Метод главных компонент

Метод главных компонент осуществляет переход от исходной совокупности признаков $\xi_1, \xi_2, \ldots, \xi_k$ к новой совокупности некоррелированных признаков $\eta_1, \eta_2, \ldots, \eta_\kappa$, каждый из которых является линейной комбинацией исходных признаков $\xi_1, \xi_2, \ldots, \xi_k$. При этом линейные комбинации выбираются таким образом, что среди всех возможных линейных нормированных комбинаций исходных признаков первая главная компонента η_1 обладает наибольшей дисперсией. Геометрически это выглядит как ориентация новой координатной оси η_1 вдоль направления наибольшей вытянутости эллипсоида рассеивания исследуемой выборки в пространстве признаков $\xi_1, \xi_2, \ldots, \xi_k$. Вторая главная компонента имеет наибольшую дисперсию среди всех оставшихся линейных преобразований, некоррелированных с первой главной компонентой. Она интерпретируется как направление наибольшей вытянутости эллипсоида рассеивания, перпендикулярное первой главной компоненте. Следующие главные компоненты определяются по аналогичной схеме.

В дальнейшем из полученных величин можно оставить только m $(m < k)$ наиболее значимых факторов, вносящих максимальный вклад в суммарную дисперсию и использовать эти величины как некие интегральные факторы, характеризующие всю совокупность признаков (если компоненты определяются последовательно, то надо определить m, при котором следует остановить дальнейший поиск) .

Пусть $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)$ – центрированная многомерная случайная величина с матрицей ковариаций A. Положим

$$\eta_i = \vec{\beta}^{(i)T}\vec{\xi} = \beta_1^{(i)}\xi_1 + \beta_2^{(i)}\xi_2 + \ldots + \beta_k^{(i)}\xi_k \ , \ i = \overline{1,m}, \ (m \le k) \qquad (1)$$

где $\vec{\beta}^{(i)}$ – векторы неизвестных коэффициентов преобразования. Будем называть величины η_i главными компонентами. Так как величины ξ_i центрированы, то $M(\eta_i) = 0, i = \overline{1,m}$, а дисперсия главных компонент определяется как:

$$D(\eta_i) = E(\eta_i \eta_i^T) = \vec{\beta}^{(i)T} E(\xi \xi^T) \vec{\beta}^{(i)} = \vec{\beta}^{(i)T} A \vec{\beta}^{(i)} \qquad (2)$$

На вектор коэффициентов преобразования $\vec{\beta}^{(i)}$ накладываем условиям нормировки:

$$\| \vec{\beta}^{(i)} \|^2 = \vec{\beta}^{(i)T}\vec{\beta}^{(i)} = 1. \qquad (3)$$

Задачу отыскания главных компонент можно свести к задаче на собственные значения и собственные векторы матрицы ковариаций A. Векторы коэффициентов преобразования $\vec{\beta}^{(i)}$, $i = \overline{1,k}$, удовлетворяющие условиям задачи поиска главных компонент, являются

собственными векторами матрицы ковариаций A, соответствующими собственным значениям $\lambda_1, \lambda_2, ..., \lambda_k$, упорядоченным по убыванию значений, причем $D(\eta_i) = \lambda_i$, а $\sum_{i=1}^{k} D(\xi_i) = \sum_{i=1}^{k} D(\eta_i) = \sum_{i=1}^{k} \lambda_i$.

Оценка главных компонент на основе выборочных данных строится на основе выборочной матрицы ковариаций. Оценки собственных значений, являющиеся собственными числами выборочной матрицы ковариаций, в случае нормального распределения генеральной совокупности являются оценками максимального правдоподобия. Если единицы измерения исходных признаков различаются или их значения сильно различаются, то лучше использовать при нахождении оценок главных компонент вместо выборочной матрицы ковариаций выборочную корреляционную матрицу (или нормированные исходные данные).

Пусть $X = \{(X_1^{(1)}, X_2^{(1)}, \ldots, X_k^{(1)}), (X_1^{(2)}, X_2^{(2)}, \ldots, X_k^{(2)}), \ldots, (X_1^{(n)}, X_2^{(n)}, \ldots, X_k^{(n)})\}$ - выборка объема n из k-мерной совокупности (матрица размеров $n \times k$).

1) Находим выборочную матрицу ковариаций $\overline{A} = \frac{1}{n} \hat{X}^T \hat{X}$, где $\hat{X}$ - центрированная выборочная матрица (из элементов каждого столбца матрицы X вычитаем среднее для этого столбца). Если исходные данные сильно различаются, следует использовать нормированную центрированную матрицу, то есть поделить элементы каждого столбца матрицы $\hat{X}$ на величину выборочного среднеквадратичного отклонения $\sqrt{\overline{D}}$ для данного столбца.
2) Находим собственные значения λ_i и собственные векторы $\vec{\beta}^{(i)}$ матрицы $\overline{A}$, упорядоченные по убыванию собственных значений. Каждый из собственных векторов должен удовлетворять условию нормировки: $\| \vec{\beta}^{(i)} \|^2 = \vec{\beta}^{(i)T} \vec{\beta}^{(i)} = 1$. Тем самым мы определяем новые компоненты η_i, связанные с исходными формулами (1). Определяем дисперсии каждой компоненты и суммы компонент.
3) Задаемся вопросом, какие из компонент можно отбросить, чтобы уменьшить размерность вектора признаков η. Заметим, что математически строгих критериев отбора не существует. Обычно используют один из следующих эвристических методов. Во первых, зная $sp\,\overline{A} = \sum_{i=1}^{k} \lambda_i$, можно выбрать те компоненты $\eta_i, i = \overline{1, m}$ из общего набора, которые бы объясняли не менее некоторой заданной доли q суммарной доли

дисперсии признаков. То есть, значимыми полагаем m первых компонент, для которых $\sum_{i=1}^{m}\lambda_i / Sp\,\overline{A} \geq q$. Обычно q полагают не менее 0,7.

Другим критерием отбора является критерий Кайзера, который предполагает использование для нахождения оценок собственных значений выборочной матрицы корреляций (можно применять и в случае использования матрицы ковариаций, следует лишь в этом случае нормировать (умножить) все собственные значения на величину $k / sp(\overline{A})$). Согласно данному критерию оставляют только те главные компоненты, дисперсия которых больше 1. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается.

Подход к оценке числа главных компонент по необходимой доле объяснённой дисперсии формально применим всегда, однако неявно он предполагает, что нет разделения на «сигнал» и «шум», и любая заранее заданная точность имеет смысл. Поэтому часто более продуктивна иная эвристика, основывающаяся на гипотезе о наличии «сигнала» (сравнительно малая размерность, относительно большая амплитуда) и «шума» (большая размерность, относительно малая амплитуда). С этой точки зрения метод главных компонент работает как фильтр: сигнал содержится, в основном, в проекции на первые главные компоненты, а в остальных компонентах пропорция шума намного выше. Вопрос, как оценить число необходимых главных компонент, если отношение «сигнал/шум» заранее неизвестно? Одним из наиболее популярных подходов является правило сломанной трости (англ. Broken stick model). Набор нормированных собственных чисел ($\lambda_i / Sp\overline{A}$, $i = \overline{1,k}$) сравнивается с распределением длин обломков трости единичной длины, сломанной в $k-1$ случайно выбранной точке (точки разлома выбираются независимо и равномерно распределенными по длине трости). Пусть l_i ($i = \overline{1,k}$) - длины полученных кусков трости, занумерованные в порядке убывания длины: $l_1 \geq l_2 \geq \ldots \geq l_k$. Тогда математическое ожидание l_i: $L_i = M(l_i) = \frac{1}{k}\sum_{j=i}^{k}\frac{1}{j}$. По правилу сломанной трости i-й собственный вектор (в порядке убывания собственных чисел λ_i) сохраняется в списке главных компонент, если $\frac{\lambda_1}{Sp\overline{A}} > L_1, \frac{\lambda_2}{Sp\overline{A}} > L_2, \ldots, \frac{\lambda_i}{Sp\overline{A}} > L_i$.

К подобным критериям относится также графический критерий каменистой осыпи Кэттелла. Критерий каменистой осыпи состоит в поиске точки, где убывание

собственных значений замедляется наиболее сильно. Справа от этой точки должна находится, по-видимому, только "факторная осыпь" ("осыпь" - это геологический термин для обломков, которые скапливаются в нижней части каменистого склона). Таким образом, число выделенных (значимых) факторов не должно превышать количество факторов слева от этой точки.

4) Записываем выражения для главных компонент (1) с найденными коэффициентами преобразования. Находим значения главных компонент (точнее оценки значений) для каждого из n наблюдений. Матрицу значений Y главных компонент можно получить как: $Y = \widehat{X}B$, где B - матрица столбцами которой являются векторы $\beta^{(i)}$, $i = \overline{1,k}$.

5) Так как $\mathrm{cov}(\xi_i, \eta_j) = \lambda_j \beta_i^{(j)}$, то для нормированных исходных данных относительный вклад признака ξ_i в дисперсию главной компоненты η_j характеризует квадрат величины $\beta_i^{(j)}$, то есть, квадрат соответствующей координаты вектора $\vec{\beta}^{(j)}$. Таким образом, для каждой компоненты η_j отбираем признаки, которым соответствуют наибольшие абсолютные значения координат вектора $\vec{\beta}^{(j)}$, и определяем их суммарную долю вклада в компоненту, суммируя соответствующие квадраты координат.

Использование МГК для устранения эффекта мультиколлинеарности

Поскольку метод главных компонент осуществляет переход от исходной совокупности признаков $\xi_1, \xi_2, \ldots, \xi_k$ к новой совокупности некоррелированных признаков $\eta_1, \eta_2, \ldots, \eta_\kappa$, каждый из которых является линейной комбинацией исходных признаков, МГК можно использовать для устранения эффекта мультиколлинеарности в регрессионных моделях. Достаточно провести регрессионный анализ результирующего показателя с факторами $\eta_1, \eta_2, \ldots, \eta_\kappa$ и отобрать значимые факторы. Поскольку факторы $\eta_1, \eta_2, \ldots, \eta_\kappa$ не коррелируют между собой, то оценки коэффициентов модели при одних и тех же факторах не будут различаться для разных моделей (с разным числом факторов). Хотя оценки дисперсий и, соответственно, уровни значимости одних и тех же коэффициентов в разных моделях будут различаться. Теоретически возможна ситуация, когда значимый коэффициент в модели с большим количеством факторов станет незначимым в модели с меньшим числом факторов и наоборот (что приведет к потере данного фактора в модели). Чтобы избежать потерь значимых факторов в этом случае, при переходе от модели с большим числом параметров к модели с меньшим числом параметров, следует отбирать факторы с запасом по

уровню значимости (то есть, например, при заданном уровне значимости 0,05 отбирать значимые факторы на уровне 0,1).

Естественно возникает вопрос - можно ли при построении регрессионной модели использовать изначально только значимые главные компоненты (значимые - в смысле метода МГК). Однозначного ответа нет. С одной стороны, МГК формирует новые компоненты по принципу максимума выделенной дисперсии, а для регрессионного анализа важны не дисперсии факторов, а их корреляции с результирующей переменной. С этой точки зрения при построении регрессионной модели надо рассматривать все главные компоненты. С другой стороны, если мы предполагаем, что информативными являются только значимые компоненты, то следует при построении регрессионной модели учитывать только их, включая остальные, мы будем искать зависимость результирующей переменной от шумовой составляющей.