**МЕТОД ГЛАВНЫХ КОМПОНЕНТ**

Метод главных компонент осуществляет переход к новой совокупности некоррелированных признаков , каждый из которых является линейной комбинацией исходных признаков . При этом линейные комбинации выбираются таким образом, что среди всех возможных линейных нормированных комбинаций исходных признаков первая главная компонента  обладает наибольшей дисперсией. Геометрически это выглядит как ориентация новой координатной оси  вдоль направления наибольшей вытянутости эллипсоида рассеивания исследуемой выборки в пространстве признаков . Вторая главная компонента имеет наибольшую дисперсию среди всех оставшихся линейных преобразований, некоррелированных с первой главной компонентой. Она интерпретируется как направление наибольшей вытянутости эллипсоида рассеивания, перпендикулярное первой главной компоненте. Следующие главные компоненты определяются по аналогичной схеме.

В дальнейшем из полученных величин можно оставить только  наиболее значимых факторов, вносящих максимальный вклад в суммарную дисперсию и использовать эти величины как некие интегральные факторы, характеризующие всю совокупность признаков.

Пусть  – центрированная многомерная случайная величина с матрицей ковариаций . Положим

 , (1)

где  – векторы неизвестных коэффициентов преобразования, , . Будем называть величины  главными компонентами (факторами). Так как величины  центрированы, то , а дисперсия главных компонент определяется как:

 (2)

На коэффициенты преобразования  накладываем условиям нормировки:

. (3)

Задача отыскания главных компонент сводится к задаче на собственные значения и собственные векторы матрицы ковариаций . Коэффициенты преобразования ,  удовлетворяющие условиям задачи поиска главных компонент являются собственными векторами матрицы ковариаций , соответствующими собственные значения  упорядоченным по убыванию значений, причем , а .

Оценка главных компонент на основе выборочных данных строится на основе выборочной матрицы ковариаций. Оценки собственных значений, являющиеся собственными числами выборочной матрицы ковариаций, в случае нормального распределения генеральной совокупности являются оценками максимального правдоподобия. Если единицы измерения исходных признаков различаются или их значения сильно различаются, то лучше использовать при нахождении оценок главных компонент вместо выборочной матрицы ковариаций выборочную корреляционную матрицу.

Зададимся вопросом, какие из компонент можно отбросить, чтобы уменьшить размерность вектора признаков . Заметим, что математически строгих критериев отбора не существует. Обычно используют один из следующих эвристических методов. Во первых, зная , можно выбрать те компоненты  из общего набора, которые бы объясняли не менее некоторой заданной доли  суммарной доли дисперсии признаков. Обычно  полагают не менее 0,7. Другим критерием отбора является критерий Кайзера, который предполагает использование для нахождения оценок собственных значений выборочной матрицы корреляций (можно применять и в случае использования матрицы ковариаций, следует лишь в этом случае нормировать (умножить) все собственные значения на величину ). Согласно данному критерию оставляют только те главные компоненты, дисперсия которых больше 1. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается.

Подход к оценке числа главных компонент по необходимой доле объяснённой дисперсии формально применим всегда, однако неявно он предполагает, что нет разделения на «сигнал» и «шум», и любая заранее заданная точность имеет смысл. Поэтому часто более продуктивна иная эвристика, основывающаяся на гипотезе о наличии «сигнала» (сравнительно малая размерность, относительно большая амплитуда) и «шума» (большая размерность, относительно малая амплитуда). С этой точки зрения метод главных компонент работает как фильтр: сигнал содержится, в основном, в проекции на первые главные компоненты, а в остальных компонентах пропорция шума намного выше. Вопрос, как оценить число необходимых главных компонент, если отношение «сигнал/шум» заранее неизвестно? Одним из наиболее популярных подходов является правило сломанной трости (англ. Broken stick model). Набор нормированных собственных чисел (, ) сравнивается с распределением длин обломков трости единичной длины, сломанной в  случайно выбранной точке (точки разлома выбираются независимо и равномерно распределенными по длине трости). Пусть  () - длины полученных кусков трости, занумерованные в порядке убывания длины: . Тогда математическое ожидание : . По правилу сломанной трости -й собственный вектор (в порядке убывания собственных чисел ) сохраняется в списке главных компонент, если . К подобным критериям относится также графический критерий каменистой осыпи Кэттелла. Критерий каменистой осыпи состоит в поиске точки, где убывание собственных значений замедляется наиболее сильно. Справа от этой точки должна находится, по-видимому, только "факторная осыпь" ("осыпь" - это геологический термин для обломков, которые скапливаются в нижней части каменистого склона). Таким образом, число выделенных факторов не должно превышать количество факторов слева от этой точки.

**МЕТОД ГЛАВНЫХ КОМПОНЕНТ В ПАКЕТЕ STATISTICA**

Для анализа главных компонент запускаем в головном меню модуль «Statistics» и в стартовой панели выбираем пункт «Multivariate Exploratory Techniques» - многомерные разведочные методы и далее - «Principal Components & Classification Analysis» - главные компоненты и классификационный анализ (рис 1.)

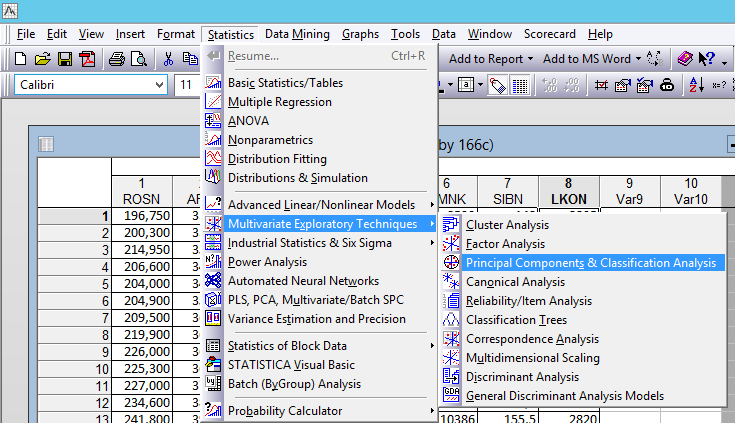


Рис.1 Запуск модуля анализа главных компонент.

В появившемся окне (рис. 2) на вкладке «Quick» выбираем переменные для анализа (рис 3) .

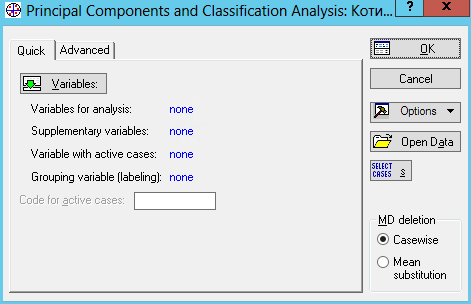


Рис. 2. Модуль анализа главных компонент, вкладка «Quick»

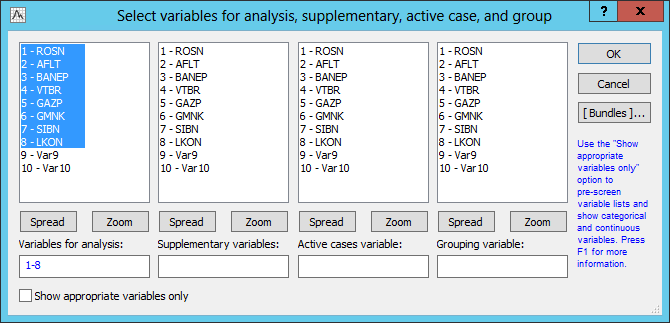


Рис 3. Выбор переменных для анализа

На вкладке «Advanced» (рис. 4) определяем на основе какой из матриц (ковариационной или корреляционной) будет проводиться анализ, а также условие нормировки для суммы квадратов отклонений, используемое при расчете выборочной матрицы ковариаций. Поскольку значения исходных переменных различаются сильно, анализ будем проводить на основе матрицы корреляций.

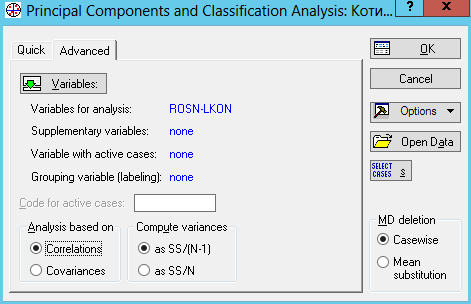


Рис 4. Выбор матрицы, на основе которой будет проводиться анализ

Нажимаем на «OK» и попадаем в модуль результатов анализа главных компонент – рис. 5. В заголовке модуля можно установить число используемых главных компонент (параметр «Number of factors»). На начальном этапе анализа можно использовать все компоненты, число которых равно числу исходных признаков.

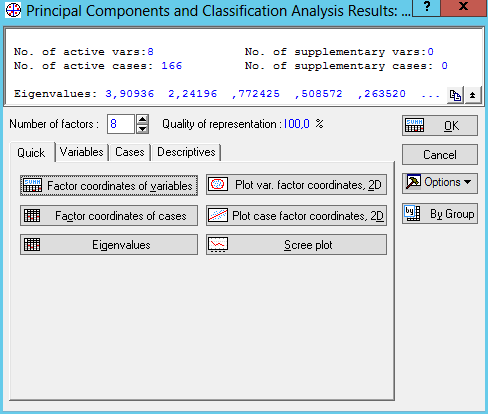


Рис. 5. Модуль результатов анализа главных компонент, вкладка «Quick»

Выберем вкладку «Quick». Нажав на кнопку «Eigenvalues», получим таблицу, содержащую собственный значения и долю выделенной дисперсии каждой компонентой и совокупностью компонент – рис. 6. График собственных значений (в зависимости от порядкового номера) получим, нажав на кнопку «Scree plot» - рис. 7.

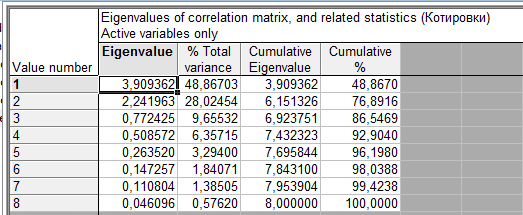


Рис. 6. Собственные значения и доли выделенной главными компонентами дисперсии.



Рис. 7. График собственных значений

Согласно графику (по критерию каменистой осыпи), следует оставить две компоненты. Также две компоненты следует оставить и по критерию Кайзера (только два собственных значения больше единицы). Согласно таблице, приведенной выше (рис. 6), первые две компоненты объясняют 76,9% общей дисперсии. Если этого не достаточно, следует выделить три компоненты, в этом случае, выделенная компонентами дисперсия, составит 86,5%.

Нажав на кнопку «Factor coordinates of variables», получим таблицу ковариаций исходных признаков и нормированных главных компонент (главных компонент с единичной дисперсией) (рис. 8).

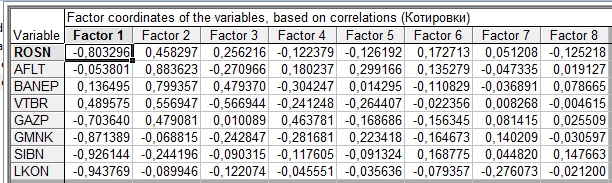


Рис. 8. Таблица ковариаций исходных признаков и нормированных главных компонент.

Ковариации исходных признаков и нормированных главных компонент есть не что иное, как координаты собственных векторов с квадратом нормы, равной собственному значению (выделенной фактором дисперсии). Если исходные данные нормированы, то факторные нагрузки это корреляции исходных признаков и главных компонент или, так называемые, факторные нагрузки.

На вкладке «Quick» можно также отобразить исходные признаки в системе любых двух выбранных факторов (рис. 9). Координатами каждого признака в данном случае являются соответствующие ковариации. Также можно изобразить поле факторных точек (рис. 10) для выбранных факторов по всем наблюдениям (на осях откладываются значения факторов для каждого наблюдения).



Рис. 9. Исходные признаки в системе координат первых двух компонент.



Рис. 10. Исходные наблюдения в пространстве первых двух компонент.

Наконец, на вкладке «Quick» можно получить значения всех компонент для каждого наблюдения, нажав на кнопку «Factor coordinates of cases» (рис. 11).

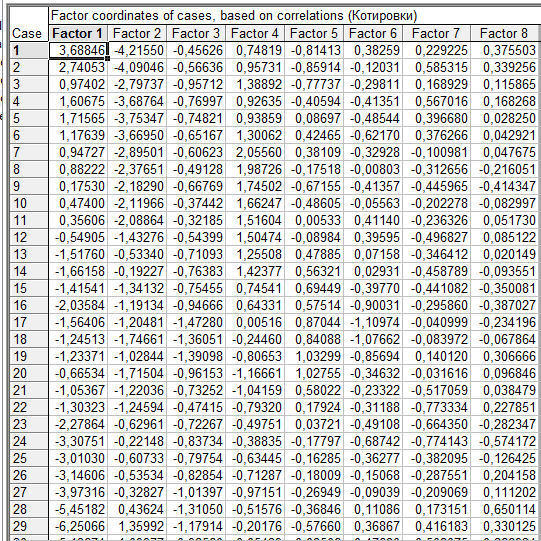


Рис. 11. Значения главных компонент для каждого наблюдения

Перейдем на вкладку «Variables» (рис. 12). Здесь можно дополнительно получить: факторные нагрузки - значения коэффициентов корреляции между исходными признаками и факторами - «Factor & variable correlation»; значения координат собственных векторов единичной длины - «Eigenvectors»; объясняемую суммарную дисперсию главными компонентами по каждому исходному признаку - «Communalities (Cosine ?)» (рис. 13).

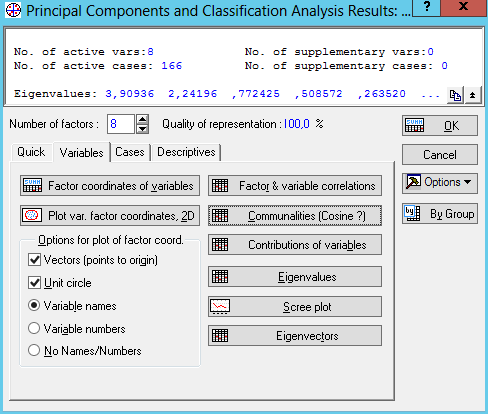


Рис. 12. Вкладка «Variables»

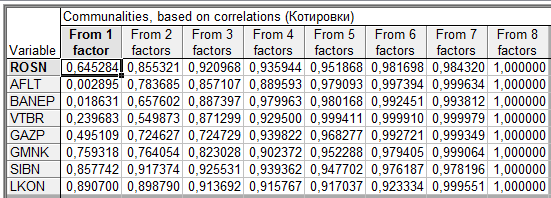


Рис. 13. Объясняемая факторами суммарная дисперсия по каждой компоненте.

На вкладке «Саses» (рис. 14) можно дополнительно получить: значения нормированных факторов - «Factor scores»; значения координат собственных векторов c нормой, равной обратной величине соответствующего собственного значения - «Factor score coefficients» (используются для получения нормированных значений факторов).

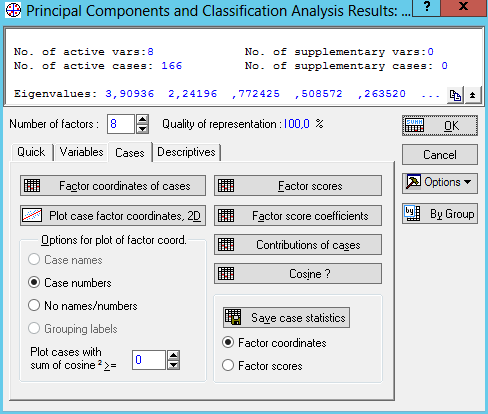


Рис. 14. Вкладка «Cases»