МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ СНИЖЕНИЯ РАЗМЕРНОСТИ БИОМЕДИЦИНСКИХ ДАННЫХ

Методические указания к лабораторной работе

САМАРА 2012

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ СНИЖЕНИЯ РАЗМЕРНОСТИ БИОМЕДИЦИНСКИХ ДАННЫХ

САМАРА 2012

УДК 519.688

Составитель: В.Н. Конюхов

Исследование методов и алгоритмов снижения размерности биомедицинских данных. Метод. указания к лабораторной работе/ Самар. гос. аэрокосм. ун-т; Сост. В.Н. Конюхов, Самара, 2012. 19с.

В методических указаниях изложены сведения об основных методах и алгоритмах снижения размерности биомедицинских данных. Показаны области применения, назначение и ограничения. Рассмотрены некоторые пакеты прикладных программ, позволяющие решать задачи снижения размерности биомедицинских данных.

Методические указания предназначены для бакалавров, обучающихся по направления подготовки 201000.62 (Биотехнические системы и технологии) и выполняющих лабораторные работы по дисциплине «Автоматизация обработки биомедицинской информации» на кафедре радиотехники и медицинских диагностических систем.

Печатаются по решению редакционно-издательского совета Самарского государственного аэрокосмического университета им. академика С.П. Королёва

Рецензент: проф. Гречишников В.М.

Цель работы: изучение методов и алгоритмов снижения размерности биомедицинских данных, а также основных задач, решаемых с помощью этих методов и ограничений, накладываемых на них.

1. КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

1.1. ОСНОВНЫЕ ПОЛОЖЕНИЯ.

Как уже отмечалось, одной из основных особенностей данных, описывающих биологический объект, является их существенная многомерность. Количество переменных, описывающих биообъект, может достигать десятков и сотен и, соответственно, число измерений нескольких тысяч. В такой ситуации зачастую трудно выделить основные закономерности, скрытые в полученных данных. Например, если переменных две- X и Y, то иногда достаточно построить на плоскости точки с координатами (X_i, Y_i), где i-число различных объектов, чтобы визуально оценить взаимосвязь между переменными. Так, если X- рост человека, а Y- вес, то, измерив значения этих переменных у группы людей, и нанеся точки на плоскость, можно увидеть, что, в среднем, чем больше рост, тем больше вес. Добавив третью переменную Z- длина рук, и построив точки с координатами (X_i, Y_i, Z_i) в трехмерном пространстве, можно заметить, что длина рук зависит от роста. Однако, если переменных больше чем три, то визуализация становиться невозможной.

В типичном клиническом исследовании число исследованных признаков бывает достаточно велико. Однако большое количество признаков (большая размерность признакового пространства) не только приводит к увеличению полноты и глубины исследования проблемы, но и маскирует имеющиеся закономерности.

Методы снижения размерности многомерного пространства позволяют без существенной потери информации перейти от первоначальной системы большого числа наблюдаемых взаимосвязанных факторов к системе существенно меньшего числа скрытых (ненаблюдаемых) факторов, определяющих вариацию первоначальных признаков.

Основными целями, для достижения которых используются методы снижения размерности пространства признаков, являются:

- достижение большей наглядности данных и возможность построения диаграмм и графиков в пространствах меньшей размерности;
- получение простых и ясных зависимостей между признаками за счет снижения числа переменных математической модели;
- резкое снижение объемов хранимой информации.

Сокращение размерности пространства исходных признаков k можно осуществить за счет выбора за счет выбора значительно меньшего числа признаков l. Выбор может

осуществляться как из числа исходных признаков, так и путем комбинирования в новые признаки исходных. При этом возможны, исходя из условий задачи, различные требования к новым признакам, обеспечивающие их оптимальность по различным критериям. Такими критериями, например, могут быть:

-сохранение в некоторых смыслах максимальной части информации, имеющейся в исходной выборке;

-обеспечение некоррелированности новых признаков;

-наименее возможное искажение геометрической структуры данных при переходе от пространства размерности k к пространству размерности l и т.д.

После того, как критерий оптимальности выбран и формализован (т.е. получена некоторая числовая характеристика снижения размерности), можно решать задачу оптимального выбора пространства размерности l.

Критерии оптимальности могут определяться как внутренней структурой данных, так и соображениями, не связанными с самими данными, а определяемыми внешними условиями. Например, требование минимального искажения геометрической структуры данных очевидно при визуализации многомерных данных в пространстве меньшей размерности. Обеспечение некоррелированности признаков обычно бывает полезной в задачах классификации.

Основными предпосылками, позволяющими эффективно решать задачу снижения размерности, являются:

-сильная коррелированность исходных признаков, что приводит к дублированию содержащихся в этих признаках информации;

-малая информативность некоторых признаков, например в том случае, если признак сильно зашумлен;

-возможность агрегирования нескольких признаков в один.

В настоящее время предложен ряд методов уменьшения размерности данных. К ним относятся методы факторного анализа, многомерного шкалирования, кластерного анализа, самоорганизующихся карт Кохонена и ряд других. В данной лабораторной работе в качестве метода сокращения пространства признаков выбран метод главных компонент.

1.2. МЕТОД ГЛАВНЫХ КОМПОНЕНТ.

Метод главных компонент - один из методов исследования структуры и снижения размерности пространства переменных. Метод главных компонент предназначен для линейного преобразования большого количества исходных переменных, коррелирующих между собой, в несколько агрегированных некоррелированных показателей. Метод главных компонент может также использоваться для исследования структуры связей между исходными переменными.

Метод главных компонент состоит в разложении (с помощью ортогонального преобразования) p-мерного случайного вектора $X_1,...,X_p$ по системе линейно независимых векторов, в качестве которой выбирается ортонормированная система собственных векторов, отвечающих собственным значениям ковариационной матрицы вектора **X**.

Пусть имеется p случайных переменных $X_1,...,X_p$ с многомерным, необязательно нормальным, совместным распределением, вектором средних $\mu^{p\times 1} = (\mu_1,...,\mu_p)'$ и ковариационной матрицей $\Sigma^{p\times p} = (\sigma_{ij})$. В некоторых случаях можно найти линейные комбинации $Y_1,...,Y_q$ переменных $X_1,...,X_p$, ($q < p$), по которым можно получить структуру зависимости между $X_1,...,X_p$. Таким образом, получается сжатое описание структуры зависимости, несущее почти всю информацию, содержащуюся в самих переменных.

<u>Суть метода главных компонент</u> состоит в том, что ищутся такие линейные комбинации исходных переменных

$$Y_1 = \sum_{j=}^{p} \alpha_{1j} X_j,...,Y_q = \sum_{j=1}^{p} \alpha_{pj} X_j, \qquad (1)$$

такие что

$$\text{cov}(Y_i Y_j) = 0, i, j = 1,..., p, i \neq j, \qquad (2)$$

$$V(Y_1) \geq V(Y_2) \geq ... \geq V(Y_p), \qquad (3)$$

$$\sum_{i=1}^{p} V(Y_i) = \sum_{i=1}^{p} \sigma_{ii}. \qquad (4),$$

где $\text{cov}(Y_i Y_j)$ - ковариация i-ой и j-ой компонент, $V(Y_j)$ - дисперсия j-й компоненты.

Из этих формул видно, что переменные $Y_1,...,Y_p$ не коррелированы и упорядочены по возрастанию дисперсии. Кроме того, общая дисперсия $V = \sum_{i=1}^{p} \sigma_{ii}$ после преобразования остается без изменений. Тогда подмножество первых q переменных Y_i будет объяснять большую часть общей дисперсии и, таким образом, получится сжатое описание структуры зависимости исходных переменных. *Метод главных компонент* состоит в определении коэффициентов $\alpha_{ij}, i, j = 1,..., p$.

Совместное распределение исходных переменных не обязательно считать многомерным нормальным. Однако такое предположение удобно, поскольку линейные комбинации нормально распределенных величин имеют в свою очередь нормальное распределение и, следовательно, полностью определяются параметрами μ и Σ. Тогда можно положить $\mu = (0,...,0)'$ и матрица Σ будет полностью описывать совместное распределение переменных $X_1,...,X_p$.

Пусть матрица Σ известна и $Y_1 = \alpha_{11}X_1 + ... + \alpha_{1p}X_p$. Требуется найти такие коэффициенты $\alpha_{11},...,\alpha_{1p}$, чтобы величина

$$V(Y_1) = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_{1i}\alpha_{1j}\sigma_{ij} \qquad (5)$$

была максимальной при $\sum_{j=1}^{p} \alpha_{1j}^2 = 1$. (Это условие обеспечивает единственность решения). Решение $\alpha_1 = (\alpha_{11},...,\alpha_{1p})'$ называется ***собственным вектором*** и соответствует максимальному ***собственному значению*** матрицы Σ. Это собственное значение равно $V(Y_1)$. Линейная комбинация $Y_1 = \alpha_{11}X_1 + ... + \alpha_{1p}X_p$ называется первой *главной компонентой* переменных $X_1,...,X_p$. Она объясняет $100V(Y_1)/V$ процентов общей дисперсии.

Положим $Y_2 = \alpha_{21}X_1 + ... + \alpha_{2p}X_p$. Надо найти такие коэффициенты $\alpha_{21},...,\alpha_{2p}$, чтобы величина

$$V(Y_2) = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_{2i}\alpha_{2j}\sigma_{ij} \qquad (6)$$

достигала максимального значения при условии $\sum_{j=1}^{p} \alpha_{2j}^2 = 1$ и $\text{cov}(Y_1, Y_2) = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_{1i}\alpha_{2j}\sigma_{ij} = 0$. Первое условие обеспечивает единственность решения, а второе – некоррелированность Y_1 и Y_2. Решение $\alpha_2 = (\alpha_{21},...,\alpha_{2p})'$ является *собственным вектором* матрицы Σ, соответствующим второму по величине *собственному значению,* а Y_2 является второй *главной компонентой* признаков $X_1,...,X_p$. Первые две главные компоненты объясняют $100[V(Y_1) + V(Y_2)]/V$ процентов общей дисперсии. После того, как получены $Y_1,...,Y_{q-1}, q = 2,...,p$, найдем переменную $Y_q = \sum_{j=1}^{p} \alpha_{qj}X_j$, такую, чтобы величина $V(Y_q) = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_{qi}\alpha_{qj}\sigma_{ij}$ достигла максимального значения при условии $\sum_{j=1}^{p} \alpha_{qj}^2 = 1$ и $\text{cov}(Y_m, Y_q) = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_{q1i}\alpha_{mj}\sigma_{ij} = 0, m = 1,...,q-1$.

В результате получим $\alpha_q = (\alpha_{q1},...,\alpha_{qp})'$ - *собственный вектор* матрицы Σ, соответствующий $q-$му по величине *собственному значению,* которое равно $V(Y_q)$. Таким образом Y_q является $q-$й *главной компонентой* признаков $X_1,...,X_p$. Переменные $Y_1,...,Y_q$ будут объяснять $100\sum_{i=}^{q} V(Y_i)/V$ процентов общей дисперсии.

Метод главных компонент допускает следующую геометрическую интерпретацию:

- вначале (при переходе от исходного вектора **X** к центрированному вектору **X=X−MX** фактически производится перенос начала координат в точку **MX**, являющуюся центром эллипсоида рассеяния случайного вектора **X**;
- затем производится поворот осей координат таким образом, чтобы новые оси координат *Of* (1), *Of* (2), …были направлены вдоль осей эллипсоида рассеяния. При этом разброс точек вдоль оси *Of* (1) должен быть не меньше, чем вдоль оси *Of* (2) и т. д. Разброс наблюдений вдоль новой оси *Of* (1) для исследователя наиболее важен, менее важен разброс вдоль оси *Of* (2), а разбросом вдоль нескольких последних осей можно пренебречь. Графически это иллюстрирует рисунок 1.

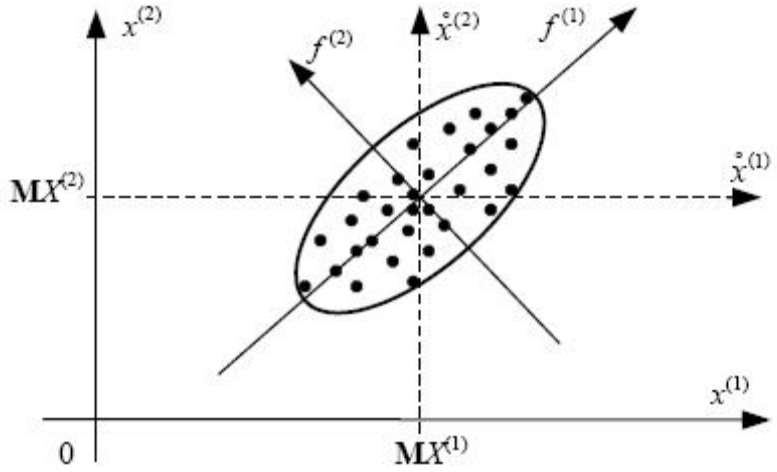

Рисунок 1- геометрическая иллюстрация метода главных компонент

Когда матрица Σ неизвестна, можно предположить, что имеется случайная выборка $X_1^{p\times 1},...,X_n^{p\times 1}$, по которой Σ оценивается выборочной ковариационной матрицей S. Для получения *оценок главных компонент* следует применить описанную выше процедуру к матрице S. В результате получаться оценки $a_{ij}, i, j = 1,..., p$ коэффициентов $\alpha_{ij}, i, j = 1,..., p$. Оценкой q-й главной компоненты будет вектор $Y_q = \sum_{j=1}^{p} a_{qj} X_j$, где $a_q = (a_{q1},...,a_{qp})'$ есть q-й собственный вектор матрицы S ($q = 1,..., p$).

1.3. ЭТАПЫ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ.

1. Подготовительный этап:
   1) Центрирование переменных – переход к $x_i^{(j)} - \overline{x}^{(j)}$

2) Нормирование переменных – переход к $\left(x_i^{(j)} - \bar{x}^{(j)}\right) \Big/ \sqrt{\sigma_{jj}}$. ***<u>Этот этап важен,</u>*** если показатели измеряются в различных единицах или имеется большой разброс значений. Если измерения проведены в одних единицах, то этот этап можно пропустить.

3) Вычисление матрицы ковариаций

$$\sum = \begin{pmatrix} \hat{\sigma}_{11} & \dots & \hat{\sigma}_{1p} \\ & \dots & \\ \hat{\sigma}_{p1} & \dots & \hat{\sigma}_{pp} \end{pmatrix},\ \hat{\sigma}_{kj} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{(k)} - \bar{x}^{(k)}\right)\left(x_i^{(j)} - \bar{x}^{(j)}\right) = KOBAP\left(x_1^{(k)},\dots,x_n^{(k)}; x_1^{(j)},\dots,x_n^{(j)}\right)$$

2. Решение характеристического уравнения $\left|\sum - \lambda E\right| = 0$:

1) Нахождение собственных чисел $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ – для симметричной положительно определенной матрицы $\sum$ все корни положительные вещественные числа

2) Нахождение для каждого корня характеристического уравнения λ_k собственного вектора $\left(\sum - \lambda_k E\right) l^{(k)} = 0$, $\left\| l^{(k)} \right\| = 1$, где Е- единичная матрица.

3. Переход к новым переменным $Z = XL$

$z^{(k)} = X\, l^{(k)}$, $k = 1,\dots,p'$ – главные компоненты.

1.4. ОПРЕДЕЛЕНИЕ КОЛИЧЕСТВА ГЛАВНЫХ КОМПОНЕНТ.

Как только получена информация о том, сколько дисперсии выделил каждый компонент, возникает вопрос – сколько новых переменных следует оставить. Это решение произвольно по своей природе. Существуют несколько критериев отбора.

<u>Критерий Кайзера.</u> Сначала вы можете отобрать только факторы, с собственными значениями, большими 1. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером и является, вероятно, наиболее широко используемым на практике.

<u>Критерий каменистой осыпи.</u> Критерий каменистой осыпи является графическим методом. При использовании этого критерия собственные значения представляются в виде графика (рис.2) и ищется точка максимального замедления убывания собственных значений. Так, исходя из рисунка 2, можно оставить две главных компоненты.

Как уже говорилось, выбор числа компонент достаточно произволен. Критерий Кайзера иногда сохраняет слишком много факторов, в то время как критерий каменистой осыпи иногда сохраняет слишком мало факторов. Оба критерия вполне хороши при

нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. На практике возникает важный дополнительный вопрос, а именно: когда полученное решение может быть содержательно интерпретировано. Поэтому обычно исследуется несколько решений с большим или меньшим числом факторов, и затем выбирается одно наиболее "осмысленное".

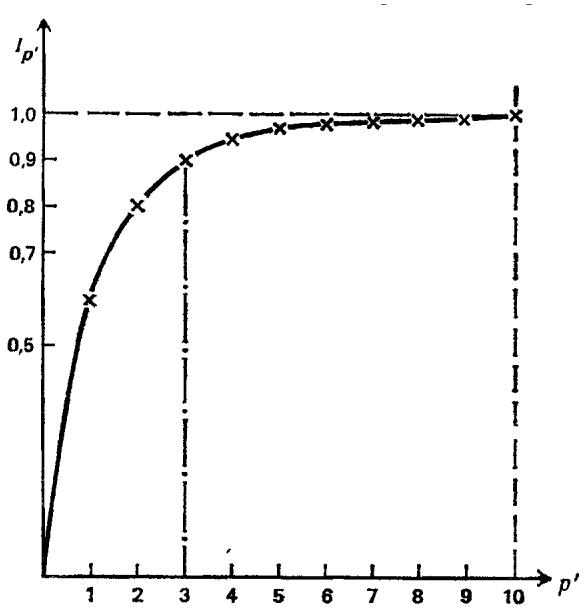

Рисунок 2- Зависимость суммарной дисперсии от числа главных компонент.

1.5. ПРИМЕР ФОРМИРОВАНИЯ ГЛАВНЫХ КОМПОНЕНТ.

1. **Исходные данные.**

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

2. Центрированные данные.

X	Y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

Исходные и центрированные данные приведены на рисунках 3 и 4 соответственно.

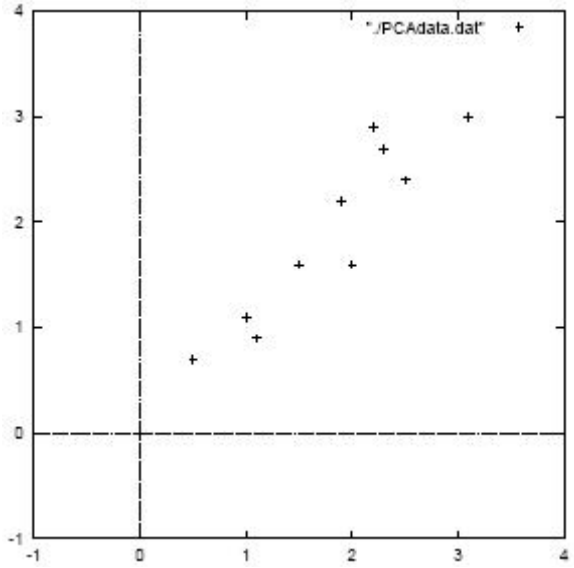

Рисунок 3- Исходные данные.

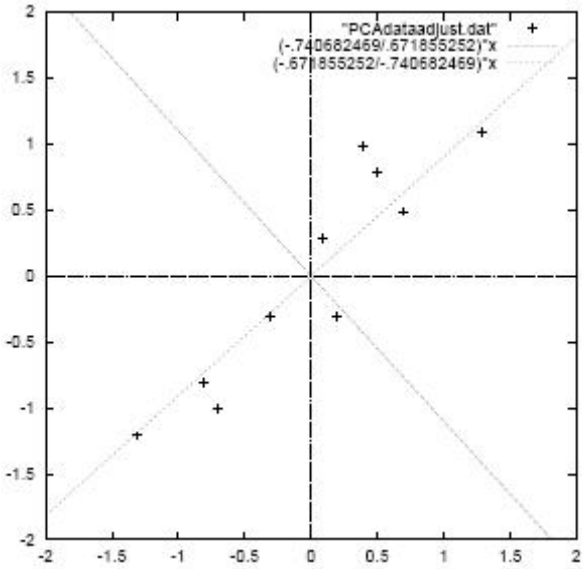

Рисунок 4- Центрированные данные.

Из рисунков 3 и 4 видно, что существует сильная корреляция между переменными X и Y.

3. Ковариционная матрица.

Команда A=cov (data) (команды пакета Matlab)

ans =

0.6166 0.6154

0.6154 0.7166

Получаем квадратную матрицу коэффициентов ковариации A.

4. Собственные числа.

eig(A)

ans =

0.0491

1.2840

Видно, что почти вся дисперсия объясняется одной компонентой.

5. Собственные векторы.

[d,c]=eigs(A)

d =

```
    0.6779   -0.7352
    0.7352    0.6779
```

c =

```
    1.2840         0
         0    0.0491
```

Матрица d- собственные векторы, c- соответствующие им собственные числа.

6. Выбор собственного вектора и формирование нового вектора данных.

Как уже отмечалось, собственные векторы, объясняющие малую часть дисперсии, отбрасываются и оставляются только собственные векторы, выделяющие большую часть дисперсии. В данном примере, очевидно, что таким вектором будет вектор с собственным числом 1.2840. Соответствующий вектор (0.6779, 0.7352). Тогда новый вектор данных находится как произведение матрицы центрированных данных на собственный вектор.

e=rot90(rot90(rot90(d)))

e = 0.7352 0.6779

0.6779 -0.7352

>> g= pca_example*e

g = 0.8394 0.1075

-1.7833 0.0016

0.9578 -0.4635

0.2627 -0.1522

1.6873 0.0731

0.8958 -0.2486

-0.0705 0.3567

-1.1446 0.0464

-0.4380 0.0178

-1.2066 0.2612

Тогда в новых координатах полученные данные будут иметь следующий вид (рис.5).

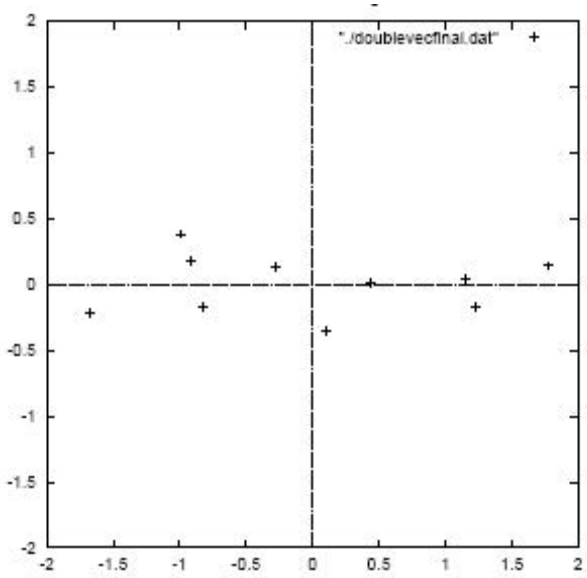

Рисунок 5- Исходные данные в новой системе координат.

Данные, соответствующие наиболее значимому собственному вектору:

-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056.

7.Обратное преобразование данных.

Получить исходные данные из преобразованных данных можно обратным преобразованием:

Исходные центрированные данные= транспонированный собственный вектор*преобразованный вектор данных.
g= g*e

g =

0.6900	0.4900
-1.3100	-1.2100
0.3900	0.9900
0.0900	0.2900
1.2900	1.0900
0.4900	0.7900
0.1900	-0.3100
-0.8100	-0.8100
-0.3100	-0.3100
-0.7100	-1.0100

Прибавив среднее значение, получим восстановленные данные (рис.6).

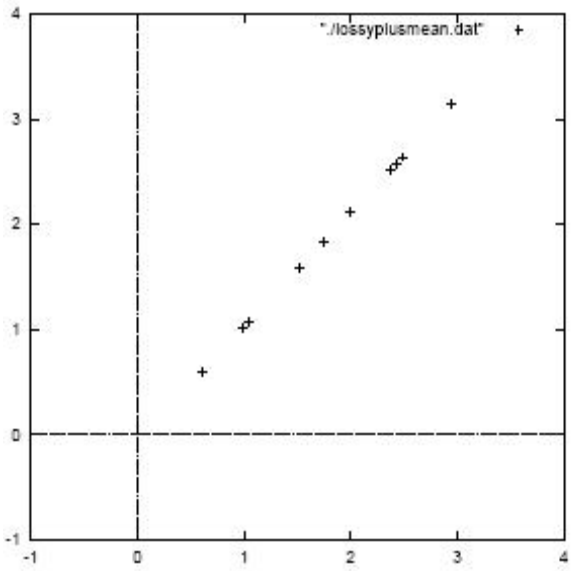

Рисунок 6- Восстановленные данные.

Фактически, данный пример показывает, что с помощью метода главных компонент была проведена низкочастотная фильтрация данных.

2.РЕАЛИЗАЦИЯ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ В ПАКЕТЕ MATLAB.

Получение главных компонент требует выполнения значительных по объему вычислений (см. приложение А). Особенно, если количество переменных, описывающих объект, достаточно большое. Так, если переменных 4, то для получения собственных чисел необходимо решать уравнение четвертой степени.

В пакете Matlab получение главных компонент интегрировано в функции [pcs,newdata,variances,t2] = princomp(A). Здесь А- матрица данных, pcs- главные компоненты, newdata- данные в новых координатах, variances- дисперсия новых данных, t2- статистическая мера, показывающая расстояние от каждого наблюдения до центра множества данных.

3. ОПИСАНИЕ ДАННЫХ И ЭКСПОРТ ДАННЫХ В ПАКЕТ MATLAB.

Данные, для выполнения лабораторной работы, находятся в файле rus_dasl24.xls. Таблица содержит 10 переменных VAR1-VAR10, расшифровка которых приведена ниже.

VAR1="'Пол" 1="МУЖ" 2="ЖЕН";
VAR2="'Возраст"

VAR3="'Группа наблюдения" 1-Контроль, 2-7 - различные патологии
VAR4="Период наблюдения" 1-До лечения, 2 - после лечения
VAR5="форсированная жизненная ёмкость легких,%
VAR6="объём форсированного выдоха, %
VAR7="пиковая объёмная скорость выдоха, %
VAR8="мгновенная объёмная скорость после выдоха 75% ФЖЕЛ,%
VAR9="мгновенная объёмная скорость после выдоха 50% ФЖЕЛ,%"
VAR10="мгновенная объёмная скорость после выдоха 25% ФЖЕЛ,%"

4. ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ.

1. Сформируйте 4 матрицы данных для переменных VAR2, VAR5-VAR10 объемом не менее 10 и не более 50 значений каждой переменной, однородные по признакам пола, заболевания и лечения. Например, матрица

VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10
1	33	1	1	123,8	111,8	90,1	92,9	103,5	101,6
1	40	1	1	127,5	115,4	94,5	97,2	96,1	107,9
1	42	1	1	119,6	108,2	89,7	88,9	98,9	94,8
1	39	1	1	122,3	111,4	90,4	92,8	99,9	101,4
1	40	1	1	126,4	115,5	89,7	96,9	100,5	106,8
1	45	1	1	120,2	109,5	89,4	88,6	97,8	95,3
1	39	1	1	125,6	111,2	90,5	92,7	95,1	101,3
1	30	1	1	121,5	110,4	89,4	95,8	101,6	105,7
1	43	1	1	119,8	110,5	88,8	89,1	100,6	96,7
1	45	1	1	127,4	111,9	90,7	92,6	99,8	101,8
1	25	1	1	123,4	109,5	88,5	94,7	95,1	104,8
1	45	1	1	123,9	109,7	90,4	90,7	97,9	97,3
1	43	1	1	124,7	111,7	90,8	92,4	103,6	101,9

имеет одинаковые значения переменных VAR1, VAR3 и VAR4 и, следовательно, удовлетворяет условиям однородности. После удаления столбцов с переменными VAR1, VAR3 и VAR4 получаем искомую матрицу

VAR2	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10
33	123,8	111,8	90,1	92,9	103,5	101,6
40	127,5	115,4	94,5	97,2	96,1	107,9
42	119,6	108,2	89,7	88,9	98,9	94,8
39	122,3	111,4	90,4	92,8	99,9	101,4
40	126,4	115,5	89,7	96,9	100,5	106,8
45	120,2	109,5	89,4	88,6	97,8	95,3
39	125,6	111,2	90,5	92,7	95,1	101,3
30	121,5	110,4	89,4	95,8	101,6	105,7
43	119,8	110,5	88,8	89,1	100,6	96,7
45	127,4	111,9	90,7	92,6	99,8	101,8
25	123,4	109,5	88,5	94,7	95,1	104,8
45	123,9	109,7	90,4	90,7	97,9	97,3
43	124,7	111,7	90,8	92,4	103,6	101,9

2. Выполните центрирование и нормирование переменных для 4-х полученных матриц. Среднее можно вычислить с помощью команды mean(x), среднеквадратичное отклонение- std(x).

3. Для 4-х матриц, полученных в пункте 2, постройте график зависимости переменной VAR6 от VAR5.

4. С помощью функции [pcs, newdata,variances,t2] = princomp(A) получите значения переменных pcs, newdata,variances для 4-х матриц, полученных в пункте 2.

5. Постройте новые полученные данные для 4-х матриц в координатах первых двух главных компонент.

6. С помощью выражения percent_explained = 100*variances/sum(variances) оцените, сколько главных компонент необходимо для описания объектов в матрицах 1-4.

7. С помощью последовательности команд

pareto(percent_explained)

xlabel('Principal Component')

ylabel('Variance Explained (%)')

постройте графики зависимости дисперсии от числа главных компонент для 4-х матриц.

5. СОДЕРЖАНИЕ ОТЧЕТА

1. Наименование и цель работы.
2. Матрицы, полученные в п.1.
3. Матрицы, полученные в п.2.
4. Графики, полученные в п.3.
5. Данные, полученные в п.4.
6. Графики, полученные в п.5.
7. Результаты п.6.
8. Графики п.7.
9. Выводы. **Примечание: выводы должны отражать полученные результаты и быть связаны с целями, для достижения которых предназначены методы снижения размерности пространства признаков. Выводы должны быть сделаны по п.4-п.7 содержания отчета.**

6. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Перечислите основные цели, для достижения которых используются методы снижения пространства признаков.
2. За счет чего можно сократить размерность пространства признаков?
3. По каким критериям отбираются или формируются новые признаки?
4. Чем определяются критерии оптимальности при формировании новых признаков?
5. Перечислите основные предпосылки, позволяющие снизить размерность пространства признаков.
6. Какие методы используются для снижения размерности пространства признаков?
7. В чем состоит метод главных компонент?
8. При каких условиях ищется линейная комбинация исходных переменных в методе главных компонент?
9. Что такое собственные векторы и собственные значения матрицы?
10. Поясните, как можно последовательно найти собственные векторы и собственные значения?

11. Сколько процентов дисперсии объясняет первая главная компонента? Приведите выражение.

12. Дайте геометрическую интерпретацию метода главных компонент.

13. Перечислите этапы реализации метода главных компонент.

14. В каких случаях необходимо нормирование переменных?

15. С помощью каких критериев можно выбрать число главных компонент?

16. Как по заданной матрице найти собственные числа и векторы?

СПИСОК ЛИТЕРАТУРЫ

1. Половко А. М., Бутусов П. Н. MATLAB для студента. — СПб.: БХВ-Петербург, 2005. - 320 с.

2. О.Ю. Реброва. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. М., МедиаСфера, 2003. 312с.

3. Новиков Д.А., Новочадов В.В. Статистические методы в медико-биологическом эксперименте (типовые случаи). Волгоград: Издательство ВолГМУ, 2005. – 84 с.

ПРИЛОЖЕНИЕ А

Пример нахождения собственных векторов и собственных значений

Пусть задана квадратная матрица $A=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, X – некоторая матрица–столбец, высота которой совпадает с порядком матрицы A. $X=\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

Во многих задачах приходится рассматривать уравнение относительно X

$$A \cdot X = \lambda \cdot X,$$

где λ – некоторое число. Понятно, что при любом λ это уравнение имеет нулевое решение $X=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}=0$.

Число λ, при котором это уравнение имеет ненулевые решения, называется *собственным значением* матрицы A, а X при таком λ называется *собственным вектором* матрицы A.

Найдём собственный вектор матрицы A. Поскольку $E \cdot X = X$, то матричное уравнение можно переписать в виде $A \cdot X = \lambda \cdot E \cdot X$ или $(A - \lambda \cdot E)X = 0$. В развёрнутом виде это

уравнение можно переписать в виде системы линейных уравнений. Действительно

$$A-\lambda\cdot E=\begin{pmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{pmatrix}.$$

И, следовательно, $\begin{cases} (a_{11}-\lambda)x_1+a_{12}x_2+a_{13}x_3=0, \\ a_{21}x_1+(a_{22}-\lambda)x_2+a_{23}x_3=0, \\ a_{31}x_1+a_{32}x_2+(a_{33}-\lambda)x_3=0. \end{cases}$

Итак, получили систему однородных линейных уравнений для определения координат x_1, x_2, x_3 вектора X. Чтобы система имела ненулевые решения необходимо и достаточно, чтобы определитель системы был равен нулю, т.е.

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{vmatrix}=0 \qquad \textit{или} \qquad \left|A-\lambda\cdot E\right|=0.$$

Это уравнение 3-ей степени относительно λ. Оно называется *характеристическим уравнением* матрицы A и служит для определения собственных значений λ.

Каждому собственному значению λ соответствует собственный вектор X, координаты которого определяются из системы при соответствующем значении λ.

Учебное издание

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ СНИЖЕНИЯ РАЗМЕРНОСТИ БИОМЕДИЦИНСКИХ ДАННЫХ

Методические указания

Составитель: Конюхов Вадим Николаевич

Самарский государственный аэрокосмический университет
имени академика С.П. Королёва.
443086 Самара, Московское шоссе, 34