7. ЛЕКЦИЯ. Метод главных компонент

7.1. Определение, вычисление и основные числовые характеристики главных компонент

Во многих задачах обработки многомерных наблюдений и, в частности, в задачах классификации исследователя интересуют в первую очередь лишь те признаки, которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного объекта к другому.

С другой стороны, не обязательно для описания состояния объекта использовать какие-то из исходных, непосредственно замеренных на нем признаков. Так, например, для определения специфики фигуры человека при покупке одежды достаточно назвать значения двух признаков (размер — рост), являющихся производными от измерений ряда параметров фигуры. При этом, конечно, теряется какая-то доля информации (портной измеряет до одиннадцати параметров на клиенте), как бы огрубляются (при агрегировании) получающиеся при этом классы. Однако, как показали исследования, к вполне удовлетворительной классификации людей с точки зрения специфики их фигуры приводит система, использующая три признака, каждый из которых является некоторой комбинацией от большого числа непосредственно замеряемых на объекте параметров.

Именно эти принципиальные установки заложены в сущность того линейного преобразования исходной системы признаков, которое приводит к главным компонентам. Формализуются же эти установки следующим образом.

Следуя общей оптимизационной постановке задачи снижения размерности (1) и полагая анализируемый признак X p-мерной случайной

$$I_{p'}\left(\tilde{Z}(X)\right) = \max_{Z \in F}\{I_{p'}(Z(X))\} \quad (1)$$

величиной с вектором средних значений $\mu = (\mu^{(1)}, \dots, \mu^{(p)})$ и ковариационной матрицей $\sum = (\sigma_{ij})$ $(i, j = 1,\ 2, \dots, p)$, вообще говоря, неизвестными, определим меру (критерий) информативности $I_{p'}(Z)$ вспомогательной p'-мерной системы показателей $Z = (z^{(1)}, \dots, z^{(p')})$ с помощью (2),

$$I_{p'}\big(Z(X)\big) = \frac{D_{z^{(1)}} + \dots + D_{z^{(p')}}}{D_{x^{(1)}} + \dots + D_{x^{(p)}}} \quad (2)$$

а класс допустимых преобразований — в виде (3). Тогда при любом

$$\sum_{v=1}^{p} c_{jv}^2 = 1, j = 1,2, \dots, p; \quad (3)$$

фиксированном $p' = 1, 2, \dots, p$ вектор искомых вспомогательных переменных $\tilde{Z}(X) = (\widetilde{z^{(1)}}(X), \dots, \widetilde{z^{(p')}}(X))$ определяется как такая линейная комбинация

$$\tilde{Z} = LX \quad (4)$$

(где матрица

$$L = \begin{pmatrix} l_{11} & \dots & l_{1p} \\ \dots & \dots & \dots \\ l_{p'1} & \dots & l_{p'p} \end{pmatrix},$$

а ее строки удовлетворяют условию ортогональности), что

$$I_{p'}\left(\widetilde{z^{(1)}}(X), \dots, \widetilde{z^{(p')}}(X)\right) = \max_{Z(X)\in F} I_{p'}\big(Z(X)\big)$$

Полученные таким образом переменные $\widetilde{z^{(1)}}(X), \dots, \widetilde{z^{(p')}}$ и называют главными компонентами вектора X. Поэтому можно дать следующее определение главных компонент.

Первой главной компонентой $\widetilde{z^{(1)}}(X)$ исследуемой системы показателей $X = (x^{(1)}, \dots, x^{(p)})$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая среди всех прочих нормированно-центрированных линейных комбинаций переменных $x^{(1)}, \dots, x^{(p)}$ обладает наибольшей дисперсией.

k-й главной компонентой $(k = 2, 3, \dots, p)$ исследуемой системы показателей $X = (x^{(1)}, \dots, x^{(p)})$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая не коррелирована с $k-1$ предыдущими главными компонентами и среди всех прочих нормированно-центрированных и не коррелированных с предыдущими $k-1$ главными компонентами линейных комбинаций переменных $x^{(1)}, \dots, x^{(p)}$ обладает наибольшей дисперсией.

З а м е ч а н и е 1 (переход к центрированным переменным). Поскольку решение задачи (а именно вид матрицы линейного преобразования L) зависит только от элементов ковариационной матрицы Σ, которые в свою очередь не изменяются при замене исходных переменных x^l переменными $x^l - c^l$ (c^l — произвольные постоянные числа), то в дальнейшем будем считать, что исходная система показателей уже центрирована, т. е. что $\mathrm{E}x^{(j)} = 0, j = 1, 2, \dots, p$. В практике этого добиваются, переходя к наблюдениям $\tilde{x}_i^{(j)} = x_i^{(j)} - \bar{x}^{(j)}$, где $\bar{x}^{(j)} = \sum_{i=1}^{n} x_i^{(j)}/n$ (для упрощения обозначений волнистую черту над центрированной переменной и над главной компонентой в дальнейшем ставить не будем).

З а м е ч а н и е 2 (переход к выборочному варианту). Поскольку в реальных задачах располагаем лишь оценками $\hat{\mu}$ и $\hat{\Sigma}$ соответственно вектора средних μ и ковариационной матрицы Σ, то во всех дальнейших рассуждениях

под $\mu^{(j)}$ понимается $\bar{x}^{(j)}$, а под σ_{kj} — выборочная ковариация $\widehat{\sigma_{kj}} = \sum_{i=1}^{n}(x_i^{(k)} - \bar{x}^{(k)})(x_i^{(j)} - \bar{x}^{(j)})/n \ (j,k = 1,2,\dots,p)$.

Вычисление главных компонент. Из определения главных компонент следует, что для вычисления первой главной компоненты необходимо решить оптимизационную задачу вида

$$\begin{cases} D(l_1, X) \to max \\ l_1 l_1' = 1 \end{cases} \tag{5}$$

где l_1 – первая строка матрицы L (4). Учитывая центрированность переменной X (т. е. $\mathrm{E}X = 0$) и то, что $\mathrm{E}(XX') = \Sigma$, имеем

$$D(l_1, X) = \mathrm{E}(l_1, X)^2 = \mathrm{E}(l_1 X X' l_1') = l_1 \Sigma l_1'$$

Следовательно, задача (5) может быть записана

$$\begin{cases} l_1 \Sigma l_1' \to max \\ l_1 l_1' = 1 \end{cases} \tag{5'}$$

Вводя функцию Лагранжа $\varphi(l_1, \lambda) = l_1 \Sigma l_1' - \lambda(l_1 l_1' - 1)$ и дифференцируя ее по компонентам вектор-столбца l_1' имеем

$$\frac{\partial \varphi}{\partial l_1'} = 2\Sigma l_1' - 2\lambda l_1'$$

что дает систему уравнений для определения l_1:

$$(\Sigma - \lambda I) l_1' = 0 \tag{6}$$

(здесь $0 = (0,0,\dots,0)'$ – p-мерный вектор-столбец из нулей).

Для того чтобы существовало ненулевое решение системы (6) (а оно должно быть ненулевым, так как $l_1 l_1' = 1$), матрица $\Sigma - \lambda I$ должна быть вырожденной, т. е.

$$|\Sigma - \lambda I| = 0 \tag{7}$$

Этого добиваются за счет подбора соответствующего значения λ. Уравнение (9) (относительно λ) называется характеристическим для матрицы Σ. Известно, что при симметричности и неотрицательной определенности матрицы Σ (каковой она и является как всякая ковариационная матрица) это уравнение имеет p вещественных неотрицательных корней $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, называемых характеристическими (или собственными) значениями матрицы Σ. Учитывая, что $Dz^{(1)} = D(l_1, X) = l_1 \Sigma l_1'$ и $l_1 \Sigma l_1' = \lambda$ (последнее соотношение следует из (6) после его умножения слева на l_1 с учетом $l_1 l_1' = 1$, получаем

$$Dz^{(1)}(X) = \lambda$$

Поэтому для обеспечения максимальной величины дисперсии переменной $z^{(1)}$ нужно выбрать из p собственных значений матрицы Σ наибольшее, т. е.

$$Dz^{(1)}(X) = \lambda_1$$

Подставляем λ_1 в систему уравнений (6) и, решая ее относительно $l_{11}, \dots, l_{1p}$, определяем компоненты вектора l_1.

Таким образом, первая главная компонента получается как линейная комбинация $z^{(1)}(X) = l_1 X$, где l_1 – собственный вектор матрицы Σ, соответствующий наибольшему собственному числу этой матрицы.

Далее аналогично можно показать, что $z^{(k)}(X) = l_k X$, где l_k – собственный вектор матрицы Σ, соответствующий k-му по величине собственному значению λ_k этой матрицы.

Таким образом соотношения для определения всех p главных компонент вектора X могут быть представлены в виде

$$Z = LX$$

где $Z = (z^{(1)}, \dots, z^{(p)})', X = (x^{(1)}, \dots, x^{(p)})'$, а матрица L состоит из строк $l_j = l_{j1}, \dots, l_{jp}, j = \overline{1,p}$, являющихся собственными векторами матрицы Σ, соответствующими собственным числам λ_j. При этом сама матрица L по построению является ортогональной, т. е.

$$LL' = L'L = I$$

Основные числовые характеристики главных компонент. Определим основные числовые характеристики (средние значения, дисперсии, ковариации) главных компонент в терминах основных числовых характеристик исходных переменных и собственных значений матрицы Σ:

а) $\mathrm{E}Z = \mathrm{E}(\mathrm{L}X) = \mathrm{L}\mathrm{E}X = 0$

б) ковариационная матрица вектора главных компонент:

$$\Sigma_Z = \mathrm{E}(ZZ') = \mathrm{E}((\mathrm{L}X)(\mathrm{L}X)') = \mathrm{E}(\mathrm{L}XX'\mathrm{L}') = \mathrm{L}\mathrm{E}(XX')\mathrm{L}' = \mathrm{L}\Sigma\mathrm{L}'$$

Умножая слева соотношения

$$(\Sigma - \lambda_k \mathrm{I})l_k' = 0 \ (k = \overline{1,p})$$

на $l_j (j = \overline{1,p})$, получаем, что

$$\mathrm{L}\Sigma\mathrm{L}' = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_p \end{pmatrix}$$

и, следовательно:

$$\Sigma_z = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_p \end{pmatrix} \quad (8)$$

Из (8), в частности, следует подтверждение взаимной некоррелированности главных компонент, а также

$$D_z^{(k)} = \lambda_k \ (k = \overline{1,p})$$

в) сумма дисперсий исходных признаков равна сумме дисперсий всех главных компонент. Действительно, $\sum_{k=1}^{p} D_z^{(k)} = S_p\Sigma_z = S_p(\mathrm{L}\Sigma\mathrm{L}') = S_p((\mathrm{LL}')\Sigma) = S_p\Sigma = \sum_{k=1}^{p} DX^{(k)}$

г) обобщенная дисперсия исходных признаков (X) равна обобщенной дисперсии главных компонент (Z). Действительно, обобщенная дисперсия вектора Z равна

$$|\Sigma_z| = |\mathrm{L}\Sigma\mathrm{L}'| = |(\mathrm{LL}')\Sigma| = |\Sigma|$$

С л е д с т в и е. Из б) и в), в частности, следует, что критерий информативности метода главных компонент (7) может быть представлен в виде

$$I_{p'}\big(Z(X)\big) = \frac{\lambda_1+\cdots+\lambda_{p'}}{\lambda_1+\cdots+\lambda_p} \qquad (7')$$

где $\lambda_1 + \cdots + \lambda_p$ – собственные числа ковариационной матрицы Σ вектора X, расположенные в порядке убывания.

Кстати, представление $I_{p'}\big(Z(X)\big)$ в виде $(7')$ дает исследователю некоторую основу, опорную точку зрения, при вынесении решения о том, сколько последних главных компонент можно без особого ущерба изъять из рассмотрения, сократив тем самым размерность исследуемого пространства.

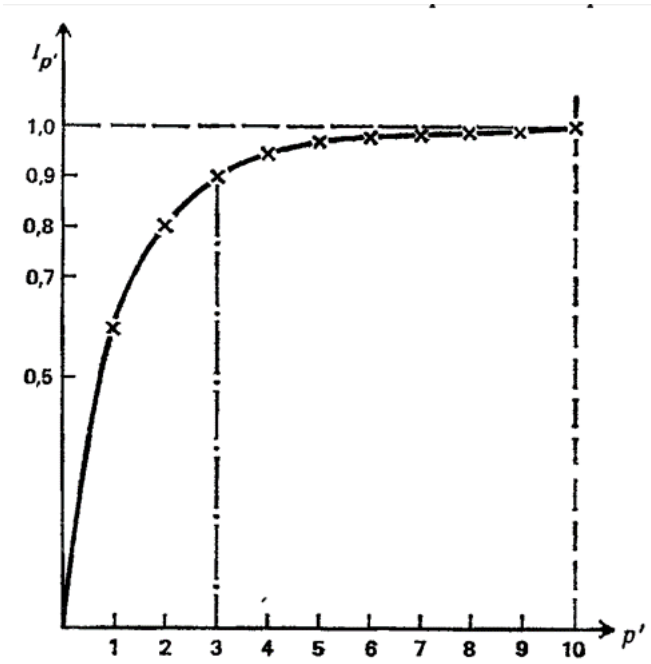

Рис. 1. Изменение относительной доли суммарной дисперсии исследуемых признаков, обусловленной первыми p' главными компонентами, в зависимости от p (случай $p = 10$)

Действительно, анализируя с помощью $(7')$ изменение относительной доли дисперсии, вносимой первыми p' главными компонентами, в зависимости от числа этих компонент, можно разумно определить число компонент, которое

целесообразно оставить в рассмотрении. Так, при изменении $I_{p'}$, изображенном на рис. 1, очевидно, целесообразно было бы сократить размерность пространства с $p = 10$ до $p' = 3$, так как добавление всех остальных семи главных компонент может повысить суммарную характеристику рассеяния не более чем на 10 %.

З а м е ч а н и е 3. Использование главных компонент оказывается наиболее естественным и плодотворным в ситуациях, в которых все компоненты $x^{(1)}, \ldots, x^{(p)}$ исследуемого вектора X имеют общую физическую природу и соответственно измерены в одних и тех же единицах. К таким примерам можно отнести исследование структуры бюджета времени индивидуумов (все $x^{(i)}$ измеряются в единицах времени), исследование структуры потребления семей (все $x^{(i)}$ измеряются в денежных единицах), исследование общею развития и умственных способностей индивидуумов с помощью специальных тестов (все $x^{(i)}$ измеряются в баллах), разного рода антропологические исследования (все $x^{(i)}$ измеряются в единицах меры длины) и т.д. Если же различные признаки $x^{(1)}, \ldots, x^{(p)}$ измеряются в различных единицах, то результаты исследования с помощью главных компонент будут существенно зависеть от выбора масштаба и природы единиц измерения. Поэтому в подобных ситуациях исследователь предварительно переходит к вспомогательным безразмерным признакам $x^{*(i)}$, например с помощью нормирующего преобразования.

$$x_v^{*(i)} = \frac{x_v^{(i)}}{\sqrt{\widehat{\sigma_{ii}}}} \quad \begin{pmatrix} i = 1, 2, \ldots, p \\ v = 1, 2, \ldots, n \end{pmatrix}$$

где σ_{ii} соответствует ранее введенным обозначениям, а затем строит главные компоненты относительно этих вспомогательных признаков X^* и их ковариационной матрицы $\widehat{\Sigma_{X^*}}$, которая, как легко видеть, является одновременно выборочной корреляционной матрицей R исходных наблюдений X_i.

З а м е ч а н и е 4. В некоторых задачах оказывается полезным понятие так называемых обобщенных главных компонент, при определении которых оговаривают более общие (чем $\sum_{j=1}^{p} l_{ij}^2 = 1$) ограничения на коэффициенты l_{ii}, т. е. требуют, чтобы

$$\sum_{k=1}^{p} \sum_{j=1}^{p} l_{ij}\, \omega_{kj} l_{ik} = 1$$

где ω_{kj} – некоторые дополнительно введенные веса. Очевидно, если $\omega_{kj} = 1$ при $k = j$ и $\omega_{kj} = 0$ при $k \neq j$, то имеем обычное условие нормировки коэффициентов l_{ij} и обычные главные компоненты. Можно показать, что при такой модификации условий нормировки коэффициенты $l_i = (l_{i1}, l_{i2}, \dots, l_{ip})$ с помощью которых обобщенные главные компоненты $z^{(i)}$ выражаются через исходные признаки $x^{(1)}, \dots, x^{(p)}$, определяются как решения уравнений

$$\left(\Sigma - \widetilde{\lambda_\iota}\Omega\right)l_i' = 0$$

где $\widetilde{\lambda_\iota}$ – i-й по величине корень уравнения $\left|\Sigma - \tilde{\lambda}\Omega\right| = 0$, а матрица $\Omega = \left(\omega_{ij}\right)$, $i, j = 1, 2, \dots, p$, – некоторая положительно определенная матрица весов. При этом, как и прежде, дисперсия обобщенной главной компоненты $z^{(i)}$ равна $\widetilde{\lambda_\iota}$, а $z^{(i)}$ и $z^{(j)}$ при $i \neq j$ взаимно Ω -не коррелированы.

Заметим, кстати, что если в качестве матрицы весов выбрать матрицу

$$\Omega = \begin{pmatrix} \sigma_{11} & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & \sigma_{pp} \end{pmatrix}$$

то, как легко показать, обобщенные компоненты (в метрике Ω), построенные по исходным признакам $x^{(1)}, \dots, x^{(p)}$, совпадут с обычными компонентами, построенными по вспомогательным безразмерным (нормированным) признакам $x^{*(1)}, \dots, x^{*(p)}$.

Проиллюстрируем определение главных компонент на численном примере.

П р и м е р 1. По данным измерений (в мм) длины $\widetilde{x^{(1)}}$, ширины $\widetilde{x^{(2)}}$ и высоты $\widetilde{x^{(3)}}$ панциря 24 особей ($n = 24$) одного из видов черепах определена выборочная ковариационная матрица

$$\hat{\Sigma} = \begin{pmatrix} 451{,}39 & 271{,}17 & 168{,}70 \\ 271{,}17 & 171{,}73 & 103{,}29 \\ 168{,}70 & 103{,}29 & 66{,}65 \end{pmatrix}$$

Решая кубическое уравнение (относительно λ) вида

$$\begin{vmatrix} 451{,}39 - \lambda & 271{,}17 & 168{,}70 \\ 271{,}17 & 171{,}73 - \lambda & 103{,}29 \\ 168{,}70 & 103{,}29 & 66{,}65 - \lambda \end{vmatrix} = 0$$

находим $\lambda_1 = 680{,}40$, $\lambda_2 = 6{,}50$, $\lambda_3 = 2{,}86$.

Подставляя последовательно численные значения λ_1, λ_2, λ_3 в систему и решая эти системы относительно неизвестных $l_i = (l_{i1}, l_{i2}, l_{i3})$ $(i = 1{,}2{,}3)$, получаем

$$l_{i1} = \begin{pmatrix} 0{,}8126 \\ 0{,}4955 \\ 0{,}3068 \end{pmatrix}, l_{i2} = \begin{pmatrix} -0{,}5454 \\ 0{,}8321 \\ 0{,}1006 \end{pmatrix}, l_{i3} = \begin{pmatrix} -0{,}2054 \\ -0{,}2491 \\ 0{,}9465 \end{pmatrix}$$

В качестве главных компонент получаем

$$z^{(1)} = 0{,}81x^{(1)} + 0{,}50x^{(2)} + 0{,}31x^{(3)};$$
$$z^{(2)} = -0{,}55x^{(1)} + 0{,}83x^{(2)} + 0{,}10x^{(3)};$$
$$z^{(3)} = -0{,}21x^{(1)} - 0{,}25x^{(2)} + 0{,}95x^{(3)};$$

Здесь под $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$ подразумеваются отклонения размеров длины $\left(x^{(1)}\right)$ ширины $\left(x^{(2)}\right)$ высоты $\left(x^{(3)}\right)$ панциря от своих средних значений.

Вычисление относительной доли суммарной дисперсии, обусловленной одной, двумя и тремя главными компонентами.

$$g(1) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0{,}9864$$
$$g(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 0{,}9958$$
$$g(3) = 1$$

Отсюда можно сделать вывод, что почти вся информация о специфике размеров панциря данного вида черепах содержится в одной лишь первой главной компоненте, которую и естественно использовать при соответствующей классификации исследуемых особей.

7.2. Центроидный метод

В качестве упрощенного аппроксимационного расчета метода главных факторов используется центроидный метод. До недавнего времени он был самым распространенным способом определения факторов. Так как раньше из-за большого объема вычислительных работ метод главных факторов применялся лишь в исключительных случаях, довольствовались центроидным методом, хотя он дает неточное и неоднозначное решение. И сегодня при отсутствии ЭВМ отдается предпочтение центроидному методу среди других процедур выделения факторов. При применении метода главных факторов и центроидного метода к одним и тем же данным выделенные первые факторы практически совпадают. Рекомендуется при использовании центроидного метода выделять на один фактор больше, чем это бы делалось с помощью метода главных факторов, и лишь после вращения заниматься интерпретацией всех факторов.

По сравнению с другими методами выделения факторов центроидный метод отличается простотой расчетов. Довольно легко воспринимается сама процедура вычислений. При этом создается впечатление, что другими

методами можно овладеть лишь после интенсивных занятий. Из-за этого дидактического преимущества метод особенно пригоден для начинающего исследователя.

Синонимом названия «центроидный метод» является «метод центра тяжести». Это название объясняет принцип метода. Положение первой координатной оси должно быть определено так, чтобы она проходила через центр тяжести скопления точек. Факторное отображение можно рассматривать как размещение m точек-переменных в r-мерном пространстве, причем отдельные точки или векторы представляют переменные. На рис. 2А схематично изображены несколько точек-переменных в двумерной системе координат. Кроме того, указана нулевая точка, в которой начинаются все векторы. Это соответствует типичной ситуации перед началом выделения факторов. Переменные представлены m точками в r-мерном пространстве, положение нулевой точки известно. Разумеется, точное значение необходимой размерности пространства неизвестно.

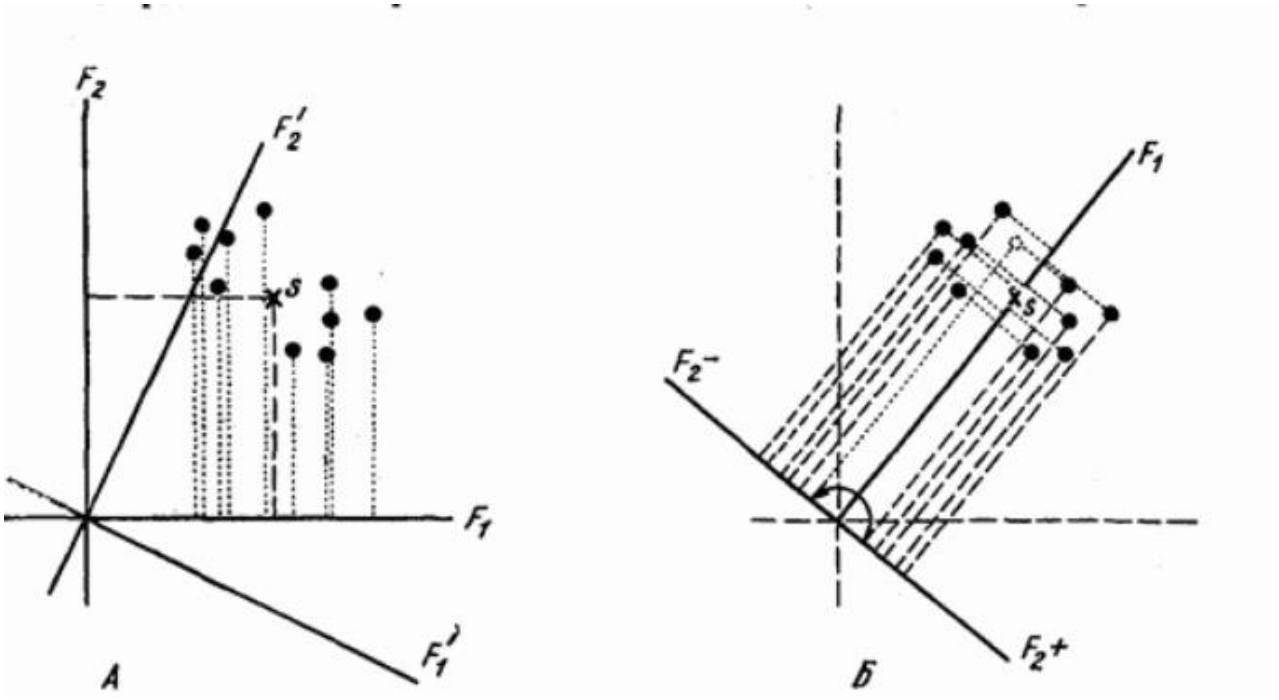

Рис. 2. Определение положения первой координатной оси с помощью центроидного метода. Диаграмма А: величина проекции центра тяжести s на F_1, является средним значением проекций всех точек на эту ось; конфигурация векторов не зависит от положения системы координат. Диаграмма Б: первая центроидная ось проводится через центр тяжести; тогда сумма остаточных проекций на ось F_2 равна нулю. Показано отражение одной точки переменной с положительной стороны на отрицательную отдельных точек.

Точки можно изобразить в очень многих ортогональных системах координат, из которых на рис. 2А представлены две — F_1F_2 и $F_1'F_2'$. Чтобы

получить однозначное положение системы координат, уславливаются, что первая ось должна проходить через центр тяжести S скопления точек-переменных. Вторая ось F_2, как показано на рис. 2Б, перпендикулярна к первой. Представим себе, что определено положение отдельных точек-переменных, центра тяжести S и нулевой точки (рис. 2А). Систему координат F_1F_2 можно повернуть так, что она, например, займет положение $F_1'F_2'$. Но результатом вращения должно быть такое ее положение, чтобы ось F_1 проходила через центр тяжести S, как показано на рис. 2Б. Это положение осей соответствует позиции факторов в центроидном решении.

В методе главных факторов для определения предпочтительной системы координат требовалось, чтобы вдоль первой оси лежал максимум дисперсии. В центроидном методе требуется, чтобы первая ось проходила через центр тяжести. Назначение обоих требований — попытаться однозначно определить положение системы координат.

Проекции точек на оси координат на рис. 2Б определяют факторные нагрузки a_{il}, которые рассчитываются по корреляционной матрице. Координаты центра тяжести можно вычислить по координатам отдельных точек.

Если в общем случае рассматривать r-мерную систему координат, то координатами центра тяжести являются выражения:

$$l/m \sum_i^m a_{i1} ; l/m \sum_i^m a_{i2} \dots l/m \sum_i^m a_{ir} \qquad (8)$$

т. е. средние значения координат отдельных точек дают координаты центра тяжести. Это можно увидеть, рассматривая 2А, на котором координаты центра тяжести отмечены пунктиром. Если теперь система координат выбрана так, что первая ось проходит через центр тяжести, то сумма проекций точек на все остальные ортогональные к ней оси равны нулю (это следует из определения центра тяжести) и тогда координаты центра тяжести S становятся равными:

$$l/m \sum_i^m a_{i1} , 0,0, \dots ,0 \qquad (9)$$

т.е.

$$\sum a_{i2} = \sum a_{i3} = \dots = \sum a_{ir} = 0$$

в чем можно убедиться по рис. 2Б для случая двумерной задачи. Сумма проекций на ось F_2 равна нулю, так как положительные и отрицательные значения проекций взаимно компенсируются. Условие (9) используется в расчетах по центроидному методу. Исходя из равенства $R_h = AA'$ можем

написать для каждого элемента k-го столбца матрицы R соответствующие выражения:

$$\begin{gathered} r_{1k} = a_{11}a_{k1} + a_{12}a_{k2} + \cdots + a_{1r}a_{kr} \\ r_{2k} = a_{21}a_{k1} + a_{22}a_{k2} + \cdots + a_{2r}a_{kr} \\ \cdots \\ r_{mk} = a_{m1}a_{k1} + a_{m2}a_{k2} + \cdots + a_{mr}a_{kr} \end{gathered} \quad (10)$$

$$\sum_i r_{ik} = a_{k1}\sum_i a_{i1} + a_{k2}\sum_i a_{i2} + \cdots + a_{kr}\sum_i a_{ir} \quad (11)$$

(11) представляет собой сумму равенств (10). Оно имеет место для каждого k-столбца корреляционной матрицы. Если теперь просуммировать все суммы столбцов, т. е. просуммировать обе части равенства (11) по всем k, то получим общую сумму T элементов корреляционной матрицы:

$$T = \sum_i \sum_k r_{ik} = \sum_k a_{k1} \sum_i a_{i1} + \sum_k a_{k2} \sum_i a_{i2} + \cdots \sum_k a_{kr} \sum_i a_{ir} \quad (12)$$

В связи с тем, что $\sum_i a_{i1} = \sum_k a_{k1}$ (перемена индекса не изменяет смысла суммирования), получаем

$$T = \sum_{ik} r_{ik} = (\sum_i a_{i1})^2 + (\sum_i a_{i2})^2 + \cdots (\sum_i a_{ir})^2 \quad (13)$$

т.е. сумма всех элементов корреляционной матрицы равна сумме квадратов сумм столбцов матрицы факторного отображения. Это равенство имеет место только для ортогонального факторного отображения.

Подставив в (13) условия (9), получим

$$T = (\sum a_{i1})^2 + 0 + 0 + \cdots \text{ или } \sqrt{T} = \sum_i a_{i1} \quad (14)$$

С учетом условия (9) равенство (11) примет вид:

$$\sum_i r_{ik} = a_{k1} \sum_i a_{i1} \quad (15)$$

Из (14) и (15) получим

$$\sum_i r_{ik} = a_{k1}\sqrt{T} \quad (16)$$

Введя обозначение $t = 1/\sqrt{T}$ выразим:

$$a_{k1} = t \sum_i r_{ik} \ \ (k = 1, \ldots, m) \quad (17)$$

или изменив индекс, получим

$$a_{i1} = t \sum_i r_{ki} \ \ (i = 1, \ldots, m)$$

Оба сомножителя правой стороны этого равенства легко определяются из R. Таким образом по (17) вычисляются нагрузки первого центроидного фактора. Равенство (14) служит для контроля правильности вычислений.

После вычисления нагрузок первого фактора по (17) определяют остаточные корреляции: $R_h - a_1 a_1' = R_1$, где a_1 является вектор-столбцом факторных нагрузок. Матрица $a_1 a_1' = R^+$ содержит так называемые воспроизведенные корреляции. R_1 дает остаточные корреляции, которые остаются после выделения первого фактора (R_1 — остаточная матрица). Если принимают решение выделить второй фактор, то повторяется та же самая вычислительная процедура по матрице остатков R_1. При этом возникает

затруднение, которое можно преодолеть с помощью некоторой уловки. Согласно определению после выделения фактора, сумма проекции всех точек на другие ортогональные оси равна нулю. Второй центр тяжести, который не совпадает с началом координат, нельзя поэтому определить и, стало быть, нельзя приступить к выделению другого фактора. Например, по рис. 2Б видно, что сумма проекций на ось F_2 равна нулю и совпадает с началом координат. Изменив знаки некоторых переменных таким образом, чтобы новый центр тяжести был удален от начала координат, создают предпосылку проведения вычислительной процедуры по выделению второго фактора. Изменение знака переменных нужно произвести так, чтобы все точки-переменные на рис. 2Б находились по одну сторону от оси F_1. Например, если изменить все отрицательные знаки на положительные, то получим новый центр тяжести, который не совпадает с началом координат и используется дальше для вычисления нагрузок второго центроидного фактора. На последующем этапе расчета изменение знака аннулируется. Изменение знаков, или так называемое «отражение» переменных, лучше всего объяснить на конкретном примере. В этом месте вычислительной процедуры центроидного метода играет определенную роль субъективизм исследователя и его опыт. Можно, конечно, выработать определенные твердые правила, исключающие субъективизм в принятии решения. Но несмотря на это, изменение знаков остается слабым местом центроидного метода. Предположим, второй фактор выделен. Затем опять определяется остаточная матрица $R_2 = R_1 - a_1 a_1'$ и принимается решение, выделять ли следующий фактор и т.д., пока последняя остаточная матрица не будет достаточно точно соответствовать нулевой матрице.

7.3. Критерии оценки числа факторов, подлежащих выделению

При применении как метода главных факторов, так и центроидного метода возникает один и тот же вопрос: когда должен быть закончен процесс выделения факторов или каким числом факторов можно удовлетвориться? Имеются различные пути решения этого вопроса, которые отчасти приводят к новым способам решения, отчасти также связаны с проблемой общности, проблемой вращения и оценкой значений факторов, т. е. с теми проблемами, которые нами пока затрагивались поверхностно.

Общепризнанного метода определения числа факторов, подлежащих выделению, не существует. Представители различных школ расходятся в мнении о том, какой метод является более достоверным и пригодным для практики. К настоящему времени разработано более двадцати способов

определения числа выделяемых факторов. В основном различают три подхода при решении задачи о числе выделяемых факторов:

1) алгебраический подход, который сводится к определению ранга R;

2) статистический подход, при котором на передний план выдвигается возможность сделать заключение на определенном уровне значимости о всей генеральной совокупности индивидуумов;

3) психометрический подход, при котором добиваются обобщения на совокупность всех переменных, и отчасти этот подход аргументирован с общенаучных позиций.

Каждый из этих подходов можно проследить в большинстве имеющихся способах оценки числа факторов, подлежащих выделению. Перечисленные подходы и их возможные комбинации определяют многообразие созданных вычислительных процедур.