УДК 528.08.519.25

Н. А. ЛОШКАРЕВ

ФАКТОРНЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ ГЕОДЕЗИЧЕСКИХ ИЗМЕРЕНИЙ

Измерительные системы в геодезии могут быть отнесены к так называемым многопараметрическим «объектам». Трудности их оптимизации связаны именно с большим числом параметров. При исследовании этих систем встает задача уменьшения числа параметров то ли путем отбрасывания некоторых из них, то ли заменой меньшим числом функций от них с условием сохранения всей информации об «объекте».

Задача может быть успешно решена методом главных компонент и факторным методом. В обоих случаях производится анализ внутренней структуры матриц ковариаций (или корреляций). В компонентном методе ковариационная матрица расчленяется на совокупность ортогональных векторов (компонент) по числу переменных. Ясно, что для точного воспроизведения корреляций между переменными требуются все компоненты, хотя значительную часть дисперсии переменных могут выделить лишь несколько из них.

В противоположность методу главных компонент в факторном анализе корреляционная матрица объясняется наличием некоторого небольшого числа предполагаемых переменных, или факторов. Если исследуемая матрица отличается от единичной, то можно поставить вопрос, существует ли случайная величина f_1, такая, что попарные корреляции между переменными равны нулю, когда влияние f_1 уже учтено. Если f_1 не объясняет корреляции, решается вопрос, существуют ли две случайные величины $f_1 f_2$, такие, что попарные корреляции переменных равны нулю, когда влияние f_1 и f_2 уже учтено, и так далее [3]. Отсюда ясно, что если метод главных компонент ориентирован на дисперсии, то факторный анализ ориентирован на ковариации (или корреляционную связь).

Если $x_1, x_2 \ldots x_p$ суть p наблюдаемых переменных, то при компонентном анализе к ним применяется ортогональное преобразование, так что получают новые p переменных $y_1, y_2 \ldots y_p$ при условии, что величины y_i имеют максимальные дисперсии и не коррелированы между собой.

В факторном анализе основное предположение состоит в том, что

$$x_i = \sum_{r=1}^{k} l_{ir} \cdot f_r + e_i; \quad (i = 1, 2, 3 \ldots p) \ldots \tag{1}$$

где f_r — простой r-й фактор, k точно задано, а e_i — остатки, представляющие источники отклонений, действующие лишь на x_i.

Заметим, что p случайных величин e_i не коррелированы как между собой, так и с k величинами f_r. Можно считать дисперсии всех f_r равными единице. Дисперсии величин e_i обозначим через v_i. Коэффициенты l_{ir} принято называть нагрузкой r-го фактора в i-й переменной. Значения l_{ir} обычно не известны и подлежат оценке.

Для вывода уравнений оценок предположим, что x_i подчиняются многомерному нормальному распределению, а их дисперсии и ковариации об-

разуют $p \times p$ матрицу $C = |c_{ij}|$. Простые факторы f_z не коррелированы. Из уравнения (1) следует, что

$$c_{ii} = \sum_{r=1}^{k} l_{ir}^2 + v_i \ldots\ldots \quad (2)$$

$$c_{ij} = \sum_{r=1}^{k} l_{ir} \cdot l_{jr}; \qquad (i \neq j) \quad (3)$$

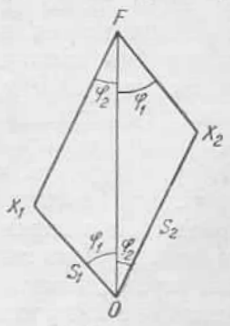
или $C = LL' + V$ ... (3) в матричной записи.

При этом $L = |l_{ir}|$ является $p \times k$ матрицей нагрузок, V — диагональная матрица с элементами v_i.

Основная модель заключается в выборе гипотезы H_0 о корреляционной матрице C. Эта матрица может быть представлена в виде суммы диагональной матрицы с положительными элементами и матрицы ранга k с положительными собственными значениями. Так как v_i не известны, это приводит к ограничению числа k вида

$$(p + k) < (p - k)^2 \ldots\ldots \quad (4)$$

Пусть $A = |a_{ij}|$ — выборочная ковариационная матрица с выборочными оценками ковариаций и дисперсий x_i с n степенями свободы (размер выборки $(n + 1)$) в качестве элементов матрицы. Задача состоит в том, чтобы получить состоятельные и эффективные оценки параметров l_{ir} и v_i, используя информацию, заключенную в A. Для решения этой задачи воспользуемся аппроксимационным центроидным методом оценки нагрузок. Этот метод, несмотря на простоту, дает оценки, весьма близкие к оценкам, получаемым по методу максимального правдоподобия [2], и достаточен для практических целей. Геометрическая модель центроидного метода может быть представлена векторами p-мерного векторного пространства, косинусы углов между которыми равны корреляциям, а длины векторов — стандартным отклонениям соответствующих переменных. Если изменить знаки переменных так, чтобы число положительных корреляций стало наибольшим, то векторы будут иметь тенденцию к группировке в пучок. После этого первый фактор (центроид) системы определяется как сумма векторов, он будет проходить через середину пучка. Учтя влияние первого центроида и изменив знаки, как и в первом случае, можно найти второй центроид и т. д. до тех пор, пока дисперсия переменных будет исчерпана.



Геометрическая интерпретация центроидного метода.

Рисунок поясняет метод при двух переменных x_1 и x_2 с дисперсиями s_1^2 и s_2^2 и коэффициентом корреляции r. Их ковариационная матрица равна

$$A = \begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{bmatrix}$$

На рисунке $OX_1 = s_1$; $OX_2 = s_2$; $\cos 0 = r$; $0 = \varphi_1 + \varphi_2$.

Суммарный вектор OF после приведения к единичной длине представляет первый фактор f_1.

Определяем нагрузки x_1 и x_2 на f_1.

$$l_{11} = s_1 \cos \varphi_1,$$
$$l_{21} = s_2 \cos \varphi_2 \ldots\ldots \quad (5)$$

Обозначая вектор нагрузок через $\mathbf{1}_1$, находим матрицу остаточных

ковариаций после учета влияния первого фактора, или

$$\begin{bmatrix} s_1^2 - (s_1 \cos\varphi_1)^2 & rs_1s_2 - s_1s_2 \cos\varphi_1 \cos\varphi_2 \\ rs_1s_2 - s_1s_2 \cos\varphi_1 \cos\varphi_2 & s_2^2 - (s_2 \cos\varphi_2)^2 \end{bmatrix} =$$

$$= \begin{bmatrix} s_1^2 \sin^2\varphi_1 & -s_1s_2 \sin\varphi_1 \sin\varphi_2 \\ -s_1s_2 \sin\varphi_1 \sin\varphi_2 & s_2^2 \sin^2\varphi_2 \end{bmatrix},$$

так как $r = \cos(\varphi_1 + \varphi_2)$.

Таким образом, в остаточной матрице сумма строк и столбцов равна нулю.

Уравнения (5) дают возможность вычислять нагрузки переменных на факторы прямо из ковариационной матрицы

$$l_{11} = s_1 \cos\varphi_1 = \frac{s_1^2 + rs_1s_2}{\sqrt{s_1^2 + s_2^2 + 2rs_1s_2}} \quad \ldots\ldots \qquad (6)$$

Здесь числитель равен сумме элементов в первом столбце матрицы A, а знаменатель — квадратному корню из суммы всех ее элементов. Для получения нагрузок на второй фактор необходимо изменить знак хотя бы одного переменного, то есть в одной строке и одном столбце. Знаки переменных изменяют до тех пор, пока получают минимальное число отрицательных значений в остаточной матрице. В некоторых случаях изменение знаков может быть целесообразным еще до оценки первых факторных нагрузок. Так как центроидные нагрузки зависят от размерностей переменных, то переменные нормируются так, что матрица A становится корреляционной. Следует заметить, что в этом случае имеются трудности при рассмотрении критериев достаточности. В центроидном методе возникает усложнение вследствие замены диагональных элементов матрицы A (то есть полных дисперсий) меньшими числами факторных дисперсий. Факторные дисперсии представляют долю полной дисперсии переменных, обусловленную действием k простых факторов.

Поскольку ни число факторов, ни факторные дисперсии не известны, то последние обычно вначале задаются. Принято выбирать их как наибольшие корреляции в каждом столбце матрицы (в ковариационной матрице наибольший коэффициент корреляции в столбце умножается на дисперсию, стоящую в том же столбце по диагонали). Дальнейшее изложение будем вести, используя пример факторного анализа данных измерения длин светодальномером МСД-1, опубликованных в «Геодезии и картографии» № 9 в 1968 году.

В качестве переменных приняты шесть разностей длин от эталонного значения в шести приемах каждой из 12 измеренных линий. В табл. 1 даны корреляции между переменными с грубоприближенными оценками факторных дисперсий по диагонали и нагрузки на первый центроид. В этой же таблице приведены элементы остаточной матрицы $A_0 - \mathbf{1} \cdot \mathbf{1}'$, где $\mathbf{1}'$ — строка нагрузок в табл. 1, а A_0 — матрица с первоначальными факторными дисперсиями. Изменяя знаки у переменных 1, 2 и 3 и заменив значения по диагонали наибольшими числами в каждом столбце, получаем табл. 2. По первым оценкам центроидных нагрузок новые оценки факторных дисперсий получают суммированием квадратов нагрузок для каждой переменной по всем k факторам.

Согласно неравенству (4) при шести переменных можно ограничиться $k = 2$, так как нахождение нагрузок на третий фактор не имеет смысла. Дальнейшие вычисления повторяют первую итерацию с той разницей, что в начальную матрицу подставляют факторные дисперсии, полученные в первом приближении.

Таблица 1

Матрица корреляций систематических погрешностей измерения длин шестью приемами

Приемы	1	2	3	4	5	6	Сумма
1	0,578 (0,033)	+0,573 (0,076)	+0,306 (—0,121)	+0,545 (—0,067)	+0,389 (—0,040)	+0,389 (+0,119)	+2,687 (0,000)
2	+0,573 (0,076)	0,735 (0,280)	+0,735 (0,342)	+0,398 (—0,164)	+0,064 (—0,250)	—0,035 (—0,284)	+2,470 (0,000)
3	+0,306 (—0,121)	+0,735 (0,342)	0,735 (—0,397)	+0,387 (—0,096)	+0,020 (—0,250)	—0,058 (—0,272)	+2,125 (0,000)
4	+0,545 (—0,067)	+0,898 (—0,164)	+0,387 (—0,096)	0,700 (0,009)	+0,570 (0,184)	+0,440 (0,134)	+3,040 (0,000)
5	+0,301 (—0,040)	+0,064 (—0,250)	+0,020 (—0,250)	+0,570 (0,184)	+0,570 (0,355)	+0,171 (0,000)	+1,696 (0,000)
6	+0,389 (0,119)	—0,035 (—0,284)	—0,058 (—0,272)	+0,440 (0,134)	+0,171 (0,000)	0,440 (0,303)	+1,347 (0,000)
Сумма	+2,687	+2,470	+2,125	+3,040	+1,696	1,347	13,365
		$\sqrt{13,365} = 3,656$					
Нагрузки на первый центроид	0,735	00,676	0,581	0,832	0,464	0,368	3,656

Таблица 2

Преобразованная остаточная матрица корреляций и вычисление нагрузок второго центроида

Приемы	1	2	3	4	5	6	Сумма
1	(0,076)	+0,076	—0,121	+0,067	+0,040	—0,119	+0,019
2	+0,076	(0,342)	+0,342	+0,164	+0,250	+0,284	+1,458
3	—0,121	+0,342	(0,272)	+0,096	+0,250	+0,272	+1,111
4	+0,067	+0,164	+0,096	(0,184)	+0,184	+0,134	+0,829
5	+0,040	+0,250	+0,250	+0,184	(0,250)	0,000	+0,974
6	—0,119	+0,284	+0,272	+0,134	0,000	(0,284)	+0,855
Сумма	0,019	+1,458	+1,111	+0,829	+0,984	+0,855	+5,246
		$\sqrt{5,246} = 2,290$					
Нагрузки на второй центроид	0,008	0,637	0,485	0,362	0,425	0,373	2,290

В [2] подробно рассмотрен вопрос упрощения последующих итераций, как только выяснится, что смена законов становится необходима на каждой стадии итерационного процесса. В табл. 3 приведены значения нагрузок и факторные дисперсии после четырех итераций.

Таблица 3

Окончательные оценки факторных нагрузок и дисперсий

Переменные	1	2	3	4	5	6
Первый центроид	0,732	0,730	0,534	0,907	0,408	0,312
Второй центроид	—0,030	—0,587	—0,538	0,359	0,375	0,443
Факторные дисперсии	0,546	0,878	0,574	0,952	0,307	0,293

По данным табл. 3 вычисляется итоговый вклад каждого фактора в суммарную дисперсию шести переменных. Для этого находят сумму квадратов

нагрузок по обоим факторам и относят ее к сумме дисперсии шести переменных (то есть *б*). Вклад первого фактора в общую дисперсию 40,7, а второго — 18,4%.

Исследование эффективности оценок, выполненное Д. Лоули [2], показало, что эффективность метода обычно высокая. Для проверки значимости остатков после исключения центроидных факторов используется критерий

$$\frac{1}{2} ntr\,(X^2)\ldots, \tag{7}$$

где $X = (V^{-1} - V^{-1}LI^{-1}L'V^{-1})(A - C)$ с числом степеней свободы

$$\frac{1}{2}(p-k)(p-k+1),$$

где $I = L'V^{-1}L$.

Заменяя неизвестные параметры их оценками и уменьшая число степеней свободы на p из-за оценки остаточных дисперсий, получаем

$$X = V^{-1}(A_0 - LM')\ldots, \tag{8}$$

где

$$M' = I^{-1}L'V^{-1}A_0\ldots, \tag{9}$$

с числом степеней свободы

$$\frac{1}{2}[(p-k)^2 - (p+k)].$$

Критерий (7) считается распределенным приближенно как χ^2, а степень приближения зависит не только от размера выборки $(n+1)$, но и от числа степеней свободы.

Окончательно критерий (7) вычисляется в виде

$$n\left(\frac{1}{2}\sum_i x_{ii}^2 + \sum_{i<j} x_{ij}\cdot x_{ji}\right). \tag{10}$$

Вследствие того, что размер выборки $(n = 11)$ слишком мал, рассмотренный пример имеет лишь методический характер. Факторный анализ эффективен при размере выборки хотя бы в несколько десятков, а обычно n колеблется от одной до нескольких сотен, так как только в этом случае элементы ковариационных (или корреляционных) матриц известны с достаточной точностью.

Один из основных моментов факторного анализа — интерпретация результатов. В нашем примере почти все корреляции положительны; это значит, что разным длинам соответствуют разные систематические погрешности, так что если систематическая погрешность первого приема выше средней по всем длинам, то и остальные пять будут выше средней и наоборот. Последнее отражает влияние данных условий измерений на величину систематической ошибки вообще, хотя одни условия более благоприятны, а другие — менее. Первый фактор, все нагрузки на который положительны, объясняет теперь эти колебания условий измерений. Когда влияние данного фактора учтено, обнаруживается различие систематических погрешностей в первых трех и трех последующих приемах.

Действие второго фактора таково, что если в первых трех полуприемах систематические погрешности выше средних, то в трех остальных они оказываются ниже средних и наоборот. Наличие этого биполярного фактора, возможно, связано с тем, что большинство длин (9 из 12) измерялось отдельными сериями по три приема.

Таким образом, применение факторного анализа для исследования геодезических измерений имеет достаточные основания, как в отношении

формы, так и в отношении содержания. Особенно перспективно применение компонентного и факторного анализов для исследования сложных, многопараметрических измерительных систем (измерение длин дальномерами разных типов и в разных условиях, измерения углов разными инструментами с разными визирными целями, разными наблюдателями и т. д.), а также для практического упрощения задач оптимизации измерительных систем.

ЛИТЕРАТУРА

1. Видуев Н. Г., Кондра Г. С. Дисперсионный анализ в теории и практике геодезических измерений. «Недра», М., 1968.
2. Лоули Д., Максвелл А. Факторный анализ как статистический метод. «Мир», М., 1967.
3. Howe W. G. Some contribution to factor analysis, USAEC Rep. ORNL-1919. 1955.

Работа поступила 24 декабря 1970 года.
Рекомендована кафедрой инженерной геодезии
Днепропетровского инженерно-строительного института.