1 Факторная модель для стандартизованных признаков

В данном разделе мы рассмотрим факторную модель для случая, когда изучаемые признаки являются случайными величинами с нулевыми математическими ожиданиями и единичными дисперсиями. Такая стандартизация признаков позволяет упростить анализ и интерпретацию результатов, так как все признаки находятся в одинаковом масштабе.

1.1 Модель и основные предположения

Обозначим через $\mathbf{z}$ случайный вектор размерности k, подчиняющийся совместному нормальному закону распределения с нулевым математическим ожиданием и ковариационной матрицей $V(\mathbf{z})$, на диагонали которой стоят единицы:

$$\mathbf{z} = \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \vdots \\ z^{(k)} \end{bmatrix}, \quad M(\mathbf{z}) = \begin{bmatrix} M(z^{(1)}) \\ M(z^{(2)}) \\ \vdots \\ M(z^{(k)}) \end{bmatrix} = \mathbf{0},$$

$$V(\mathbf{z}) = M(\mathbf{z}\mathbf{z}^T) = \begin{bmatrix} 1 & M(z^{(1)}z^{(2)}) & \dots & M(z^{(1)}z^{(k)}) \\ M(z^{(2)}z^{(1)}) & 1 & \dots & M(z^{(2)}z^{(k)}) \\ \vdots & \vdots & \ddots & \vdots \\ M(z^{(k)}z^{(1)}) & M(z^{(k)}z^{(2)}) & \dots & 1 \end{bmatrix},$$

что означает $\mathrm{Var}(z^{(j)}) = 1$ для всех $j = 1, \dots, k$.

Пусть $\mathbf{f}$ и $\mathbf{u}$ – случайные векторы размерности m и k соответственно, также подчиняющиеся совместным нормальным законам распределения с нулевыми математическими ожиданиями:

$$M(\mathbf{f}) = \mathbf{0}, \quad M(\mathbf{u}) = \mathbf{0}, \tag{1}$$

и единичными ковариационными матрицами:

$$V(\mathbf{f}) = M(\mathbf{f}\mathbf{f}^T) = \mathbf{E}_m, \quad V(\mathbf{u}) = M(\mathbf{u}\mathbf{u}^T) = \mathbf{E}_k, \tag{2}$$

где $\mathbf{E}_m$ и $\mathbf{E}_k$ – единичные матрицы размерности $m \times m$ и $k \times k$ соответственно. Кроме того, векторы $\mathbf{f}$ и $\mathbf{u}$ не коррелируют друг с другом:

$$M(\mathbf{f}\mathbf{u}^T) = \mathbf{0}, \quad M(\mathbf{u}\mathbf{f}^T) = \mathbf{0}. \tag{3}$$

Основным предположением факторной модели является то, что каждая из величин $z^{(j)}$, $j = 1, \dots, k$, линейно зависит от общих факторов $f^{(s)}$,

$s = 1, \ldots, m$, и характерного фактора $u^{(j)}$:

$$\begin{cases} z^{(1)} = a_1^{(1)} f^{(1)} + a_1^{(2)} f^{(2)} + \cdots + a_1^{(m)} f^{(m)} + d_1 u^{(1)}, \\ z^{(2)} = a_2^{(1)} f^{(1)} + a_2^{(2)} f^{(2)} + \cdots + a_2^{(m)} f^{(m)} + d_2 u^{(2)}, \\ \vdots \\ z^{(k)} = a_k^{(1)} f^{(1)} + a_k^{(2)} f^{(2)} + \cdots + a_k^{(m)} f^{(m)} + d_k u^{(k)}. \end{cases} \tag{4}$$

Здесь коэффициенты $a_j^{(s)}$ и d_j называются общими и характерными факторными нагрузками соответственно. В матричной форме модель (4) записывается как:

$$\mathbf{z} = \mathbf{A}\mathbf{f} + \mathbf{D}\mathbf{u}, \tag{5}$$

где $\mathbf{A}$ – матрица общих факторных нагрузок размерности $k \times m$, а $\mathbf{D}$ – диагональная матрица характерных факторных нагрузок размерности $k \times k$.

1.2 Корреляционная матрица и воспроизведенная матрица корреляций

Поскольку $\mathrm{Var}(z^{(j)}) = 1$ для всех $j = 1, \ldots, k$, ковариационная матрица $V(\mathbf{z})$ совпадает с матрицей корреляций $\mathbf{R}$:

$$V(\mathbf{z}) = \mathbf{R} = \begin{bmatrix} 1 & r_{z^{(1)}z^{(2)}} & \ldots & r_{z^{(1)}z^{(k)}} \\ r_{z^{(2)}z^{(1)}} & 1 & \ldots & r_{z^{(2)}z^{(k)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{z^{(k)}z^{(1)}} & r_{z^{(k)}z^{(2)}} & \ldots & 1 \end{bmatrix}.$$

В рамках факторной модели (5) корреляционная матрица $\mathbf{R}$ может быть представлена в виде:

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \mathbf{D}^2. \tag{6}$$

Матрица $\mathbf{A}\mathbf{A}^T$ называется воспроизведенной матрицей корреляций и обозначается как $\mathbf{R}^*$:

$$\mathbf{R}^* = \mathbf{A}\mathbf{A}^T = \begin{bmatrix} 1 - d_1^2 & r_{z^{(1)}z^{(2)}} & \ldots & r_{z^{(1)}z^{(k)}} \\ r_{z^{(2)}z^{(1)}} & 1 - d_2^2 & \ldots & r_{z^{(2)}z^{(k)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{z^{(k)}z^{(1)}} & r_{z^{(k)}z^{(2)}} & \ldots & 1 - d_k^2 \end{bmatrix}. \tag{7}$$

Элементы на диагонали матрицы $\mathbf{R}^*$ представляют собой общности $h_j^2 = 1 - d_j^2$, которые показывают долю дисперсии признака $z^{(j)}$, объясненную общими факторами.

1.3 Подробное рассмотрение свойств факторной модели

В этом разделе мы подробно рассмотрим свойства факторной модели, приведем доказательства некоторых из них и выведем соответствующие формулы.

1.3.1 Свойство 1A: Общности и характерности признаков

Формулировка: Общности и характерности признаков связаны уравнением:

$$h_j^2 + d_j^2 = 1, \quad j = 1, \ldots, k. \tag{8}$$

Общность h_j^2 определяется как сумма квадратов общих факторных нагрузок:

$$h_j^2 = \left(a_j^{(1)}\right)^2 + \cdots + \left(a_j^{(m)}\right)^2, \tag{9}$$

а характерность d_j^2 – как квадрат характерной факторной нагрузки.

Доказательство: Из факторной модели (4) для каждого признака $z^{(j)}$ имеем:

$$z^{(j)} = a_j^{(1)} f^{(1)} + \cdots + a_j^{(m)} f^{(m)} + d_j u^{(j)}.$$

Поскольку $\mathrm{Var}(z^{(j)}) = 1$, а факторы $f^{(s)}$ и $u^{(j)}$ некоррелированы и имеют единичные дисперсии, дисперсия $z^{(j)}$ может быть выражена как:

$$\mathrm{Var}(z^{(j)}) = \left(a_j^{(1)}\right)^2 + \cdots + \left(a_j^{(m)}\right)^2 + d_j^2 = h_j^2 + d_j^2 = 1.$$

Таким образом, $h_j^2 + d_j^2 = 1$.

1.3.2 Свойство 2A: Общности и характерности находятся в пределах от нуля до единицы

Формулировка: Общности и характерности находятся в пределах от нуля до единицы:

$$0 \le h_j^2 \le 1, \quad 0 \le d_j^2 \le 1, \quad j = 1, \ldots, k. \tag{10}$$

Доказательство: Из свойства 1A следует, что $h_j^2 = 1 - d_j^2$. Поскольку d_j^2 – это квадрат характерной факторной нагрузки, он не может быть отрицательным, и, следовательно, $0 \le d_j^2 \le 1$. Аналогично, $h_j^2 = 1 - d_j^2$ также находится в пределах от 0 до 1.

1.3.3 Свойство 3A: Вклад общего фактора в суммарную дисперсию признаков

Формулировка: Сумма квадратов элементов матрицы $\mathbf{A}$ по столбцу с номером s показывает вклад общего фактора $f^{(s)}$ в суммарную дисперсию признаков $z^{(j)}$. Сумма квадратов всех элементов матрицы $\mathbf{A}$ равна сумме общностей и показывает долю суммарной дисперсии, объясненную общими факторами.

Доказательство: Рассмотрим сумму квадратов элементов матрицы $\mathbf{A}$ по столбцу с номером s:

$$\sum_{j=1}^{k} \left(a_j^{(s)}\right)^2.$$

Эта сумма показывает, насколько сильно фактор $f^{(s)}$ влияет на все признаки $z^{(j)}$. Сумма квадратов всех элементов матрицы $\mathbf{A}$ равна:

$$\sum_{s=1}^{m} \sum_{j=1}^{k} \left(a_j^{(s)}\right)^2 = \sum_{j=1}^{k} h_j^2,$$

что соответствует сумме общностей, то есть доле суммарной дисперсии, объясненной общими факторами.

1.3.4 Свойство 4A: Доля суммарной дисперсии, объясненная общими факторами

Формулировка: Доля суммарной дисперсии признаков, объясненная общими факторами, вычисляется как:

$$\delta = \frac{\sum_{j=1}^{k} h_j^2}{k}. \tag{11}$$

Доказательство: Суммарная дисперсия всех признаков $z^{(j)}$ равна k, так как $\mathrm{Var}(z^{(j)}) = 1$ для всех j. Доля суммарной дисперсии, объясненная общими факторами, равна сумме общностей, деленной на суммарную дисперсию:

$$\delta = \frac{\sum_{j=1}^{k} h_j^2}{k}.$$

1.3.5 Свойство 5A: Воспроизведенная матрица корреляций

Формулировка: Воспроизведенная матрица корреляций может быть записана в виде:

$$\mathbf{R}^* = \mathbf{A}\mathbf{A}^T = \begin{bmatrix} h_1^2 & r_{z^{(1)}z^{(2)}} & \dots & r_{z^{(1)}z^{(k)}} \\ r_{z^{(2)}z^{(1)}} & h_2^2 & \dots & r_{z^{(2)}z^{(k)}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{z^{(k)}z^{(1)}} & r_{z^{(k)}z^{(2)}} & \dots & h_k^2 \end{bmatrix}. \tag{12}$$

Доказательство: Из факторной модели (5) и свойств ковариационных матриц следует:

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \mathbf{D}^2.$$

Поскольку $\mathbf{D}^2$ – диагональная матрица с элементами d_j^2, то $\mathbf{A}\mathbf{A}^T$ представляет собой матрицу корреляций, где на диагонали стоят общности $h_j^2 = 1 - d_j^2$, а вне диагонали – корреляции между признаками.

1.3.6 Свойство 6A: Корреляция признака и общего фактора

Формулировка: Элемент $a_j^{(s)}$ матрицы $\mathbf{A}$ показывает корреляцию признака $z^{(j)}$ и общего фактора $f^{(s)}$:

$$\text{Cov}(z^{(j)}, f^{(s)}) = r_{z^{(j)}f^{(s)}} = a_j^{(s)}, \quad j = 1, \dots, k; \quad s = 1, \dots, m. \tag{13}$$

Доказательство: Из факторной модели (4) и свойств ковариации следует:

$$\text{Cov}(z^{(j)}, f^{(s)}) = \text{Cov}\left(a_j^{(1)} f^{(1)} + \dots + a_j^{(m)} f^{(m)} + d_j u^{(j)}, f^{(s)}\right) = a_j^{(s)},$$

так как $\text{Cov}(f^{(s)}, f^{(s)}) = 1$ и $\text{Cov}(f^{(s)}, f^{(t)}) = 0$ при $s \neq t$, а также $\text{Cov}(f^{(s)}, u^{(j)}) = 0$.

1.3.7 Свойство 7A: Матрица корреляции признаков и общих факторов

Формулировка: Матрица факторных нагрузок $\mathbf{A}$ совпадает с матрицей корреляции признаков $\mathbf{z}$ и общих факторов $\mathbf{f}$:

$$\mathbf{R}(\mathbf{z}, \mathbf{f}) = \mathbf{A}. \tag{14}$$

Доказательство: Из свойства 6A следует, что каждый элемент матрицы $\mathbf{A}$ равен корреляции соответствующего признака и общего фактора. Таким образом, матрица $\mathbf{A}$ совпадает с матрицей корреляции $\mathbf{R}(\mathbf{z}, \mathbf{f})$.

1.4 Ортогональное вращение факторов

Для улучшения интерпретации факторов может быть использовано ортогональное вращение. Пусть $\mathbf{Q}$ – ортогональная матрица, тогда новые общие факторы $\tilde{\mathbf{f}}$ и новая матрица факторных нагрузок $\tilde{\mathbf{A}}$ определяются как:

$$\tilde{\mathbf{f}} = \mathbf{Q}^T\mathbf{f}, \quad \tilde{\mathbf{A}} = \mathbf{A}\mathbf{Q}.$$

Методы ортогонального вращения, такие как Quartimax и Varimax, позволяют максимизировать интерпретируемость факторов. Метод Quartimax максимизирует дисперсию квадратов факторных нагрузок:

$$\Phi(\tilde{\mathbf{A}}) = \frac{1}{km}\sum_{s=1}^{m}\sum_{j=1}^{k}\left(\tilde{a}_j^{(s)}\right)^2 - \left[\frac{1}{km}\sum_{s=1}^{m}\sum_{j=1}^{k}\tilde{a}_j^{(s)}\right]^2.$$

Метод Varimax максимизирует сумму дисперсий квадратов факторных нагрузок для каждого фактора:

$$\Phi(\tilde{\mathbf{A}}) = \frac{1}{k}\sum_{s=1}^{m}\left[\sum_{j=1}^{k}\left(\tilde{a}_j^{(s)}\right)^2 - \left(\frac{1}{k}\sum_{j=1}^{k}\tilde{a}_j^{(s)}\right)^2\right].$$

Эти методы реализованы в большинстве статистических пакетов и позволяют получить более интерпретируемую факторную структуру.

2 Пример факторного анализа

Рассмотрим применение факторного анализа на примере исследования потребительских предпочтений при выборе горнолыжных курортов. Опрос проводился среди отдыхающих, которые оценивали важность различных характеристик курорта по 9-балльной шкале: 1 – «неважно», 9 – «очень важно». В исследовании использовались следующие переменные:

- **цена** – стоимость одного дня катания;
- **подъем** – скорость подъемников;
- **снег** – толщина снежного покрытия на трассе;
- **влажность** – влажность снежного покрытия;
- **трасса** – протяженность трасс на курорте.

2.1 Матрица корреляций

На первом этапе анализа была построена матрица корреляций $\hat{\mathbf{R}}$ для пяти переменных (см. Рис. 1). Этот шаг является ключевым для понимания взаимосвязей между переменными. Матрица корреляций позволяет оценить степень линейной зависимости между каждой парой переменных. Значимые на уровне 0.5 корреляции выделены более темным или светлым цветом, что помогает визуально идентифицировать сильные связи. Наличие значимых корреляций между переменными является основанием для применения методов факторного анализа, так как это указывает на возможность существования скрытых факторов, которые влияют на наблюдаемые переменные.

2.2 Оценка общностей и дисперсий главных компонент

Первоначальные оценки общностей для пяти стандартизованных переменных $Z^{(1)}$ (цена), $Z^{(2)}$ (подъем), $Z^{(3)}$ (снег), $Z^{(4)}$ (влажность), $Z^{(5)}$ (трасса) были получены методом регрессионного анализа и составили 0.557, 0.658, 0.458, 0.477, 0.502 соответственно. Общности показывают долю дисперсии каждой переменной, которая может быть объяснена общими факторами. Этот шаг важен для понимания того, насколько каждая переменная связана с общими факторами. Высокие значения общностей указывают на то, что переменные хорошо объясняются выделенными факторами.

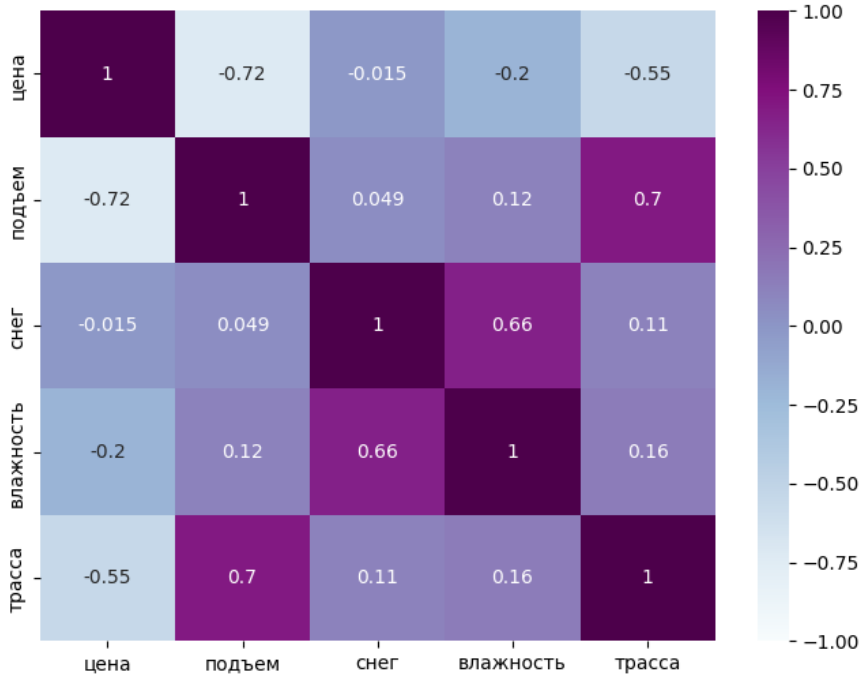

Рис. 1: Матрица корреляции $\hat{\mathbf{R}}$

Далее были рассчитаны дисперсии главных компонент, извлекаемых из матрицы корреляций. Они оказались равными 2.417, 1.572, 0.480, 0.315, 0.216 (см. Рис. 2). На основе критерия Кайзера (который рекомендует оставлять компоненты с дисперсией больше 1) и критерия «каменистой осыпи» было принято решение о построении модели с двумя общими факторами. Критерий Кайзера и метод «каменистой осыпи» помогают определить оптимальное количество факторов, которые следует оставить для анализа. Это позволяет сократить размерность данных, сохраняя при этом максимальное количество информации.

2.3 Матрица факторных нагрузок

Методом главных факторов была получена оценочная матрица факторных нагрузок $\hat{\mathbf{A}}$ для двух общих факторов (см. Табл. 1). Матрица факторных нагрузок показывает, насколько каждая переменная связана с каждым из факторов. Этот шаг позволяет понять, какие переменные вносят наибольший вклад в каждый фактор. В таблице также приведены оценки общностей h_j^2 и вклады общих факторов в объяснение суммарной дисперсии стандартизованных переменных.

Выделенные общие факторы объясняют 67.18% суммарной диспер-

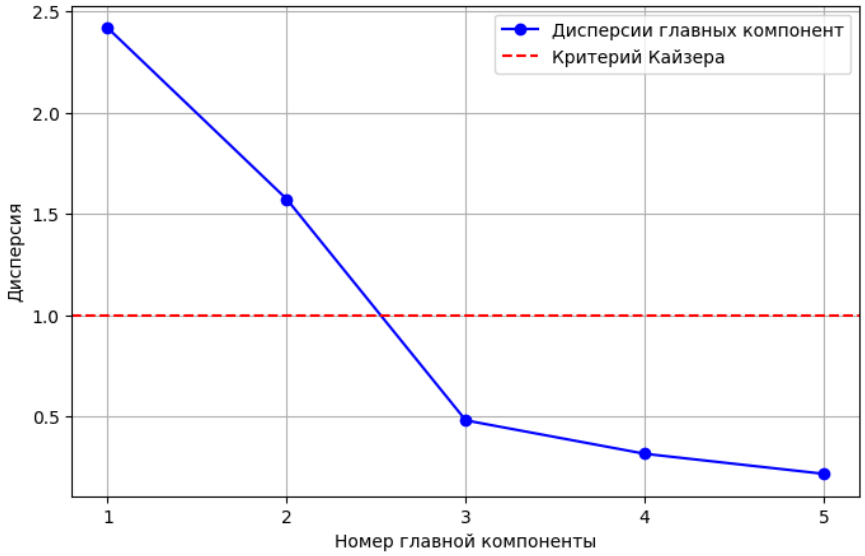

Рис. 2: Дисперсии главных компонент

Таблица 1: Матрица факторных нагрузок $\hat{\mathbf{A}}$

	$f^{(1)}$	$f^{(2)}$	h_j^2
$Z^{(1)}$ (цена)	-.741	.174	.580
$Z^{(2)}$ (подъем)	.912	-.280	.910
$Z^{(3)}$ (снег)	.263	.760	.647
$Z^{(4)}$ (влажность)	.377	.737	.686
$Z^{(5)}$ (трасса)	.721	-.130	.537
Вклады	2.113	1.246	3.359

сии стандартизованных переменных:

$$\delta = \frac{3.359}{5} = 0.6718.$$

Все общности превышают 50%, что свидетельствует о хорошем качестве факторного решения:

$$h_j^2 > 0.5, \quad j = 1, \ldots, 5.$$

2.4 Воспроизведенная матрица корреляций и матрица остатков

Для проверки качества факторного решения были построены воспроизведенная матрица корреляций $\hat{\mathbf{R}}^* = \hat{\mathbf{A}}\hat{\mathbf{A}}^T$ (см. Рис. 3) и матрица остат-

ков $\hat{\mathbf{R}} - \hat{\mathbf{R}}^*$ (см. Рис. 4). Матрица остатков показывает разности между исходной и воспроизведенной матрицами корреляций. Этот шаг позволяет оценить, насколько хорошо факторная модель воспроизводит исходные корреляции. Все элементы матрицы остатков по модулю не превышают 0.05, что подтверждает точность воспроизведения корреляций двумя факторами.

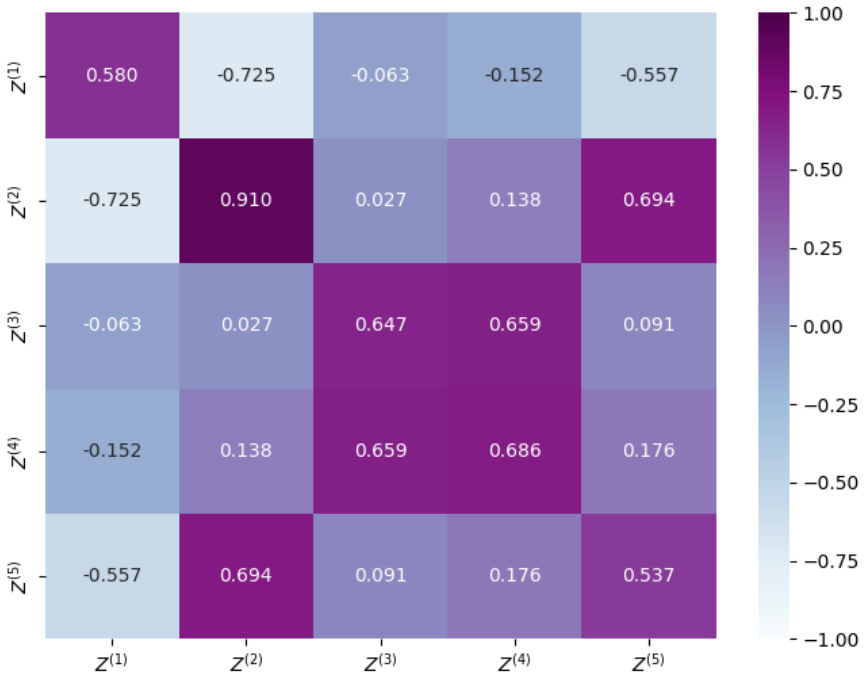

Рис. 3: Воспроизведенная матрица корреляций $\hat{\mathbf{R}}^* = \hat{\mathbf{A}}\hat{\mathbf{A}}^T$

2.5 Диаграмма факторных нагрузок

Для визуализации результатов факторного анализа была построена диаграмма факторных нагрузок (см. Рис. 5). На диаграмме каждая переменная представлена точкой, координаты которой соответствуют нагрузкам на первый и второй общие факторы. Этот шаг помогает наглядно представить, как переменные связаны с факторами. Например, переменная $Z^{(2)}$ (подъем) имеет высокую положительную нагрузку на первый фактор и отрицательную на второй, что указывает на сильную связь с первым фактором.

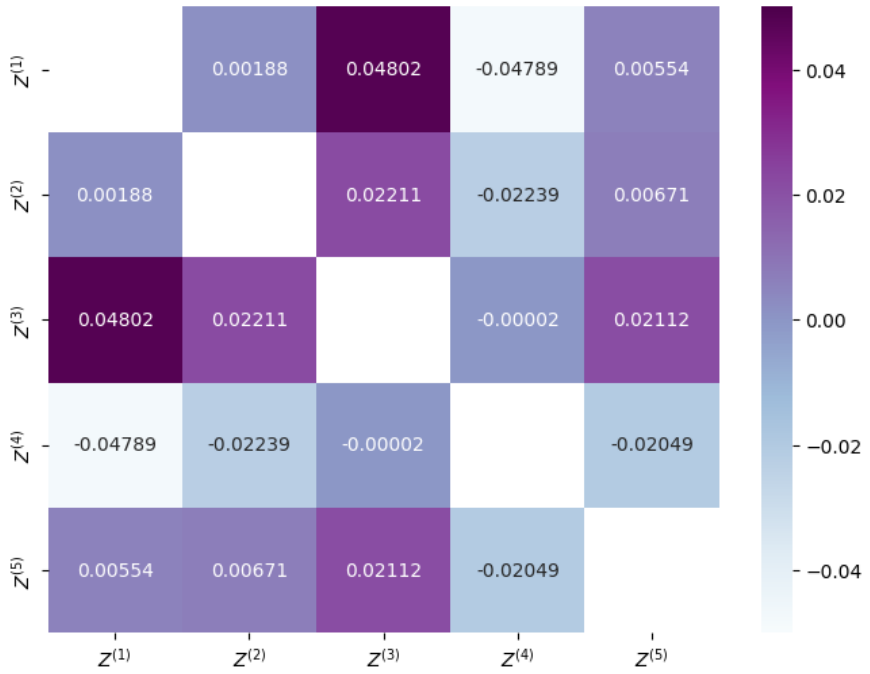

Рис. 4: Матрица остатков $\hat{\mathbf{R}} - \hat{\mathbf{R}}^*$

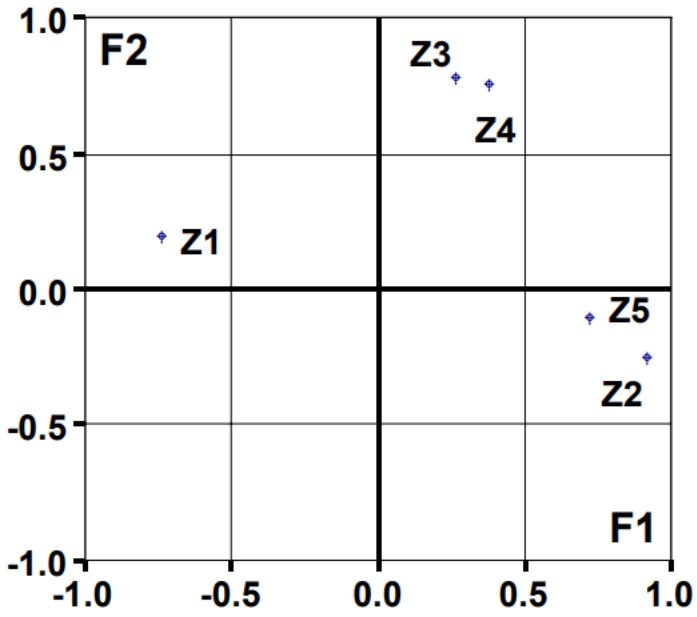

Рис. 5: Диаграмма факторных нагрузок

2.6 Интерпретация факторов

На основе матрицы факторных нагрузок и диаграммы можно интерпретировать выделенные факторы:

1. **Первый фактор** сильно положительно коррелирован с переменными $Z^{(2)}$ (подъем) и $Z^{(5)}$ (трасса) и отрицательно – с $Z^{(1)}$ (цена). Это указывает на то, что высокий уровень первого фактора связан с предпочтением технических характеристик курорта (скорость подъемников и протяженность трасс) и меньшим вниманием к цене.

2. **Второй фактор** сильно положительно коррелирован с переменными $Z^{(3)}$ (снег) и $Z^{(4)}$ (влажность), что свидетельствует о важности состояния снежного покрова для потребителей.

2.7 Вращение факторов

Для улучшения интерпретации факторов было выполнено ортогональное вращение методом Varimax. Вращение осуществляется с помощью ортогональной матрицы $\mathbf{Q}$, которая преобразует исходные факторы в новые, сохраняя их ортогональность. Этот шаг позволяет упростить интерпретацию факторов, увеличивая нагрузки на «свои» переменные и уменьшая на «чужие». Преобразование задается формулами:

$$\tilde{\mathbf{f}} = \mathbf{Q}^T \mathbf{f},$$
$$\tilde{\mathbf{A}} = \mathbf{A}\mathbf{Q}$$

где $\tilde{\mathbf{f}}$ – новые факторы, а $\tilde{\mathbf{A}}$ – новая матрица факторных нагрузок.

В результате вращения была получена новая матрица факторных нагрузок (см. Табл. 2) и соответствующая диаграмма (см. Рис. 6). После вращения нагрузки факторов на «свои» переменные увеличились, что упрощает интерпретацию.

Таблица 2: Новая матрица факторных нагрузок $\tilde{\mathbf{A}}$

	$\tilde{f}^{(1)}$	$\tilde{f}^{(2)}$	h_j^2
$Z^{(1)}$ (цена)	-.758	-.067	.580
$Z^{(2)}$ (подъем)	.954	.020	.910
$Z^{(3)}$ (снег)	.011	.804	.647
$Z^{(4)}$ (влажность)	.127	.818	.686
$Z^{(5)}$ (трасса)	.725	.103	.537
Вклады	2.028	1.331	3.359

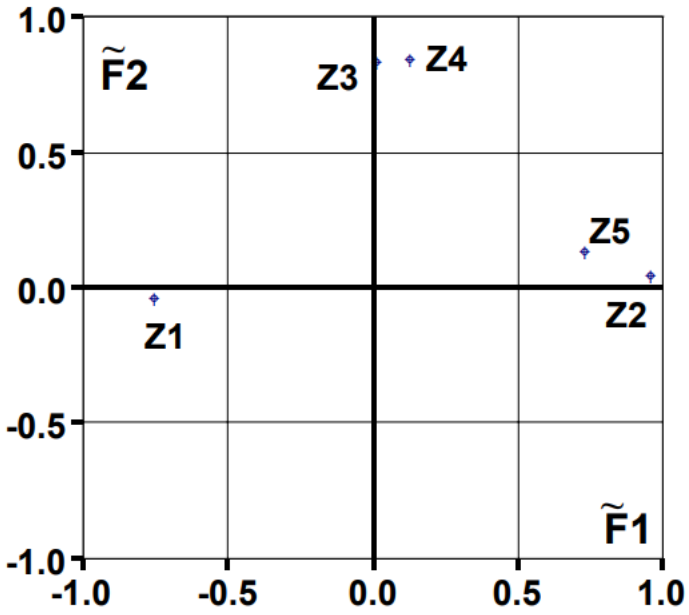

Рис. 6: Диаграмма факторных нагрузок после вращения

2.8 Оценка значений факторов

Для оценки значений общих факторов для каждого опрошенного используется регрессионный метод:

$$\mathbf{F} = \mathbf{Z}\hat{\mathbf{R}}^{-1}\hat{\mathbf{A}},$$

где $\mathbf{Z}$ – матрица стандартизованных исходных данных, $\mathbf{F}$ – матрица факторных значений, а $\hat{\mathbf{R}}^{-1}\hat{\mathbf{A}}$ – матрица расчета факторных значений (см. Табл. 3). Этот шаг позволяет получить количественные оценки факторов для каждого наблюдения, что может быть полезно для дальнейшего анализа, например, для кластеризации или регрессионного анализа.

Таблица 3: Матрица расчета факторных значений $\hat{\mathbf{R}}^{-1}\hat{\mathbf{A}}$

	$\tilde{f}^{(1)}$	$\tilde{f}^{(2)}$
$Z^{(1)}$ (цена)	-.122	-.003
$Z^{(2)}$ (подъем)	.791	-.183
$Z^{(3)}$ (снег)	-.067	.464
$Z^{(4)}$ (влажность)	.040	.517
$Z^{(5)}$ (трасса)	.105	.027

Значения факторов для каждого опрошенного вычисляются по фор-

мулам:

$$\tilde{F}_i^{(1)} = -0.122Z_i^{(1)} + 0.791Z_i^{(2)} - 0.067Z_i^{(3)} + 0.040Z_i^{(4)} + 0.105Z_i^{(5)},$$

$$\tilde{F}_i^{(2)} = -0.003Z_i^{(1)} - 0.183Z_i^{(2)} + 0.464Z_i^{(3)} + 0.517Z_i^{(4)} + 0.027Z_i^{(5)},$$

где $Z_i^{(j)}$ – стандартизованное значение переменной j для опрошенного i.

2.9 Заключение

Факторный анализ позволил выявить два основных фактора, влияющих на выбор горнолыжного курорта:

1. **Компромисс между ценой и качеством курорта** (первый фактор).
2. **Состояние снежного покрова** (второй фактор).

Вращение факторов методом Varimax улучшило интерпретацию модели, увеличив нагрузки факторов на соответствующие переменные. Полученные результаты могут быть использованы для разработки маркетинговых стратегий и улучшения предложений горнолыжных курортов. Каждый шаг факторного анализа был направлен на то, чтобы максимально точно выявить и интерпретировать скрытые факторы, влияющие на выбор курорта, что позволяет принимать обоснованные решения на основе данных.