**Краткое изложение теории компонентного анализа** (по Прилож.А к докт. дисс.)

Как известно, компонентный анализ (КА) – один из методов многомерной статистики, в его основу положена гипотеза: наблюдаемые или измеряемые параметры являются лишь косвенными характеристиками изучаемого объекта или явления. На самом же деле существуют внутренние (скрытые, не измеряемые непосредственно) параметры или свойства, число которых мало и которые определяют значения наблюдаемых параметров [73]. Эти внутренние параметры, называемые главными компонентами (ГК), как предполагается, сохраняют всю информацию, содержащуюся во множестве наблюдаемых переменных.

И хотя ГК заранее нам не известны, КА ставит задачу представить наблюдаемые параметры в виде линейных комбинаций ГК и определить их, т.е. для каждого объекта указать значение каждой ГК. В таком случае модель КА может быть записана в виде [73, 101]:

**Y**[*n*×*m*] = **F**[*n*×*m*]×**A**[*m*×*m*]

где **Y**[*n*×*m*] представляет собой совокупность всех *n* наблюдаемых значений всех *m* параметров, **F**[*n*×*m*] – матрица, включающая совокупность всех *n* получаемых значений всех *m* ГК, это искомая матрица значений новых переменных в каждой точке опробования, а **A**[*m*×*m*] – так называемая матрица компонентных нагрузок – матрица НГК, или весовая матрица, она является связующим звеном между старыми и новыми переменными.

Как видим, в этом матричном уравнении две неизвестные матрицы – **A**[*m*×*m*] и **F**[*n*×*m*], а так как из одного уравнения можно найти только одно неизвестное, для определения второго требуется какое-то дополнительное условие. Таким дополнительным условием, исходной предпосылкой анализа, является наличие взаимосвязи между несколькими одновременно наблюдаемыми переменными.

В качестве количественной меры связи между двумя переменными используется коэффициент корреляции. Он может принимать значения от -1 до +1. При этом если он приближается к 0, это свидетельствует об отсутствии линейной связи, и чем более он близок к +1 или –1, тем более тесная линейная связь существует между переменными.

Все вычисленные коэффициенты корреляции между каждой парой переменных располагаются соответствующим образом в корреляционной матрице. В ней содержится важная информация о взаимоотношениях переменных с учетом влияния помех, причиной которых может явиться, например, неоднородность материала. При анализе такой корреляционной матрицы получают структуру искомых гипотетических величин – матрицу НГК (**A**[*m*×*m*]), которые находятся в определенных взаимоотношениях с переменными.

Эта структура находится в результате математических преобразований корреляционной матрицы, вычисленной по исходным результатам наблюдений. Преобразования основаны на теореме, использующей симметричность матрицы коэффициентов корреляции [73], и некоторых несложных операциях над матрицами, в результате которых получают:

**R**[*m*×*m*] = **U**′[*m*×*m*]×**Λ**[*m*×*m*]×**U**[*m*×*m*]

**A**[*m*×*m*] = **U**′[*m*×*m*]×**Λ1/2**[*m*×*m*]

В последних уравнениях – ортогональная матрица, столбцами которой являются собственные векторы матрицы , а **** – диагональная матрица, составленная из собственных чисел матрицы коэффициентов корреляции , соответствующих собственным векторам, причем элементы в матрице ****[*m*×*m*] расположены в порядке убывания: 1>2>...>m>0.

Таким образом, матрицу **A**[*m*×*m*] можно считать определенной, если известны собственные вектора и собственные числа матрицы **R**[*m*×*m*]. Для вычисления собственных значений и собственных векторов симметричных матриц существует много стандартных алгоритмов.

Так решается основная проблема компонентного анализа – определение матрицы весовых коэффициентов (НГК), учитывающих тесноту связи между признаками и главными компонентами (ГК).

Модель компонентного анализа предполагает точное определение, как компонентных нагрузок (НГК), так и значений ГК для каждой точки опробования (для каждого объекта). Поскольку 1>2>...>m>0, последние в этом ряду *i* вносят в суммарную дисперсию небольшой вклад. Поэтому на практике оставляют обычно небольшое число ГК, если на их долю приходится достаточно большой процент суммарной дисперсии параметров, которая равна следу матрицы ****[*m*×*m*], или размерности матрицы **R**[*m*×*m*], то есть числу параметров. Следовательно, можно выбрать небольшое число ГК (*q*<*m*) и добиться значительного облегчения описания признаков. Поскольку обычно *q/m =* 0,8÷0,9, то есть чаще всего ограничиваются 80-90% суммарной дисперсии, хотя эту величину можно устанавливать в зависимости от целей исследования и меньше этой величины, и больше.

После нахождения матрицы НГК (иногда говорят «весовых» нагрузок) определяют второе неизвестное – матрицу значений ГК в каждой точке опробования (для каждого объекта) – по уравнению записи модели.

В качестве ГК выбирается ассоциация исходных признаков, вошедших неё со статистически значимыми величинами нагрузок на соответствующие признаки. И хотя на первом этапе анализа (определение матрицы **A**[*m*×*m*]) по выделенным ГК можно сделать некоторые выводы об условиях, определяющих ход исследуемых процессов, матрица значений ГК **F**[*n*×*m*] представляет собой важный и полезный результат [70], особенно для картируемых величин. Первый этап компонентного анализа можно считать завершающим в том случае, когда его применяют для подтверждения выдвигаемых гипотез, то есть когда мы заранее предполагаем определенную иерархию признаков или их сочетаний, и выделенные ассоциации признаков (компоненты) подтверждают это предположение.

ГК отражают не простую сумму параметров, описывающих систему – они являются результатом системного взаимодействия этих параметров, тем новым свойством, которое появляется при построении системы. А процесс их предметной интерпретации заключается в выявлении общих причин, вызывающих "параллельное" или "антипараллельное" изменение измеряемых параметров. Свертывание информации заключается в том, что число действующих на систему факторов всегда меньше числа параметров, являющихся их проявлением, примером чему может служить проявление климатического фактора через температурный режим, зональность уровенного и химического режима грунтовых вод, зональность ландшафтов и т.д.

Построение графиков распределения значений компонент по объектам в пространстве самих компонент (например, в плоскости 1-ой и 2-ой главных компонент) дает способ группировки объектов по общим свойствам, то есть появляется возможность решения задач классификации и выделения "особых" зон (это фактически решение задачи кластерного анализа). С другой стороны, полученные значения компонент для каждого объекта можно рассматривать как новые переменные, которые при ортогональном решении не коррелируют между собой, а это является важным обстоятельством при использовании других статистических процедур, например, при построении уравнений регрессии на главных компонентах.

Таким образом, второй этап компонентного анализа, или результат решения обратной компонентной задачи, представляет значительный интерес. Это решение находится по вычисленной предварительно матрице компонентных нагрузок и нормированной матрице исходных данных по следующему уравнению:

**F**[*n*×*m*] = **Y**[*n*×*m*]×**A-1**[*m*×*m*]

Это уравнение справедливо лишь для случая, когда вычисляется полная матрица компонентных нагрузок **A**[*m*×*m*], и обратную матрицу можно найти. Если же ищутся только первые *q* компонент, удовлетворяющие заданной точности (**A**[*m*×*q*]), то алгоритмы решения несколько усложняются. Запишем теперь исходное уравнение модели компонентного анализа через матрицы значений ГК и матрицу компонентных нагрузок с сокращенным количеством ГК:

**Y**[*n*×*m*] = **F**[*n*×*q*]×**A**′[*q*×*m*]

В этой записи для сокращённого количества учитываемых ГК необходимо использовать транспонированную матрицу НГК. Чтобы из новой записи модели КА получить **F**[*n*×*q*], проведём небольшие преобразования:

**Y**[*n*×*m*]×**A**[*m*×*q*] = **F**[*n*×*q*]×**A**′[*q*×*m*]×**A**[*m*×*q*], или

**F**[*n*×*q*] = **Y**[*n*×*m*]×**A**[*m*×*q*]×(**A**′[*q*×*m*]×**A**[*m*×*q*])-1

Теперь учтём, что

**A**′[*q*×*m*]×**A**[*m*×*q*] = **Λ**[*q*×*q*],

получим:

**F**[*n*×*q*] = **Y**[*n*×*m*]×**A**[*m*×*q*] ×**Λ-1**[*q*×*q*]

Полученное решение есть окончательный результат КА в смысле математических построений, далее предстоит содержательная, предметная интерпретация выделенных ГК, которая, как отмечается во всех теоретических работах по КА, представляет главную трудность в применении этого метода.

73 Алгоритмы и программы восстановления зависимостей / В.Н. Вапник, Т.Г. Гладкова, В.А. Кащеев и др. – М.: Наука, 1984. – 815 с.