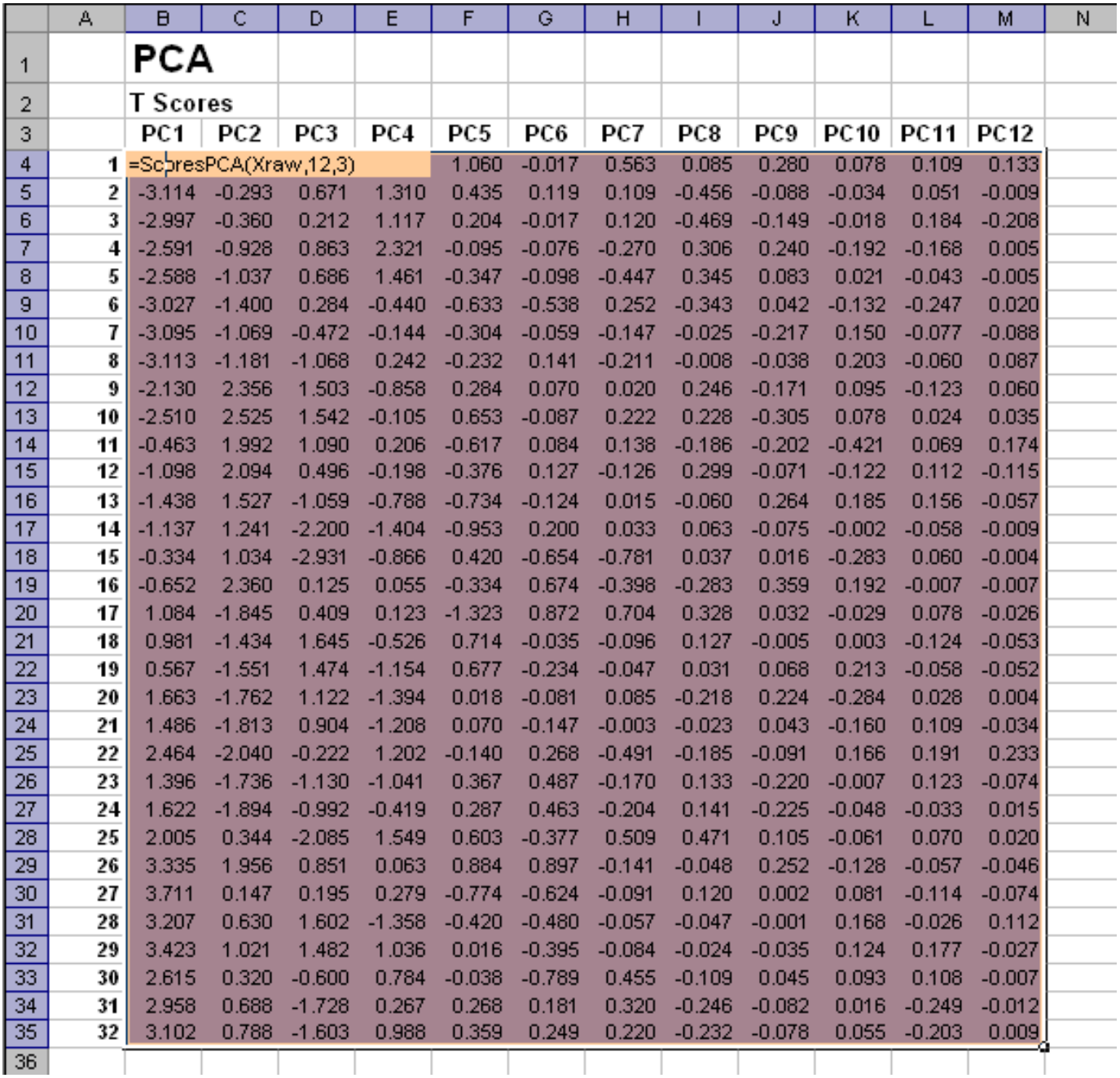
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
39			**Autoscaled Data**											
40			**Height**	**Weight**	**Hair**	**Shoes**	**Age**	**Income**	**Beer**	**Wine**	**Sex**	**Strength**	**Region**	**IQ**
41	**1**	**MN**	2.47	1.81	-0.98	2.08	1.43	1.97	1.88	-0.34	-0.98	2.25	-0.98	-1.24
42	**2**	**MN**	1.08	1.29	-0.98	1.05	-0.15	0.62	1.11	-0.60	-0.98	1.43	-0.98	1.22
43	**3**	**MN**	0.98	1.22	-0.98	1.05	0.27	0.73	0.78	-0.68	-0.98	1.30	-0.98	0.98
44	**4**	**MN**	0.88	1.02	-0.98	0.54	0.06	0.29	1.64	-1.35	-0.98	0.48	-0.98	2.04
45	**5**	**MN**	0.68	1.02	-0.98	0.79	0.16	0.29	1.53	-1.39	-0.98	0.34	-0.98	1.14
46	**6**	**MN**	0.98	1.09	-0.98	0.54	0.27	0.85	1.05	-1.75	-0.98	1.16	-0.98	-0.83
47	**7**	**MN**	0.68	1.15	-0.98	1.05	0.90	1.07	1.16	-1.00	-0.98	0.89	-0.98	-0.50
48	**8**	**MN**	0.68	1.09	-0.98	1.05	1.21	1.63	1.24	-0.84	-0.98	0.61	-0.98	-0.17
49	**9**	**MS**	1.18	1.15	-0.98	1.31	-0.89	-1.28	0.50	0.98	-0.98	1.43	0.98	-0.50
50	**10**	**MS**	1.38	1.29	-0.98	1.56	-0.78	-1.22	0.55	0.94	-0.98	1.84	0.98	0.32
51	**11**	**MS**	0.39	0.03	-0.98	0.28	-0.89	-1.06	-0.45	0.57	-0.98	0.61	0.98	0.40
52	**12**	**MS**	0.68	0.50	-0.98	0.79	-0.15	-0.94	-0.15	0.88	-0.98	0.48	0.98	-0.01
53	**13**	**MS**	0.78	0.69	-0.98	0.79	0.79	0.40	-0.57	0.59	-0.98	0.20	0.98	-0.83
54	**14**	**MS**	0.29	0.23	-0.98	0.54	1.64	0.96	-0.60	0.92	-0.98	0.07	0.98	-1.57
55	**15**	**MS**	0.19	0.17	0.98	0.54	2.16	1.18	-0.71	1.12	-0.98	-0.20	0.98	-0.83
56	**16**	**MS**	0.48	0.69	-0.98	0.54	-0.47	-0.38	-0.51	1.54	-0.98	-0.07	0.98	0.24
57	**17**	**FN**	-0.71	-1.15	-0.98	-1.00	-0.26	0.06	0.23	-1.08	0.98	-0.89	-0.98	-0.26
58	**18**	**FN**	-0.31	-0.29	0.98	-0.49	-1.20	-0.83	0.69	-0.66	0.98	-0.07	-0.98	-0.42
59	**19**	**FN**	-0.11	-0.03	0.98	-0.23	-1.10	-0.61	0.65	-0.82	0.98	0.07	-0.98	-1.08
60	**20**	**FN**	-0.41	-0.89	0.98	-1.00	-1.10	-0.50	0.01	-0.86	0.98	-0.48	-0.98	-1.41
61	**21**	**FN**	-0.51	-0.82	0.98	-0.75	-0.78	-0.44	0.12	-0.92	0.98	-0.48	-0.98	-1.24
62	**22**	**FN**	-1.60	-1.15	0.98	-1.00	-0.26	0.51	-0.16	-0.80	0.98	-1.57	-0.98	0.98
63	**23**	**FN**	-0.91	-0.95	0.98	-0.49	0.69	0.73	0.06	0.05	0.98	-0.75	-0.98	-1.16
64	**24**	**FN**	-1.11	-1.02	0.98	-0.75	0.58	0.73	0.17	-0.15	0.98	-0.89	-0.98	-0.59
65	**25**	**FS**	-0.51	-0.95	0.98	-0.75	1.53	0.73	-0.88	0.61	0.98	-0.75	0.98	1.63
66	**26**	**FS**	-0.71	-1.02	0.98	-1.00	-1.41	-1.50	-1.10	2.29	0.98	-0.89	0.98	0.65
67	**27**	**FS**	-1.50	-1.22	0.98	-1.52	-0.47	-1.06	-1.43	-0.23	0.98	-1.57	0.98	0.32
68	**28**	**FS**	-1.01	-0.95	0.98	-1.00	-1.73	-1.84	-1.18	0.09	0.98	-0.89	0.98	-1.08
69	**29**	**FS**	-1.11	-0.95	0.98	-1.00	-1.52	-1.78	-1.29	0.29	0.98	-1.02	0.98	1.39
70	**30**	**FS**	-0.81	-0.89	0.98	-1.00	0.16	-0.16	-1.42	-0.05	0.98	-0.75	0.98	0.89
71	**31**	**FS**	-1.21	-1.08	0.98	-1.26	0.69	0.45	-1.47	1.30	0.98	-0.89	0.98	0.40
72	**32**	**FS**	-1.31	-1.08	0.98	-1.26	0.58	0.40	-1.45	1.34	0.98	-1.02	0.98	1.14
73	**mean**		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
74	**STD**		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
75														

Рис. 4.4. Автошкалированные данные в примере People

В принципе, данные можно было бы не преобразовывать явно, на листе, а оставить как есть. Ведь стандартные хемометрические процедуры, собранные в надстройке *Chemometrics* могут центрировать и шкалировать данные при выполнении вычислений. Однако матрица автошкалированных данных понадобится нам при вычислении остатков.

4.5 Вычисление счетов и нагрузок

Для построения PCA декомпозиции можно воспользоваться стандартными функциями `ScoresPCA` и `LoadingsPCA`, имеющимися в надстройке *Chemometrics*. Мы вычислим все 12 возможных главных компонент. В качестве первого аргумента используется исходный, не преобразованный массив данных, поэтому последний аргумент в обеих функциях равен `3` – автошкалирование.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		**PCA**												
2		T Scores												
3		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	
4	1	=ScoresPCA(Xraw,12,3)				1.060	-0.017	0.563	0.085	0.280	0.078	0.109	0.133	
5	2	-3.114	-0.293	0.671	1.310	0.435	0.119	0.109	-0.456	-0.088	-0.034	0.051	-0.009	
6	3	-2.997	-0.360	0.212	1.117	0.204	-0.017	0.120	-0.469	-0.149	-0.018	0.184	-0.208	
7	4	-2.591	-0.928	0.863	2.321	-0.095	-0.076	-0.270	0.306	0.240	-0.192	-0.168	0.005	
8	5	-2.588	-1.037	0.686	1.461	-0.347	-0.098	-0.447	0.345	0.083	0.021	-0.043	-0.005	
9	6	-3.027	-1.400	0.284	-0.440	-0.633	-0.538	0.252	-0.343	0.042	-0.132	-0.247	0.020	
10	7	-3.095	-1.069	-0.472	-0.144	-0.304	-0.059	-0.147	-0.025	-0.217	0.150	-0.077	-0.088	
11	8	-3.113	-1.181	-1.068	0.242	-0.232	0.141	-0.211	-0.008	-0.038	0.203	-0.060	0.087	
12	9	-2.130	2.356	1.503	-0.858	0.284	0.070	0.020	0.246	-0.171	0.095	-0.123	0.060	
13	10	-2.510	2.525	1.542	-0.105	0.653	-0.087	0.222	0.228	-0.305	0.078	0.024	0.035	
14	11	-0.463	1.992	1.090	0.206	-0.617	0.084	0.138	-0.186	-0.202	-0.421	0.069	0.174	
15	12	-1.098	2.094	0.496	-0.198	-0.376	0.127	-0.126	0.299	-0.071	-0.122	0.112	-0.115	
16	13	-1.438	1.527	-1.059	-0.788	-0.734	-0.124	0.015	-0.060	0.264	0.185	0.156	-0.057	
17	14	-1.137	1.241	-2.200	-1.404	-0.953	0.200	0.033	0.063	-0.075	-0.002	-0.058	-0.009	
18	15	-0.334	1.034	-2.931	-0.866	0.420	-0.654	-0.781	0.037	0.016	-0.283	0.060	-0.004	
19	16	-0.652	2.360	0.125	0.055	-0.334	0.674	-0.398	-0.283	0.359	0.192	-0.007	-0.007	
20	17	1.084	-1.845	0.409	0.123	-1.323	0.872	0.704	0.328	0.032	-0.029	0.078	-0.026	
21	18	0.981	-1.434	1.645	-0.526	0.714	-0.035	-0.096	0.127	-0.005	0.003	-0.124	-0.053	
22	19	0.567	-1.551	1.474	-1.154	0.677	-0.234	-0.047	0.031	0.068	0.213	-0.058	-0.052	
23	20	1.663	-1.762	1.122	-1.394	0.018	-0.081	0.085	-0.218	0.224	-0.284	0.028	0.004	
24	21	1.486	-1.813	0.904	-1.208	0.070	-0.147	-0.003	-0.023	0.043	-0.160	0.109	-0.034	
25	22	2.464	-2.040	-0.222	1.202	-0.140	0.268	-0.491	-0.185	-0.091	0.166	0.191	0.233	
26	23	1.396	-1.736	-1.130	-1.041	0.367	0.487	-0.170	0.133	-0.220	-0.007	0.123	-0.074	
27	24	1.622	-1.894	-0.992	-0.419	0.287	0.463	-0.204	0.141	-0.225	-0.048	-0.033	0.015	
28	25	2.005	0.344	-2.085	1.549	0.603	-0.377	0.509	0.471	0.105	-0.061	0.070	0.020	
29	26	3.335	1.956	0.851	0.063	0.884	0.897	-0.141	-0.048	0.252	-0.128	-0.057	-0.046	
30	27	3.711	0.147	0.195	0.279	-0.774	-0.624	-0.091	0.120	0.002	0.081	-0.114	-0.074	
31	28	3.207	0.630	1.602	-1.358	-0.420	-0.480	-0.057	-0.047	-0.001	0.168	-0.026	0.112	
32	29	3.423	1.021	1.482	1.036	0.016	-0.395	-0.084	-0.024	-0.035	0.124	0.177	-0.027	
33	30	2.615	0.320	-0.600	0.784	-0.038	-0.789	0.455	-0.109	0.045	0.093	0.108	-0.007	
34	31	2.958	0.688	-1.728	0.267	0.268	0.181	0.320	-0.246	-0.082	0.016	-0.249	-0.012	
35	32	3.102	0.788	-1.603	0.988	0.359	0.249	0.220	-0.232	-0.078	0.055	-0.203	0.009	
36														

Рис. 4.5. Вычисление матрицы счетов

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
2			P Loadings												
3			PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	
4		Height	=LoadingsPCA(Xraw,12,3)				0.186	-0.124	0.268	0.118	0.729	-0.307	0.255	-0.065	
5		Weight	-0.381	0.111	0.068	0.033	0.100	-0.192	-0.224	-0.219	0.190	0.572	-0.415	-0.395	
6		Hair	0.338	-0.150	-0.079	-0.114	0.660	-0.489	-0.368	-0.081	0.041	-0.140	0.007	0.078	
7		Shoes	-0.378	0.151	0.001	-0.066	0.152	-0.031	-0.234	0.171	-0.280	0.376	0.685	0.183	
8		Age	-0.143	-0.061	-0.720	0.055	-0.029	-0.165	0.043	0.435	-0.174	-0.134	-0.036	-0.430	
9		Income	-0.190	-0.287	-0.586	0.085	0.063	0.137	0.129	-0.434	0.180	0.167	-0.038	0.492	
10		Beer	-0.325	-0.308	0.188	0.040	0.231	0.239	-0.170	0.567	-0.015	-0.049	-0.420	0.350	
11		Wine	0.124	0.554	-0.212	-0.125	0.415	0.638	-0.120	-0.040	0.024	-0.054	-0.095	-0.093	
12		Sex	0.352	-0.232	0.052	-0.051	0.313	0.098	0.580	0.254	0.078	0.529	0.084	-0.124	
13		Strength	-0.365	0.112	0.135	-0.081	0.336	-0.160	0.512	-0.258	-0.530	-0.232	-0.165	-0.001	
14		Region	0.144	0.595	-0.130	-0.022	-0.151	-0.402	0.161	0.265	0.050	0.180	-0.255	0.476	
15		IQ	0.044	0.123	0.062	0.969	0.180	-0.010	0.024	0.001	-0.006	-0.033	0.076	-0.010	
16															

Рис. 4.6. Вычисление матрицы нагрузок

В этом пособии все PCA вычисления проводятся в книге *People.xls* на листе `MVA`. Для удобства читателя эти же результаты продублированы на листе `PCA` как числа, без ссылки на надстройку *Chemometrics.xla*. Остальные листы рабочей книги связаны не с данными на листе `MVA`, с данными на листе `PCA`. Поэтому файл *People.xls* можно использовать даже тогда, когда надстройка *Chemometrics.xla* не установлена на компьютере.

4.6 График счетов

Посмотрим на графики счетов, которые показывают, как расположены образцы в проекционном пространстве.

На графике младших счетов PC1–PC2 (Рис. 4.7) мы видим четыре отдельные группы, разложенные по четырем квадрантам: слева – женщины (F), справа – мужчины (M), сверху – юг (S), а снизу – север (N). Из этого сразу становится ясен смысл первых двух направлений PC1 и PC2. Первая компонента разделяет людей по полу, а вторая – по месту жительства. Именно эти факторы наиболее сильно влияют на разброс свойств.

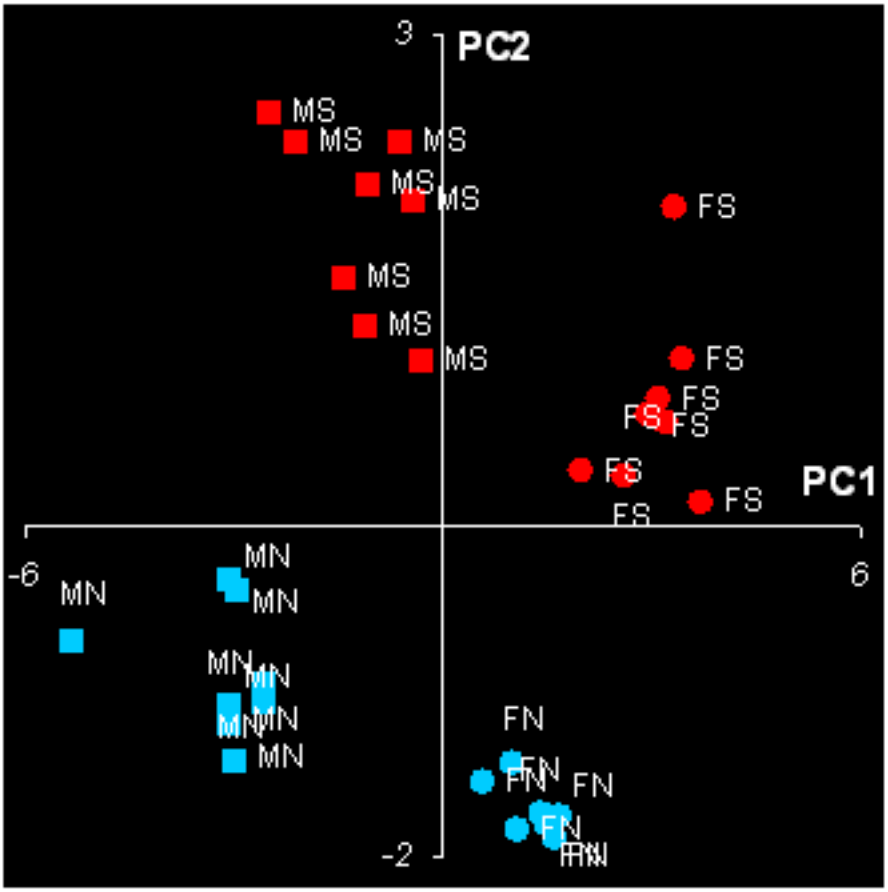


Рис. 4.7. График счетов (PC1 – PC2) с обозначениями, использованными ранее на Рис 4.2

Продолжим изучение, построив график старших счетов PC3– PC4 (Рис. 4.8).

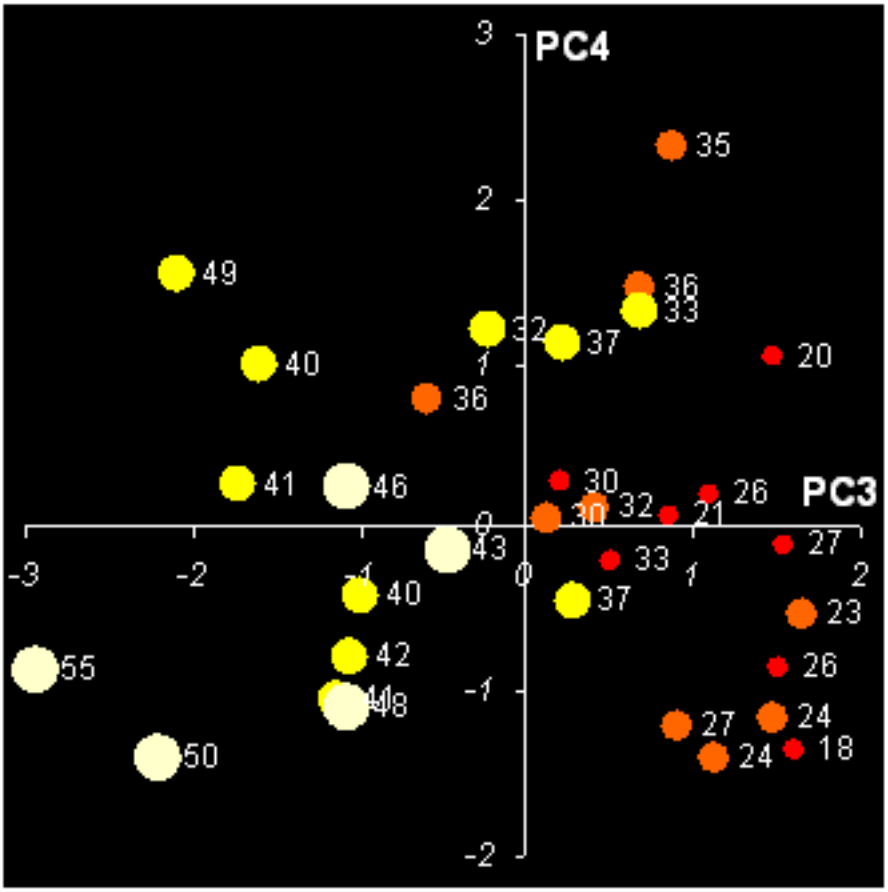


Рис. 4.8. График счетов (PC3 – PC4) с новыми обозначениями: размер и цвет символов отражает доход – чем больше и светлее, тем он больше. Числа представляют возраст

Здесь уже не видно таких отчетливых групп. Тем не менее, внимательно исследовав этот график совместно с таблицей исходных данных, можно, после некоторых усилий, сделать вывод о том, что PC3 отделяет старых/богатых людей от молодых/бедных. Чтобы сделать это более очевидным, мы изменили обозначения. Теперь каждый человек показан кружком, цвет и размер которого меняется в зависимости от дохода – чем больше и светлее, тем больше доход. Рядом показан возраст каждого объекта. Как видно, возраст и доход уменьшается слева направо, т.е. вдоль PC3. А вот смысл PC4 нам по–прежнему не ясен.

4.7 Графики нагрузок

Чтобы разобраться с этим, построим соответствующие графики нагрузок. Они подскажут нам, какие переменные и как связаны между собой, что влияет на что.

Из графика младших компонент мы сразу видим, что переменные рост (*Height*), вес (*Weight*), сила (*Strength*) и обувь (*Shoes*) образуют компактную группу в правой части графика. Они практически сливаются, что означает их тесную положительную корреляцию. Переменные волосы (*Hair*) и пол (*Sex*) находятся в

другой группе, лежащей по диагонали от первой группы. Это свидетельствует о высокой отрицательной корреляции между переменными из этих групп, например, силой (*Strength*) и полом (*Sex*). Наибольшие нагрузки на вторую компоненту имеют переменные вино (*Wine*) и регион (*Region*), также тесно связанные друг с другом. Переменная доход (*Income*) лежит на первом графике напротив переменной регион (*Region*), что отражает дифференциацию состоятельности: Север–Юг. Можно заметить также и антитезу переменных пиво (*Beer*) – регион/вино (*Region*/*Wine*).

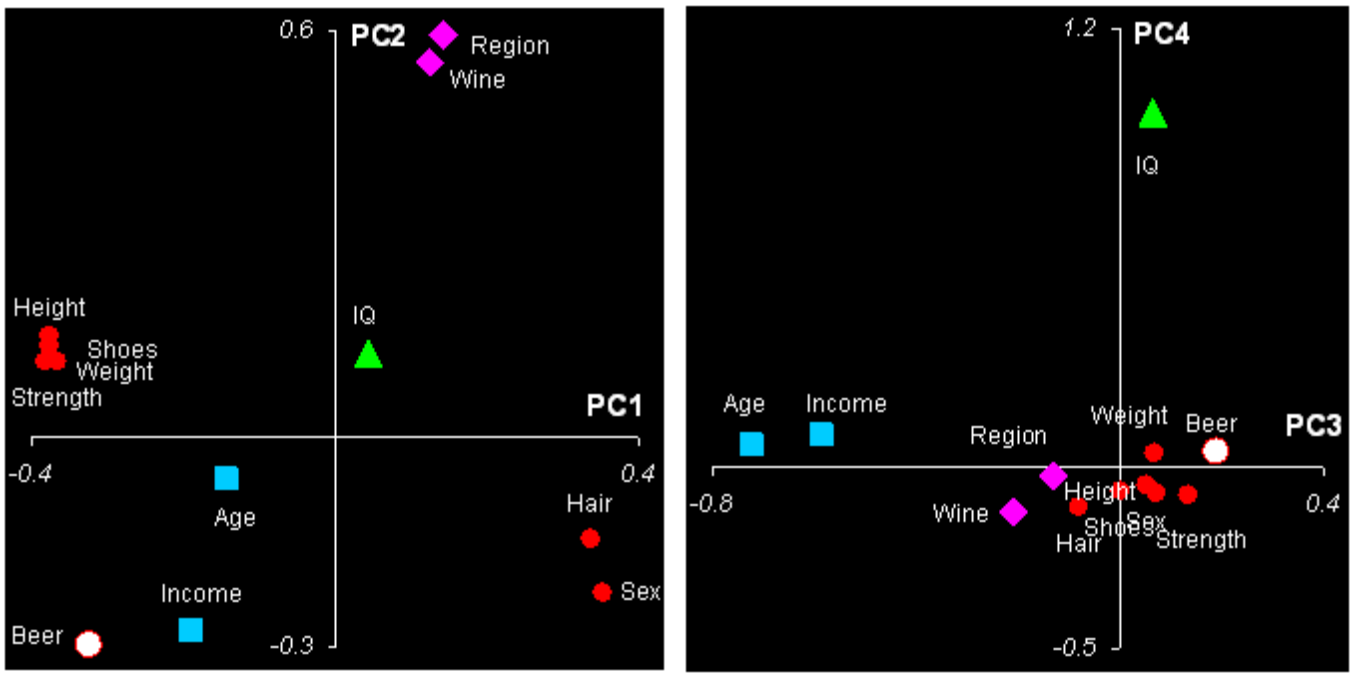


Рис. 4.9. Графики нагрузок: PC1–PC2 и PC3–PC4

Из второго графика мы видим большие нагрузки переменных возраст (*Age*) и доход (*Income*) на ось PC3, что соответствует графику счетов на Рис. 4.7. Рассмотрим, переменные пиво (*Beer*) и *IQ*. Первая из них имеет большие нагрузки как на PC1, так и на PC2, фактически формируя диагональ взаимоотношений между объектами на графике счетов. Переменная *IQ* не обнаруживает связи с другими переменным, так как ее значения близки к нулю для нагрузок первых трех PC, и проявляет она себя только в четвертой компоненте. Мы видим, что значения *IQ* не зависят от места жительства, физиологических характеристик и пристрастий к напиткам.

Впервые PCA был применен еще в начале 20–го века в психологических исследованиях, когда верили, что такие показатели, как *IQ* или криминальное поведение можно объяснить с помощью индивидуальных физиологических и социальных характеристик. Если сравнить результаты PCA с графиками, построенными нами ранее для пар переменных, видно, что PCA сразу дает всеобъемлющее представление о структуре данных, которое можно “охватить одним взглядом” (точнее, с помощью четырех графиков). Поэтому, одна из наиболее сильных сторон PCA в исследовании структур данных –

это переход от большого числа не связанных между собой графиков пар переменных к очень небольшому числу графиков счетов и нагрузок.

4.8 Исследование остатков

Сколько главных компонент нужно использовать в этом примере? Для ответа на вопрос нужно исследовать, как изменяется качество описания при увеличении числа PC. Заметим, что в этом примере мы не будем проводить проверку – в этом нет необходимости, т.к. PCA модель нужна только для исследования данных. Она не будет использоваться далее для прогнозирования, классификации, и т.п.

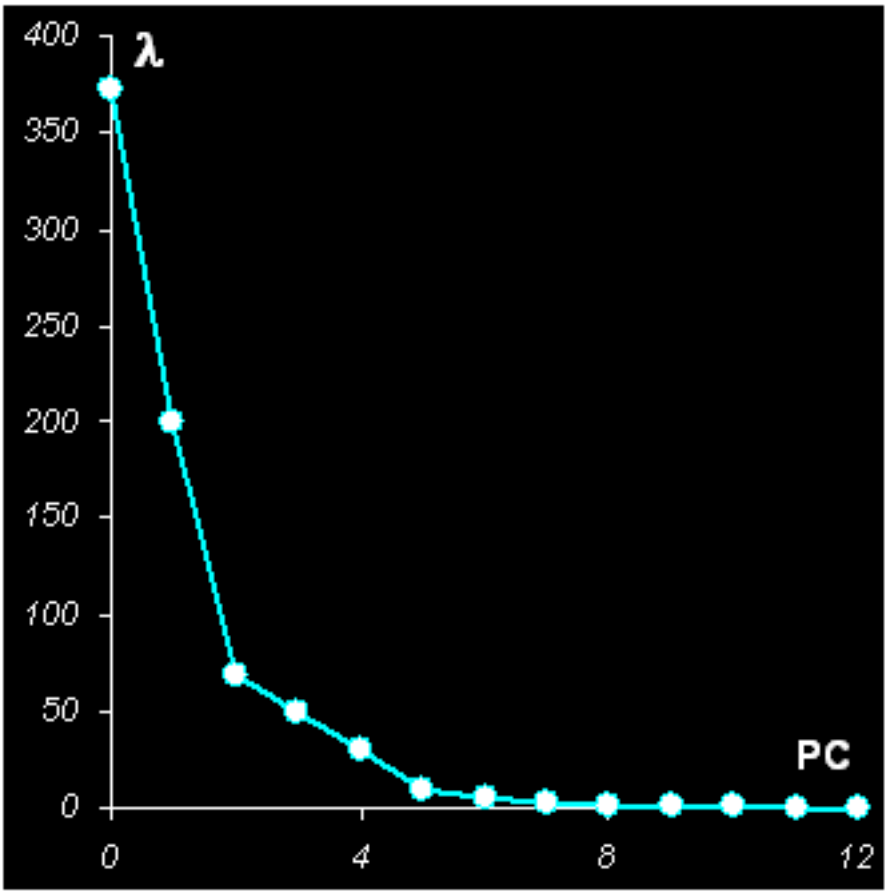


Рис. 4.10. Графики собственных значений

На Рис. 4.10 показано, как, в зависимости от числа PC, меняются собственные значения λ . Видно, что около PC=5 происходит изменение в их поведении. Для расчета показателей *TRV* и *ERV* можно получить матрицу остатков **E** для каждого числа главных компонент A и вычислить требуемые показатели. Пример такого расчета для значения $A = 4$ приведен на листе `Residuals`.

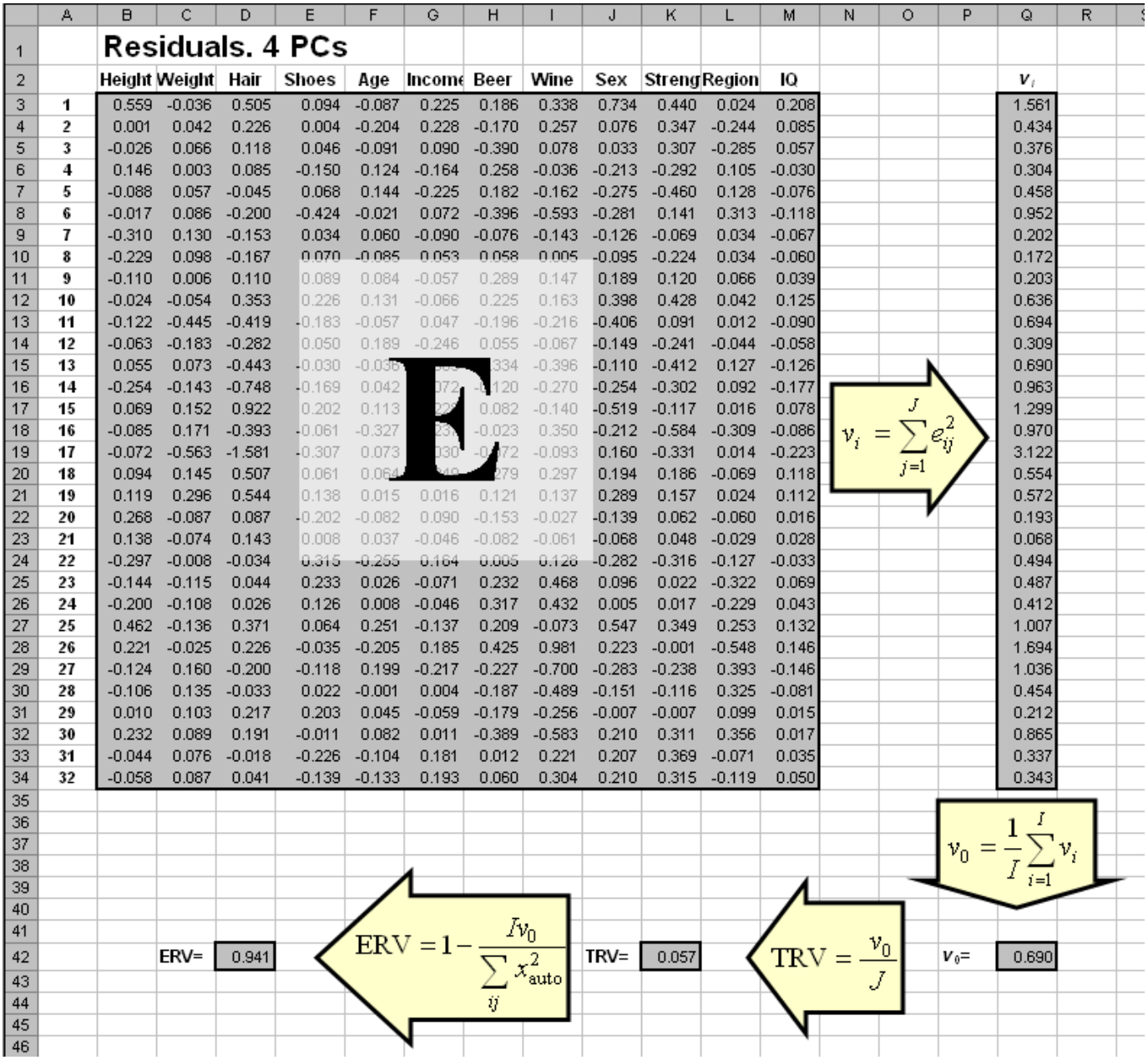


Рис. 4.11. Анализ остатков

Однако те же характеристики можно получить и проще, если воспользоваться соотношениями

$$TRV(A) = \frac{1}{IJ}\left(\lambda_0 - \sum_{a=1}^{A} \lambda_a\right)$$

$$ERV(A) = 1 - \frac{TRV(A)}{TRV(0)}$$

Эти величины представлены на Рис. 28

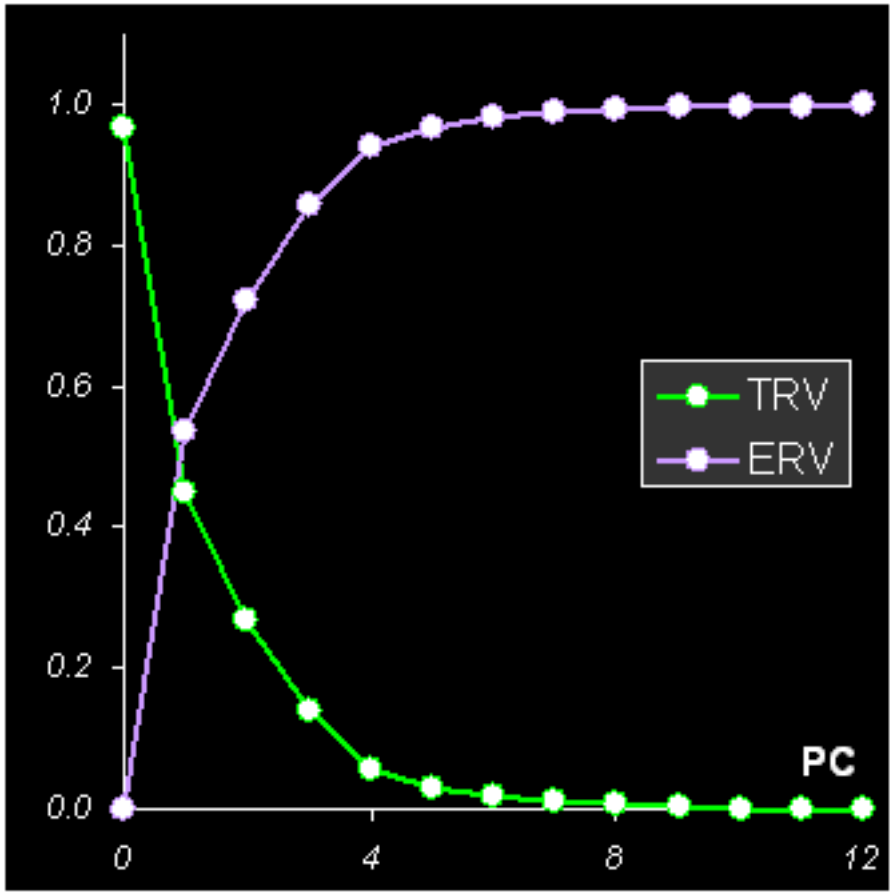


Рис. 4.12. Графики полной (TRV) и объясненной (ERV) дисперсии остатков

Из этих зависимостей видно, что для описания данных достаточно четырех PC – они моделируют 94% данных, или, иными словами, шум, оставшийся после проекции на четырехмерное пространство PC1–PC4, оставляет всего 6% от исходных данных.

5 Заключение

Рассмотренный пример позволил взглянуть лишь на малую часть возможностей, предоставляемых PCA–моделированием. Мы рассмотрели задачу исследования данных, которая не предполагает дальнейшего использования построенной модели для предсказания или классификации.

Метод PCA дает основу разнообразным методам, применяемым в хемометрике. В задачах классификации – это метод SIMCA, в задачах калибровки – это метод PCR, в задачах разделения кривых – это EFA, WFA и т.д.