Ким Дж. О.

Факторный, дискриминантный и кластерный анализ

Москва
«Книга по Требованию»

Дж.-О. Ким, Ч. У. Мьюллер

ФАКТОРНЫЙ АНАЛИЗ: СТАТИСТИЧЕСКИЕ МЕТОДЫ И ПРАКТИЧЕСКИЕ ВОПРОСЫ

Jae-On Kim, Charles W. Mueller. **Factor Analysis:** Statistical Methods and Practical Issues (Eleventh Printing, 1986).

ПРЕДИСЛОВИЕ

Настоящая работа является продолжением книги Джэй-Он Кима и Чарльза У. Мьюллера «Введение в факторный анализ: что это такое и как им пользоваться», также опубликованной в серии «Quantitative Applications in the Social Sciences». Последняя является введением в метод факторного анализа; в ней даются ответы на вопросы читателя: «Для чего используется факторный анализ?» и «Какие предположения делаются при использовании этого метода?», но не затрагиваются вопросы применения факторного анализа к конкретным данным. В работе «Факторный анализ: статистические методы и практические вопросы» более подробно рассматриваются специфические примеры анализа данных, различные виды факторного анализа и ситуации, когда его применение наиболее полезно. Различие между конфирматорным и разведочным факторным анализом здесь обсуждается более детально, чем во «Введении в факторный анализ». Например, рассматриваются различные критерии для факторного вращения. Особенно полезным является обсуждение различных форм косоугольных вращений и интерпретации коэффициентов в факторном анализе. Дж.-О. Ким и Ч. У. Мьюллер также ставят вопрос о числе факторов, фигурирующих в разведочном факторном анализе, разбирают методы проверки гипотез в конфирматорном анализе и рассматривают проблему вычисления значений факторов. Предлагается словарь специальных терминов, а также ответы на вопросы, наиболее часто возникающие у пользователей факторного анализа, которые могут предостеречь их от многих ошибок. Математический аппарат достаточно скромный — приводятся только сведения из матричной алгебры.

ISBN 0-8039-1161-1

Факторный анализ использовался в экономических задачах, в которых наличие сильно коррелированных параметров приводило к неверным результатам в регрессионном анализе. Ученые, занимающиеся общественно-политическими проблемами, сопоставляли всевозможные признаки наций с разными политическими и социально-экономическими характеристиками, пытаясь определить, какие из них наиболее важны при классификации наций (например, благосостояние и численность); социологи определяли «дружественные группы», изучая группы людей, симпатизирующих именно друг другу (а не другим индивидуумам). Психологи использовали метод факторного анализа для определения того, как люди воспринимают всевозможные «стимулы» и классификации людей в группы, соответствующие различным реакциям, а издатели применяли факторный анализ для изучения способов связывать отдельные элементы языка*.

Как утверждают авторы, их работа не охватывает всех аспектов факторного анализа, так как он постоянно развивается. Тем не менее если читатель получит достаточно полное представление о том, как этот метод может быть использован, то можно считать, что авторы выполнили свою задачу.

Е. М. Асланер, редактор серии

---

* Более подробно одно из таких исследований описано в разд. «Кластерный анализ». — *Примеч. ред.*

I. ВВЕДЕНИЕ

Основная концепция факторного анализа проста и несложна для изучения. Тем не менее существует несколько причин, по которым овладение методом для практического использования может быть достаточно трудным. Во-первых, для понимания принципов статистического оценивания, как правило, требуется бóльшая искушенность в математике, чем это необходимо для понимания постановки задачи. Во-вторых, в литературе были описаны многочисленные методы получения факторных решений, и даже относительно простая компьютерная программа, вероятно, предусматривает различные варианты на каждой стадии анализа. Эти обстоятельства могут ошеломить начинающего и вызвать затруднения даже у специалиста. В-третьих, практическая задача почти всегда является более сложной, чем предполагается в факторной модели. Например, (1) организация измерений некоторых или всех переменных не соответствует требованиям, принятым в факторном анализе; (2) некоторые предположения модели, такие, как независимость ошибок измерений, могут не выполняться для определенных данных или (3) могут существовать второстепенные факторы, идентификация которых непосредственно не нужна, но присутствие которых влияет на идентификацию основных общих факторов. Трудность состоит в том, что исследователь должен в конце концов принять по собственному усмотрению некоторые «внестатистические» решения. К счастью, как будет показано, эти трудности преодолимы.

Исследователь для решения проблемы в большей или меньшей степени должен положиться на существующие компьютерные программы, которые часто предусматривают различные варианты вычислений, принятые по умолчанию. Последние устраивают пользователя по крайней мере до тех пор, пока задача не потребует некоторых изменений. Более того, по мере знакомства с разнообразными вариантами факторного анализа становится ясно, что различия между ними большей частью поверхностны. Фактически это разнообразие обусловлено расхождением в небольшом числе основных предположений.

Еще более существенно, что применение различных методов и критериев к одним и тем же данным приводит к эквивалентным, с практической точки зрения, результатам. Короче говоря,

читателю не обязательно изучать и использовать все варианты немедленно. Вместе с тем необходимо, чтобы пользователь знал наиболее распространенные алгоритмы факторного анализа и осознавал с самого начала тот факт, что большинство проблем не имеет единственного, окончательного (или наилучшего) решения.

Надеемся, что читатель имеет общее представление о концепции факторного анализа, а также знаком с различием между неоднозначностью вывода скрытой (латентной) факторной структуры из наблюдаемых ковариаций (логическая задача) и разбросом значений оценок параметров генеральной совокупности по выборке (статистическая задача). Хотя при получении решения задачи факторного анализа эти две проблемы в целом переплетаются, важно представлять концептуальные различия. Прежде чем мы изложим статистические методы и практические вопросы, нам кажется, что будет полезно обратиться к основам факторного анализа.

ОБЗОР ОСНОВ ФАКТОРНОГО АНАЛИЗА

В факторном анализе предполагается, что наблюдаемые переменные являются линейной комбинацией некоторых латентных (гипотетических или ненаблюдаемых) факторов. Некоторые из этих факторов допускаются общими для двух и более переменных, а другие — характерными для каждого параметра в отдельности. Характерные факторы — ортогональны друг другу (по крайней мере в разведочном факторном анализе). Следовательно, характерные факторы не вносят вклад в ковариацию между переменными. Другими словами, только общие факторы, число которых предполагается гораздо меньшим числа наблюдаемых переменных, вносят вклад в ковариацию между ними.

Принимаемая в факторном анализе линейная система такова, что структура ковариаций может быть идентифицирована без ошибок, если известна матрица нагрузок латентных факторов. Тем не менее однозначное восстановление латентной факторной структуры исходя из наблюдаемой ковариационной структуры всегда проблематично. Эта неопределенность не имеет никакого отношения к статистическому оцениванию и должна разрешаться с помощью «внестатистических» постулатов: принципа факторной причинности и принципа экономии.

При использовании этих постулатов и свойств линейной системы можно точно идентифицировать латентную факторную структуру путем исследования результирующей ковариационной матрицы, если структура не является слишком сложной и если она удовлетворяет требованиям простой факторной структуры. Модель с двумя общими факторами (рис. 1) может быть восстановлена из матрицы корреляций, представленной в нижнем треугольнике табл. 1. Любая компьютерная программа (какой бы алгоритм в ней ни был заложен) позволяет достаточно хорошо восстановить данную модель[1].

На практике тем не менее на исследуемую матрицу корреляций оказывают влияние различные случайные и неслучайные ошибки, и в результате она будет отлична от корреляционной матрицы, обусловленной факторной структурой генеральной совокупности. Над главной диагональю табл. 1 помещены элементы корреляционной матрицы, вычисленной для выборки объема 100 с использованием факторного отображения, приведенного на рис. 1 (т. е. с использованием матрицы корреляции под диагональю табл. 1). Обратите внимание на отличие между соответствующими наддиагональными и поддиагональными элементами таблицы и на тот факт, что каждая выборочная корреляционная матрица будет отличаться в некоторой степени от корреляционной матрицы для генеральной совокупности и от любой другой выборочной матрицы для других выборок из той же самой генеральной совокупности. Таким образом, на практике невозможно получить точную структуру факторной модели, можно только пытаться найти оценки параметров факторной структуры, с использованием определенных статистических и (или) практических критериев.

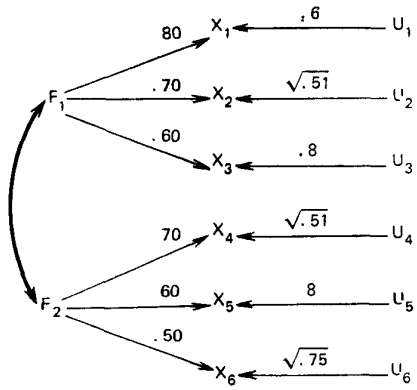

Рис. 1. Граф факторной структуры с шестью переменными и двумя косоугольными общими факторами, где наблюдаемые переменные означают:

X_1 — правительство должно тратить больше средств на школы;
X_2 — правительство должно тратить больше средств на сокращение процента безработных;
X_3 — правительство должно контролировать большой бизнес;
X_4 — правительство должно устранять сегрегацию через занятость населения;
X_5 — правительство должно обеспечивать национальным меньшинствам соответствующую квоту рабочих мест;
X_6 — правительство должно выполнять программу борьбы с кризисами

При решении задач разведочного факторного анализа исследователь обычно делает три шага: (1) подготовка соответствующей ковариационной матрицы; (2) выделение первоначальных (ортогональных) факторов и (3) вращение с целью получения окончательного решения. Подчеркнем, что исходную информацию для факторного анализа получить сравнительно просто.

ОСНОВНЫЕ АЛГОРИТМЫ И МЕТОДЫ

В зависимости от задач исследователя следует воспользоваться либо *разведочным,* либо конфирматорным факторным анализом. В обоих случаях существуют три основных этапа: подготовка соответствующей матрицы ковариаций, выделение перво-

Таблица 1

Коэффициенты корреляции для генеральной совокупности (поддиагональные элементы) и модельной выборки объема 100 (наддиагональные элементы), относящиеся к модели с двумя общими факторами, представленной на рис. 1

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	—	0,6008	0,4984	0,1920	0,1939	0,3466
X_2	0,560	—	0,4749	0,2196	0,1912	0,2979
X_3	0,480	0,420	—	0,2079	0,2010	0,2445
X_4	0,224	0,196	0,168	—	0,4334	0,3197
X_5	0,192	0,168	0,144	0,420	—	0,4207
X_6	0,160	0,140	0,120	0,350	0,300	—

начальных факторов и вращение с целью получения окончательного решения. Хотя на практике для получения окончательного решения не всегда требуются все эти шаги (особенно при проверке специальных гипотез), тем не менее удобно обсуждать разнообразие методов факторного анализа в связи с данными этапами. Таким образом, первая часть этой работы так или иначе касается этих трех этапов анализа.

Перед проведением факторного анализа необходимо решить: использовать ли как исходную матрицу ковариации (корреляции) между переменными или использовать корреляции между индивидуумами (объектами). В данной работе мы будем обсуждать только первый из этих подходов*.

На первом этапе может применяться модель *общих факторов,* а также *анализ главных компонент,* цель которого отлична от цели факторного анализа. В то же время оба метода широко используются и являются эффективными способами исследования «взаимосвязей» между переменными. Основное отличие между этими двумя методами заключается в том, что главные компоненты являются линейными функциями от наблюдаемых переменных, в то время как общие факторы не выражаются через комбинацию наблюдаемых переменных. Альтернативой анализу первоначальных факторов служит анализ образов-факторов, в котором предполагается, что наблюдаемые переменные выбраны из бесконечного множества переменных, причем вводятся «образы-факторы», являющиеся линейными комбинациями переменных. Сопоставление этих подходов будет рассмотрено ниже. Кроме того, существует несколько путей выделения первоначальных факторов. Из них в этой работе рассматриваются следующие: 1) решение, получаемое методом максимального правдоподобия (включая канонический факторный анализ); 2) решение по ме-

* Второй подход, так называемый *Q*-техника, кратко рассматривается в разд. «Кластерный анализ» — *Примеч. ред.*

тоду наименьших квадратов (включая метод минимальных остатков и метод главных факторов с итерациями по общностям) и 3) альфа-факторный анализ. Последний может рассматриваться либо как вариант метода с общими факторами, либо как альтернативная стратегия.

Шаг, связанный с вращением, включает два варианта: ортогональное и косоугольное вращение. Косоугольные вращения в свою очередь подразделяются на те, которые основаны на прямом упрощении матрицы коэффициентов факторного отображения, и на те, которые используют упрощение матрицы нагрузок на вторичные оси. Внутри этих вариантов существует множество подвариантов. О большинстве из них мы поговорим в следующих разделах. Вопрос о числе факторов рассматривается отдельно, что связано с необходимостью обсудить несколько эмпирических правил, которые многие практики находят полезными.

В разделе, посвященном конфирматорному факторному анализу, будет дано понятие эмпирического подтверждения факторных моделей, а затем мы проиллюстрируем его на двух простых, но важных практических примерах.

Далее мы обсудим вопрос вычисления значений факторов. Этот раздел помещен после обсуждения конфирматорного факторного анализа, поскольку используются некоторые его результаты.

В заключительном разделе рассматривается широкий спектр проблем в форме вопросов и ответов, причем многие из них в основном тексте вовсе не обсуждались. Здесь мы также даем некоторые практические советы для решений, по которым пока нет единого мнения.

Словарь, приложенный в конце работы, служит не для точного определения каждого термина, а лишь дает удобный способ представления контекста, в котором этот термин встречается.

И наконец, ссылки не предназначены ни для отражения исторического развития методов факторного анализа, ни для обзора последних достижений в этой области. Мы пользовались источниками, которые считали ценными, с точки зрения нашего собственного понимания предмета.

II. МЕТОДЫ ВЫДЕЛЕНИЯ ПЕРВОНАЧАЛЬНЫХ ФАКТОРОВ

Основная цель выделения первичных факторов в разведочном факторном анализе заключается в определении минимального числа общих факторов, которые удовлетворительно воспроизводят корреляции между наблюдаемыми переменными. При отсутствии ошибок измерений и случайности в выборке, а также при выполнении принципа факторной причинности, для заданной корреляционной матрицы существует точное соответствие между минимальным числом общих факторов и рангом редуцированной

корреляционной матрицы. (В редуцированной корреляционной матрице общности помещаются на главную диагональ.) Иными словами, в случае отсутствия ошибок в соответствии факторной модели данным число общих факторов и общности могут быть сколь угодно точно вычислены с помощью исследования ранга редуцированной корреляционной матрицы. Если же выборка является случайной, то проблема усложняется и возникает задача найти критерий, с помощью которого можно было бы оценить минимально необходимое число общих факторов. Но поскольку основной критерий определения минимального числа общих факторов заключается в хорошей воспроизводимости наблюдаемых корреляций с помощью отобранных факторов, то задачу можно переформулировать следующим образом: определить правило остановки при выделении общих факторов. Эта задача сводится к определению момента, когда расхождение между вычисленными и наблюдаемыми корреляциями может быть приписано случайности выборки.

Мы начнем с описания основной стратегии, которая является общей для ряда методов выделения. Она включает проверку гипотез о минимальном числе общих факторов, необходимых для воспроизведения наблюдаемых корреляций. При отсутствии априорных данных следует обратиться к однофакторной модели. Эта «гипотеза» (достаточности одного фактора) проверяется с помощью критерия, применяя который можно узнать, достигнуто ли удовлетворительное расхождение между предполагаемой моделью и данными. Если расхождение статистически значимо, то оценивается модель с еще одним дополнительным фактором и снова применяется критерий. Этот процесс продолжается до тех пор, пока расхождение не сможет быть приписано случайности выборки. Следует заметить, что реальные компьютерные программы могут явно не делать такую последовательную оценку, но принцип выделения первых k факторов, которые согласуются с наблюдаемыми ковариациями, остается в силе.

Хотя принцип этой основной стратегии прост, его применение — разнообразно, поскольку есть различные критерии наилучшего соответствия (или минимальной невязки). Существуют два главных метода решения, в которых фигурируют общие факторы: 1) метод максимального правдоподобия [Lawley, Maxwell, 1971; Jöreskog, 1967; Jöreskog, Lawley, 1968], варианты которого сводятся к каноническому факторному анализу [Rao, 1955] и к алгоритмам, основанным на минимизации детерминантов матрицы частных коэффициентов корреляции [Browne, 1968]; 2) метод наименьших квадратов, варианты которого включают метод главных осей с итерациями по общности [Thomson, 1934] и метод минимальных остатков [Harman, 1976]. Кроме того, существуют еще три основных метода выделения: 1) альфа-факторный анализ [Kaiser, Gaffrey, 1965]; 2) анализ образов [Guttman, 1953; Harris, 1962] и 3) анализ главных компонент [Hotelling, 1933].

ГЛАВНЫЕ КОМПОНЕНТЫ, СОБСТВЕННЫЕ ЗНАЧЕНИЯ И ВЕКТОРА

Мы начинаем обсуждение именно с анализа главных компонент по двум причинам: во-первых, он послужит в качестве базовой модели, с которой будут сравниваться и сопоставляться методы, где используются общие факторы. Во-вторых, он представляется наиболее простым для введения таких особых понятий, как корни характеристического уравнения (собственные числа) и собственные вектора, и дает возможность выявить их роли в алгоритмах факторного анализа. (Мы не отказывается от стремления применять наиболее простой математический аппарат, но знакомство с подобной терминологией необходимо для использования многих компьютерных программ. Мы настоятельно рекомендуем читателям ознакомиться с основными определениями.)

Анализ главных компонент — это метод преобразования данной последовательности наблюдаемых переменных в другую последовательность переменных. Наиболее простой способ пояснить внутреннюю логику метода сводится к его изучению в двумерном случае. Предположим, что есть две переменные X и Y с совместным нормальным распределением.

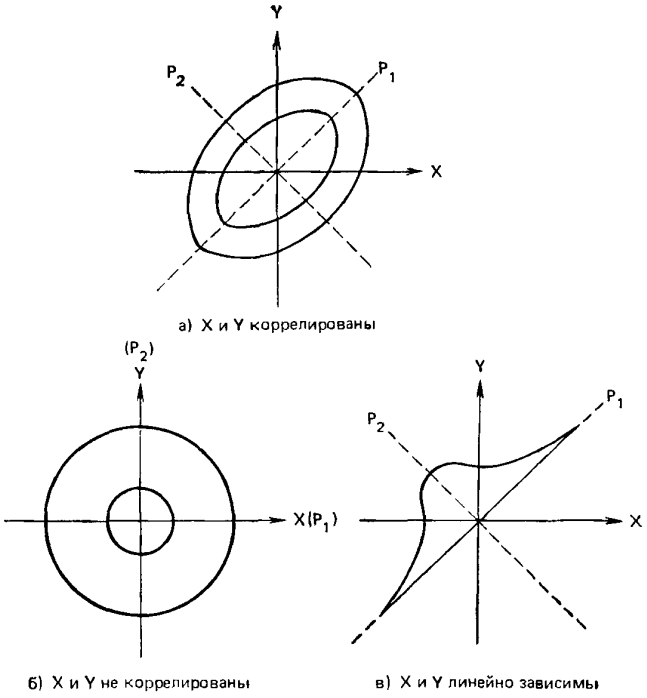

Рис. 2. Главные оси двумерных распределений

Совместное нормальное распределение величин, имеющих положительную корреляцию, представлено на рис. 2 с помощью кривых равных вероятностей. Эти кривые показывают, что благодаря положительной связи между X и Y данные представляют кластер, в котором бо́льшие величины X имеют тенденцию соответствовать бо́льшим величинам Y (и наоборот). Таким образом, в большинстве случаев точки попадают в первый и третий квадранты, и реже — во второй и четвертый. Кривые равных вероятностей имеют форму эллипсов, две оси которых изображены пунктирными линиями. Главная ось (P_1) проходит по линии, вдоль которой располагается основная часть данных; вторая ось (P_2) — по линии, вдоль которой расположена меньшая часть данных.

Теперь предположим, что нужно представить точки в терминах только одной размерности (оси). В этом случае естественно выбрать ось P_1, потому что в целом она ближе описывает данные наблюдений. Тогда первая главная компонента есть не что иное, как представление точек, расположенных вдоль выбранной главной оси. Например, точка с единичными значениями X и Y будет иметь координату, большую 1 по оси P_1 и меньшую 1 по оси P_2. Если мы описываем каждую точку в терминах P_1 и P_2 (в новой системе координат), потери информации не произойдет. Тем не менее можем сказать, что первая ось (и первая компонента) является более информативной в описании точек, так как связь между X и Y становится сильнее. В том случае, когда X и Y связаны линейной зависимостью, первая главная компонента будет содержать всю информацию, необходимую для описания каждой точки. Если X и Y независимы, то главная ось отсутствует и анализ главных компонент не способствует даже минимальному сокращению (сжатию) результатов наблюдений.

Понятие главных осей относится не только к нормальным распределениям. В общем случае главная ось задается линией, для которой сумма квадратов расстояний до всевозможных точек минимальна. Сравнение анализа главных компонент с принципом наименьших квадратов поможет объяснить это определение. При нахождении линии регрессии ($\hat{Y}=a+bX$) методом наименьших квадратов мы минимизируем сумму квадратов расстояний между Y и $\hat{Y}$, т. е. минимизируем* $(Y-\hat{Y})$, где расстояние измеряется по линии, параллельной оси Y и перпендикулярной оси X. При нахождении главной оси мы минимизируем расстояние** от точки до оси (т. е. расстояние по перпендикуляру к главной оси, а не к оси X). Это отличие показано на рис. 3. (В [Malinvand, 1970] описан метод наименьших квадратов с помощью ортогональной регрессии.)

---

* Более точно минимизируется среднее значение квадрата такой невязки. — *Примеч. ред.*

** Более точно минимизируется среднее значение квадрата этого расстояния. — *Примеч. ред.*