**Снижение размерности признакового пространства.**

Есть несколько методов решения этой задачи:

* метод главных компонент,
* факторный анализ.

**Метод главных компонент.**

**Основная модель МГК.**

Предпосылки появления:

1) Многие признаки существенно коррелированы.

2) Некоторые признаки обладают достаточно малой дисперсией, т.е. при переходе от одного объекта к другому почти не изменяются, и, поэтому, малоинформативны.

3) Возможно существуют новые признаки (может быть даже непосредственно не измеряемые).

**Определение и свойства главных компонент.**

Пусть дана матрица объект – признак 

Векторы x1, x2, ..., xp - это измеряемые признаки.

**Главными компонентами называются новые признаки y1, y2, ..., yp, обладающие свойствами:**

**1) Главная компонента - это линейная комбинация исходных измеряемых признаков **

**2) Главные компоненты ортогональны между собой, т.е. некоррелированы cov(yi, yj)=0, если i≠j.**

**3) Главные компоненты упорядочены по мере убывания дисперсии D(y1)≥D(y2)≥…≥D(yp).**

**Вычислительная процедура МГК.**

Введем вектор 

Это вектор весов i-ой главной компоненты. Тогда i-ая главная компонента в векторной форме выглядит следующим образом:

yi(N×1)=X(N×p)×ci(p×1), 

yi=Xci

Рассмотрим первую главную компоненту:

y1=X×c1



Пусть ковариационная матрица ∑ имеет вид:

, σij- ковариация между xi и xj

Тогда можно показать, что:

D(y1)= C1

В общем виде:

D(yi)= сi

На вектор весов наложим ограничение, состоящее в том, что сумма квадратов весов каждой компоненты равна 1.



сi=1, i = 1,...,p.

**Задача определения первой главной компоненты.**

Найти такой ненулевой вектор , что D(y1)=с1 максимальна по с1, при условии сi=1.

Эта задача на условный экстремум решается с помощью метода множителей Лагранжа.

Г(λ1)=с1-λ1(с1-1)

Это выражение должно быть максимальным по с1.

Берем производную по с1.

с1-λ1с1=0 Отсюда следует, что (-λ1E)с1= 0

Получили однородную систему р линейных уравнений с р неизвестными. Она имеет нетривиальное решение с1, если |-λ1E|=0. Таким образом, λ1- это собственное число ковариационной матрицы.

Из предыдущего уравнения получаем:

с1=λ1с1 .

Обе части этого равенства умножим слева на :

С1=λ1С1

λ1=С1=D (y1).

**Определение второй главной компоненты.**

Необходимо найти вектор С2= такой, что:

1) D(y2)= с2 максимальна среди оставшихся главных компонент

2) с2=1

3) cov(y2,y1)=0

Вектор с2 является собственным вектором матрицы . Он отвечает наибольшему из оставшихся собственных чисел ковариационной матрицы, т.е. собственному числу λ2.

Аналогичен смысл векторов с3, с4 и т.д.

**Задача определения i-ой главной компоненты:**

Нужно найти вектор сi= такой, что:

1) D(yi)=сi, максимальна среди оставшихся главных компонент

2) сi=1,

3) cov(yi,yj)=0, j=.

**Схема МГК исследования.**

1) Вычисляем ковариационную матрицу , или корреляционную R.

Замечание: Главная компонента изменяется при линейном преобразовании матрицы, поэтому главные компоненты матриц  и R могут существенно отличаться. Решение о том, с какой матрицей работать принимает исследователь в зависимости от размерности исходных признаков. Если все признаки измеряются в одинаковых единицах, то принципиальной разницы в использовании матриц нет. Если признаки имеют различную физическую природу, то лучше работать с матрицей R.

2) Вычисляем собственные числа матрицы R и упорядочиваем их по убыванию λ1≥λ2≥...≥λp, а также вычисляем собственные вектора с1, с2,..., сp, соответствующие этим собственным числам.

3) Вычисляем



Замечание: На практике имеет место следующий факт: для положительно определенной симметричной матрицы сумма элементов, стоящих на главной диагонали, равна сумме всех ее собственных чисел.



Т.е. при переходе от исходной системы признаков к главным компонентам, суммарная дисперсия не изменяется.



4) Выбор числа новых признаков.

Как только I(р')>ε, то р' равняется количеству слагаемых в числителе.

5) Интерпретация и анализ полученных главных компонент.

**Числовые характеристики главных компонент.**

Главная компонента 

Характеристики главных компонент:

1) M(yi)=0

2) D(yi)=сi

3) 

Т.е. ковариационная матрица главной компоненты имеет вид:



**Модели факторного анализа.**

**Статистическая модель главных компонент.**

**Определение: основное соотношение факторного анализа:**

**,**

**xj- стандартизированные измеряемые признаки (M=0, D=1);**

**fk- стандартизованные общие факторы;**

**j- центрированные, но не нормированные специфические факторы (М=0, D≠1),**

**дополненное следующими двумя предположениями:**

**1) общие факторы не коррелированны между собой.**

****

**2) общие факторы и факторные нагрузки таковы, что суммарная дисперсия специфических факторов минимальна.**

****

**-называется статистической моделью главных компонент (СМГК).**



Таким образом, близость между совокупностью измеряемых признаков и совокупностью общих факторов в СМГК понимается в смысле суммы квадратов евклидовых расстояний между соответствующими векторами.

**Смысл факторных нагрузок.**

Специфический фактор 

Общие факторы и факторные нагрузки выбираются из условия:



Найдем нагрузку j-го параметра на s-ый вектор.

Пусть имеется вектор a(x)=(a1 (x),..,an (x))







(j,fs)=0, j=1,...,p; s=1,...,m

**В СМГК специфические факторы и общие факторы не коррелируют между собой.**

;

xj,fs)-sj N=0

****

То есть |sj |≤1.

**Таким образом, смысл факторных нагрузок**  состоит в том, что это коэффициент корреляции между измеряемым признаком и общим фактором.

**Следствие: Специфический фактор и вычисленный признак **

**не коррелируют между собой.**

Доказательство:



**Дисперсия измеряемого признака.**



Так как xj – стандартизованные признаки, то







Нужно, чтобы общность давала больший вклад, а специфичность меньший.

**Другая формулировка СМГК.**

**Определение: основное соотношение факторного анализа:**

**,**

**xj- стандартизированные измеряемые признаки (M=0, D=1);**

**fk- стандартизованные общие факторы;**

**j- центрированные, но не нормированные специфические факторы (М=0, D≠1),**

**дополненное следующими двумя предположениями:**

**1) общие факторы не коррелированны между собой.**

****

**2\*) общие факторы и факторные нагрузки таковы, что**

****

**Доказательство условия 2\*:**

1=s2()+s2(j) просуммируем по j.

Получим





При этом факторные нагрузки выбираются из условия (fk,j)=0 или kj=rxj,fk). Таким образом, в СМГК близость между совокупностью измеряемых признаков и совокупностью общих факторов может пониматься в смысле суммы квадратов парных коэффициентов корреляции.