1 Методы кластерного анализа

2 Иерархические алгоритмы

3 Расстояния между кластерами

Возможные расстояния:

1. **Расстояние "ближайшего соседа"** (одиночная связь):

$$\rho_{\min}(K_i, K_j) = \min_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно расстоянию между самыми ближними объектами кластеров.

2. **Расстояние "дальнего соседа"** (полная связь):

$$\rho_{\max}(K_i, K_j) = \max_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно расстоянию между самыми дальними объектами кластеров.

3. **Невзвешенное попарное среднее**:

$$\rho_{\text{mean}}(K_i, K_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно среднему между всеми парами расстояний.

4. **Взвешенное попарное среднее**:

$$\rho_{\text{mean2}}(K_i, K_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \rho(k_1 \cdot x_i, k_2 \cdot x_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \frac{k_1 x_i + k_2 x_j}{k_1 + k_2}.$$

   Равно среднему между всеми парами расстояний с учетом весов k_1, k_2, равных ёмкости кластеров.

5. **Незвешенный центроидный метод**: расстояние между кластерами равно расстоянию между их центрами тяжести, где центр тяжести есть среднее арифметическое всех объектов в кластере:

$$x_c = \frac{\sum_{i=1}^{k} 1 \cdot x_i}{k} = \text{mean}_{x_i \in K_i}(x_i)$$

6. **Взвешенный центроидный метод**. Как я понял, это расстояние между центрами, но с учетом весов, как в пункте 4. Но вообще начиная с пункта 4 нигде нет формул, так что не факт, что они у меня правильные.

7. **Метод Варда**. Целевой функцией является внутригрупповая сумма квадратов:

$$\text{SS} = \sum_i \sum_{x_j \in K_i} \rho(x_j, x_i),$$

   где сумма берётся по всем кластерам, x_i – среднее по кластеру K_i. Объединение в кластеры на каждой итерации происходит так, чтобы увеличение этой функции было минимальным.

4 Процедуры эталонного типа

Наряду с иерархическими методами кластеризации существуют итеративные (k-средних), суть которых заключается в следующем (возможны модификации):

1. Среди объектов x_i некоторым образом выбираются k штук – центры будущих кластеров.
2. Объекты, не отнесённые к какому-либо кластеру, приписываются тому кластеру, до которого будет наименьшее расстрояние.
3. Центры кластеров пересчитываются.
4. Для всех точек пересматривается их принадлежность к кластеру.
5. Пункты 3-4 повторяются, пока точки могут менять принадлежность.

На этом сайте хорошо показано, как работает метод k-средних: `https://www.naftaliharris.com/blog/visualizing-k-means-clustering/`

5 Методика дискриминантного анализа

Основной целью дискриминации является нахождение такой линейной комбинации переменных (дискриминантных переменных), которая бы оптимально разделила рассматриваемые группы. Линейная функция:

$$d_{km} = \beta_0 + \sum_{i=1}^{p} \beta_i x_{ikm}, m = 1, \ldots, n \tag{1}$$

называется **канонической дискриминантной функцией** с неизвестными коэффициентами. Здесь d – значение дискриминантной функции для m-го объекта в группе k. С геометрической точки зрения дискриминантные функции определяют гиперповерхности в p-мерном пространстве. В частном случае при $p = 2$ она является прямой, а при $p = 3$ – плоскостью.

Коэффициенты первой канонической дискриминантной функции выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично определяются и другие функции. Если число групп равно g, то число канонических дискриминантных функций будет на единицу меньше числа групп. Однако по многим причинам практического характера полезно иметь одну, две или же три дискриминантных функций. Тогда графическое изображение объектов будет представлено в одно-, двух- и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных переменных p велико по сравнению с числом групп g.

Рассмотрим этапы ДА более подробно На первом этапе проведения ДА изучаются межгрупповые различия и ищутся ответы на следующие вопросы: возможно ли, используя заданный набор дискриминантных переменных,отличить одну группу от другой; насколько хорошо эти переменные позволяют провести различение; какие из них наиболее информативны. Одним из способов отбора информативных дискриминантных переменных является пошаговый дискриминантный анализ. Логика пошагового ДА такова: вначале определяется та переменная, для которой средние значения в априорно заданных группах "наиболее различны". На каждом следующем шаге рассматриваются условные распределения оставшихся переменных и определяется та, для которой средние значения в группах «наиболее различны». Процесс завершается, когда ни одна из оставшихся переменных не вносит значимого вклада в различение групп (статистика Уилкса, расстояние Махаланобиса и др.). Процедура пошагового ДА предполагает также проверку (в начале каждого шага) всех дискриминантных переменных на соответствие двум условиям: необходимой точности вычисления (толерантности) и превышению заданного уровня различения (с этой целью используют статистики F-ввода

и F-исключения). Статистика F-ввода оценивает улучшение различения благодаря использованию данной переменной по сравнению с различением, достигнутым с помощью уже отобранных переменных. Статистика F-исключения определяет значимость ухудшения различения после удаления переменной из списка уже отобранных переменных. На заключительном шаге статистика F-исключения может быть использована для оценки дискриминантных возможностей отобранных переменных. Переменная с наибольшим значение F-исключения дает наибольший вклад в различение, достигнутое посредством других переменных.

На втором этапе проведения ДА отобранное подмножество наиболее информативных переменных используется для вычисления дискриминантных функций. Исследователь получает одну или несколько дискриминантных функций.

Далее определяется все ли вычисленные канонические дискриминантные функции полезны для описания межгрупповых различий. С этой целью используется собственное значение и относительное процентное содержание вычисленных функций (% объясненной дисперсии), коэффициент канонической корреляции, а также тест равенства средних значений канонических дискриминантных функций в группах. На относительный вклад отдельных дискриминантных переменных в значение каждой дискриминантной функции указывают стандартизированные коэффициенты. Соответственно, чем больше стандартизированный коэффициент, тем больше вклад переменной. Самым лучшим показателем информативности отобранных дискриминантных переменных и полезности применения дискриминантной функции для интерпретации межгрупповых различий является, конечно, процент правильно распознанных объектов с использованием вычисленных дискриминантных функций (классификация с учителем). Число правильно распознанных новых объектов (как в целом, так и по отдельным группам) свидетельствует о соответствии дискриминантной модели эмпирическим данным.

6 Что характеризует Лямбда Уилкса?

Лямбда Уилкса — это отношение меры внутригрупповой изменчивости к мере общей изменчивости. Внутригрупповая изменчивость — часть общей, и это означает, что лямбда Уилкса может принимать значения от 0 (группы полностью однородны) до 1 (разделение объектов на группы не приводит к тому, что внутригрупповая изменчивость оказывается меньше общей). Итого, чем меньшее значение имеет лямбда Уилкса, тем лучшим оказывается разделение на группы при дискриминантном анализе.

7 Что показывают квадраты расстояний Махаланобиса?

8 Какое максимальное число канонических дискриминантных функций допустимо в дискриминантном анализе?

9 Какую информацию дают стандартизованные и структурные коэффициенты дискриминантной функции?

Стандартизованные коэффициенты – это то же самое, что и нестандартизованные, но полученные по стандартизованным начальным данным. *Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных: если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.*

Структурные коэффициенты определяются коэффициентами взаимной корреляции между отдельными переменными и дискриминантной функцией. Если относительно некоторой переменной абсолютная величина коэффициента велика, то вся информация о дискриминантной функции заключена в этой переменной.

Структурные коэффициенты по своей информативности несколько отличаются от стандартизованных коэффициен-тов. *Стандартизованные коэффициенты показывают вклад переменных в значение дискриминантной функции.* Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных. Такое распределение величины стандартизованного коэффициента объясняется тем, что при их вычислении учитывается влияние всех переменных. *Структурные же коэффициенты являются парными корреляциями и на них не влияют взаимные зависимости прочих переменных.*

10 Опишите процедуру отбора переменных с помощью стандартизованных и структурных коэффициентов

Никакой инфы про эту процедуру "отбора переменных с помощью стандартизованных и структурных коэффициентов"я не нашёл. Структурные коэффициенты вообще показывают корреляцию между переменной и дискриминантной функцией, что позволяет *интепретировать* результаты, а не отбирать переменные. Можете сами попробовать поискать что-то про это в копии её методички: `http://masters.donntu.org/2005/kita/kapustina/library/discr_an.htm`.

Про стандартизованные коэффициенты есть такая инфа: "В рамках алгоритмического подхода список информативных переменных формируется автоматически в соответствии с тем или иным алгоритмом пошагового дискриминантного анализа. Предусмотрены два его варианта: «Forward stepwise» и «Backward stepwise» — метод последовательного пополнения и метод последовательного исключения списка информативных признаков. В первом случае выбирается переменная, вносящая наибольший вклад в межгрупповые различия, и далее к ней на каждом шаге анализа присоединятся другие информативные переменные. Метод исключения основан на альтернативной процедуре: из полного списка признаков на каждом шаге анализа последовательно исключаются малозначимые переменные."При этом величина вклада переменной в различия определяется как раз величиной стандартизованного коэффициента.

Пошаговый дискриминантный анализ.

Вероятно, наиболее общим применением дискриминантного анализа является включение в исследование многих переменных с целью определения тех из них, которые наилучшим образом разделяют совокупности между собой.

Модель. Другими словами, вы хотите построить "модель позволяющую лучше всего предсказать, к какой совокупности будет принадлежать тот или иной образец. В следующем рассуждении термин "в модели"будет использоваться для того, чтобы обозначать переменные, используемые в предсказании принадлежности к совокупности; о неиспользуемых для этого переменных будем говорить, что они "вне модели".

Пошаговый анализ с включением. В пошаговом анализе дискриминантных функций модель дискриминации строится по шагам. Точнее, на каждом шаге просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

Пошаговый анализ с исключением. Можно также двигаться в обратном направлении, в этом случае все переменные будут сначала включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в предсказания. Тогда в качестве результата успешного анализа можно сохранить только "важные"переменные в модели, то есть те переменные, чей вклад в

дискриминацию больше остальных.

11 Какова интерпретация канонического коэффициента корреляции?

Общее число дискриминантных функций не превышает числа дискриминантных переменных и, по крайней мере, на единицу меньше числа групп. Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу λ_i, вторая обеспечивает максимальное различение после первой и т. д. Различительную способность i-й функции оценивают по относительной величине в процентах собственного числа λ_i от суммы всех λ.

Коэффициент канонической корреляции. Другой характеристикой, позволяющей оценить полезность дискриминантной функции является коэффициент канонической корреляции r_i. Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1. Будем считать, что группы составляют одно множество, а другое множество образуют дискриминантные переменные. Коэффициент канонической корреляции для i-й дискриминантной функции определяется формулой:

$$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}.$$

Чем больше величина r_i, тем лучше разделительная способность дискриминантной функции.

12 В каком случае учет априорных вероятностей может сильно изменить результаты классификации?

Априорные вероятности оказывают наибольшее влияние при перекрытии групп и, следовательно, многие объекты с большой вероятностью могут принадлежать ко многим группам. Если группы сильно различаются, то учет априорных вероятностей практически не влияет на результат классификации, поскольку между классами будет находиться очень мало объектов

13 Методика факторного анализа

Факторный анализ – раздел статистического многомерного анализа, объединяющий методы оценки размерности множества наблюдаемых переменных посредством исследования структуры ковариационных или корреляционных матриц.

Факторный анализ (ФА) представляет собой совокупность методов, которые на основе реально существующих связей анализируемых признаков, связей самих наблюдаемых объектов, позволяют выявлять скрытые (неявные, латентные) обобщающие характеристики организационной структуры и механизма развития изучаемых явлений, процессов. Основное предположение факторного анализа заключается в том, что корреляционные связи между большим количеством наблюдаемых переменных можно объяснить существованием меньшего числа гипотетических переменных или факторов, называемых общими скрытыми факторами. Далее их будем называть просто факторами. Факторный анализ может быть:

- разведочным — он осуществляется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузках;

- конфирматорным, предназначенным для проверки гипотез о числе факторов и их нагрузках.

Фактор – латентная переменная, конструируемая таким образом, чтобы можно было объяснить корреляцию между набором некоторых имеющихся переменных. **Нагрузка** – корреляция между исходной переменной и фактором.

Общей моделью факторного анализа служит следующая линейная зависимость:

$$x_i = \sum_{j=1}^{m} a_{ij} F_j + b_i U_i + \varepsilon_i, i = 1, \ldots, k, \tag{2}$$

где F_j – общие факторы, x_i – наблюдаемые переменные, U_i – характерные факторы, ε_i – случайные ошибки.

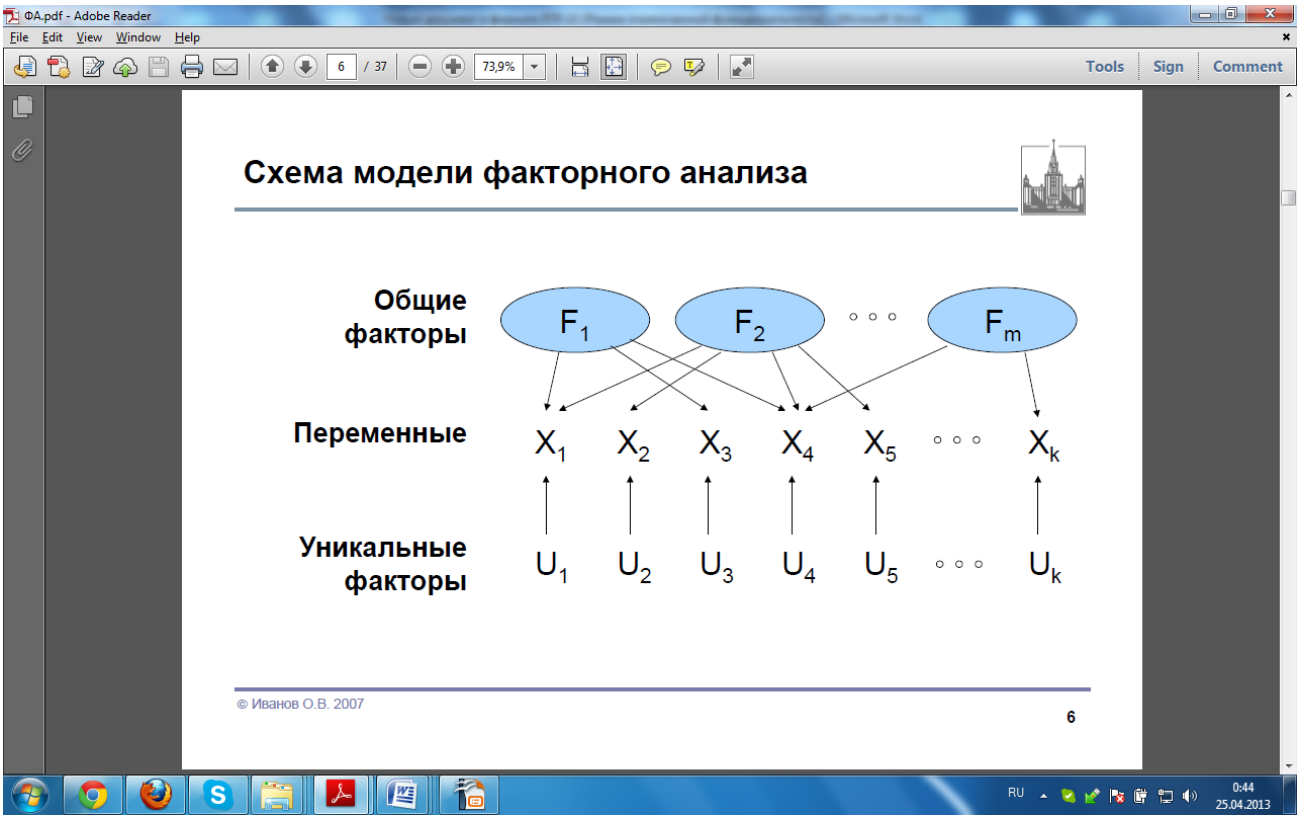


Рис. 1: Схема

Главными целями факторного анализа являются:

- сокращение числа переменных (редукция данных);
- определение структуры взаимосвязей между переменными, т.е. классификация переменных.

Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят:

- все признаки должны быть количественными;
- число наблюдений должно быть не менее чем в два раза больше числа переменных;
- выборка должна быть однородна.

Для проведения факторного анализа предлагается методика, включающая в себя следующие этапы:

1. сбор исходных статистических данных и подготовка корреляционной (ковариационной) матрицы;

2. выделение общих скрытых факторов;

3. вращение факторной структуры;

4. содержательная интерпретация результатов факторного анализа.

Алгоритмы факторного анализа основываются на использовании редуцированной матрицы парных корреляций (ковариаций). **Редуцированная матрица** – это матрица парных коэффициентов корреляции, на главной диагонали которой расположены не единицы (оценки) полной корреляции или оценки полной дисперсии, а их редуцированные, несколько уменьшенные величины – значения оценок общностей. Общность характеризует вклад данного признака в суммарную общность процесса При этом постулируется, что в результате анализа будет объяснена не вся дисперсия изучаемых признаков (объектов), а ее некоторая часть, обычно большая. Оставшаяся необъясненная часть дисперсии — это характерность, возникающая из-за специфичности наблюдаемых объектов, или ошибок, допускаемых при регистрации явлений, процессов, т.е. ненадежности вводных данных.

14 Суть задачи вращения общих факторов

Задача вращения общих факторов решается с целью улучшения их *интерпретируемости*. Производится попытка достижения простой структуры, в которой каждая переменная характеризуется преобладающим влиянием какого–то одного фактора. Факторные нагрузки могут быть изображены в виде диаграммы рассеяния, на которой каждая переменная представлена точкой. Можно повернуть оси в любом направлении без изменения относительного положения точек. При этом действительные координаты точек, то есть факторные нагрузки, изменяются.

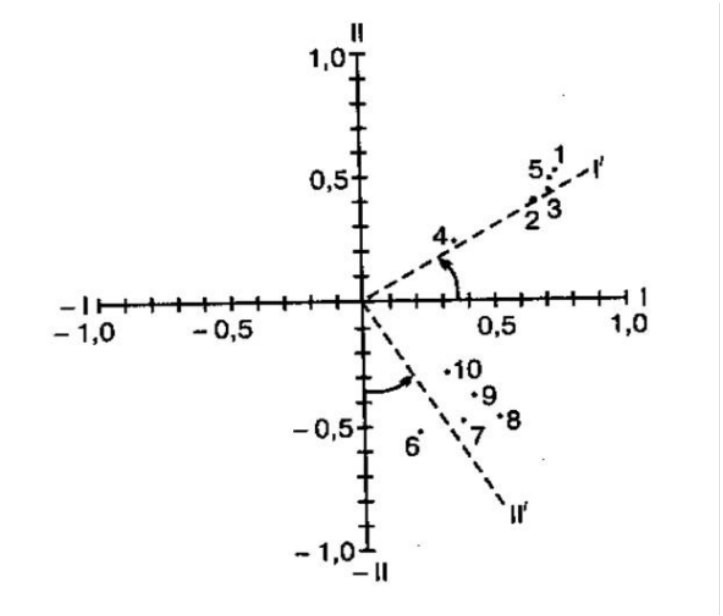


Рис. 2: Вращение факторов

Существуют различные методы вращения факторов. Целью этих методов является получение понятной (интерпретируемой) матрицы нагрузок, то есть факторов, которые ясно отмечены высокими нагрузками для некоторых переменных и низкими – для других.

Примеры методов вращения:

- **Варимакс**. Ортогональный метод вращения, минимизирующий число переменных с высокими нагрузками на каждый фактор.
- **Метод косоугольного (неортогонального) вращения**. Самое косоугольное решение соответствует дельте, равной 0 (по умолчанию). По мере того, как дельта отклоняется в отрицательную сторону, факторы становятся более ортогональными.
- **Квартимакс**. Метод вращения, который минимизирует число факторов, необходимых для объяснения каждой переменной.
- **Эквимакс**. Метод вращения, объединяющий методы варимакс, упрощающий факторы, и квартимакс, упрощающий переменные. Минимизируется число переменных с большими факторными нагрузками и число факторов, требуемых для объяснения переменной.
- **Промакс-вращение**. Косоугольное вращение в предположении, что факторы могут коррелировать между собой. Оно производится быстрее, чем вращение типа косоугольного вращения, поэтому оно полезно для больших наборов данных.

Примечание. КОСОУГОЛЬНОЕ ВРАЩЕНИЕ — такая трансформация факторного пространства, которая предполагает возможность проведения факторных осей под углами друг к другу, отличающимися от 90 градусов (от ортогональности)

15 Критерий Кайзера

Критерий Кайзера или критерий собственных чисел: отбираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером (Kaiser, 1960), и является, вероятно, наиболее широко используемым.

16 Критерий Каменистой осыпи

Критерий каменистой осыпи является графическим методом, где собственные значения представляются в виде простого графика. *Необходимо найти такое место на графике, где убывание собственных значений слева направо максимально замедляется.* Справа от этой точки находится, по-видимому, только "факторная осыпь"; "осыпь— это геологический термин для обломков, которые скапливаются в нижней части каменистого склона.

По графику видно, что только три фактора влияют на анализ. Их и необходимо оставить. Но иногда такое количество факторов не покрывает необходимый процент выборки и тогда берется больше факторов.

Рис. 3: График собственных значений общих факторов

17 Метод главных компонент

Метод главных компонент — это технология многомерного статистического анализа, используемая для сокращения размерности пространства признаков с минимальной потерей полезной информации, которая представляет собой ортогональное линейное преобразование, которое отображает данные из исходного пространства признаков в новое пространство меньшей размерности.

При этом первая ось новой системы координат строится таким образом, чтобы дисперсия данных вдоль неё была бы максимальна. Вторая ось строится ортогонально первой так, чтобы дисперсия данных вдоль неё, была бы максимальной их оставшихся возможных и т.д. Первая ось называется первой главной компонентой, вторая – второй и т. д.

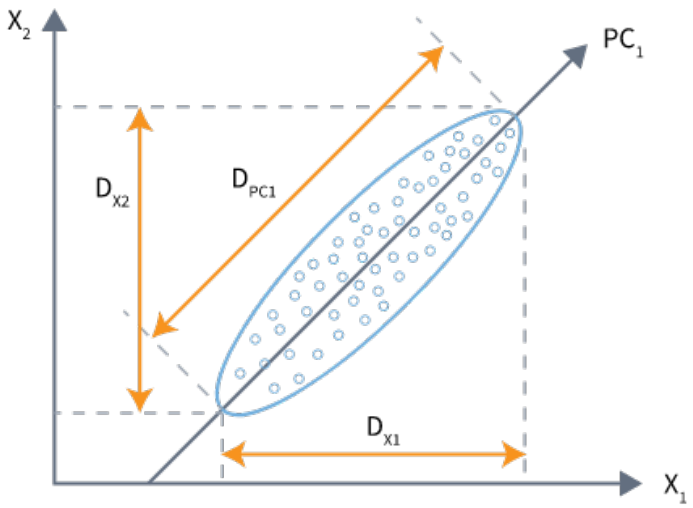
На рисунке показано снижение размерности исходного 2-мерного пространства (X1,X2) с помощью метода главных компонент до 1-мерного. Первая главная компонента PC1 ориентирована вдоль направления наибольшей вытянутости эллипсоида рассеяния точек объектов исходного набора данных в пространстве признаков, т.е. с ней связана наибольшая дисперсия.

На рисунке, также, несложно увидеть, что проекция дисперсии данных на ось первой главной компоненты DPC1, больше, чем её проекции на исходные оси DX1 и DX2, но меньше их суммы. Т.е. с помощью первой главной компоненты выразить всю дисперсию данных не удалось. Поэтому строят вторую, третью и т.д. главные компоненты, пока они суммарно не отразят (почти) всю дисперсию.

Смысл метода заключается в том, что с каждой главной компонентой связана определённая доля общей дисперсии исходного набора данных (её называют нагрузкой). В свою очередь, дисперсия, являющаяся мерой изменчивости данных, может отражать уровень их информативности.

Задача метода главных компонент заключается в том, чтобы построить новое пространство признаков меньшей размерности, дисперсия между осями которой будет перераспределена так, чтобы максимизировать дисперсию по каждой из них. Для этого выполняется последовательность следующих действий:

1. Вычисляется общая дисперсия исходного пространства признаков. Это нельзя сделать простым



суммированием дисперсий по каждой переменной, поскольку они, в большинстве случаев, не являются независимыми. Поэтому суммировать нужно взаимные дисперсии переменных, которые определяются из ковариационной матрицы.

2. Вычисляются собственные векторы и собственные значения ковариационной матрицы, определяющие направления главных компонент и величину связанной с ними дисперсии.

3. Производится снижение размерности. Диагональные элементы ковариационной матрицы показывают дисперсию по исходной системе координат, а её собственные значения – по новой. Тогда разделив дисперсию, связанную с каждой главной компонентой на сумму дисперсий по всем компонентам, получаем долю дисперсии, связанную с каждой компонентой. После этого отбрасывается столько главных компонент, чтобы доля оставшихся составляла 80-90%.

Основными ограничениями метода главных компонент являются:

- невозможность смысловой интерпретации компонент, поскольку они "вбирают"в себя дисперсию от нескольких исходных переменных;

- метод может работать только с непрерывными данными.