

Лекция 4. Сокращение признакового пространства. Метод главных компонент

Во многих задачах обработки и анализа биомедицинских данных исходное число p рассматриваемых, т. е. измеряемых на исследуемых объектах, признаков довольно велико. Для визуализации картины, простоты интерпретации и упрощения расчетов очень часто необходимо представить каждое из наблюдений в виде набора чисел, состоящего из *существенно меньшего* (чем p) числа признаков. При этом оставшиеся признаки могут либо выбираться из числа исходных, либо определяться по какому-либо правилу по совокупности исходных признаков, например, как линейные комбинации последних.

Главные компоненты представляют собой новое множество исследуемых признаков

$$y^{(1)}, y^{(2)}, \dots, y^{(p)},$$

каждый из которых получен в результате некоторой линейной комбинации, непосредственно измеренных на объектах, исходных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$. Полученные в результате такого преобразования новые признаки $y^{(1)}, y^{(2)}, \dots, y^{(p)}$ могут обладать рядом удобных статистических свойств. Например, быть упорядоченными по степени рассеяния в изучаемой совокупности объектов: первый признак обладает наибольшей степенью рассеяния, т. е. наибольшей дисперсией, второй – меньшей и т.д.

Рассмотрим метод вычисления главных компонент. Будем предполагать, что исследуемые наблюдения $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ извлечены из некоторой p -мерной генеральной совокупности, определяемой соответствующей вероятностной мерой. Однако для приводимых здесь понятий из всех характеристик исследуемой генеральной совокупности существенное значение имеет лишь ковариационная матрица $\Sigma = [\sigma_{ij}]$, где

$$\sigma_{ij} = \mathbf{M}(x^{(i)} - a^{(i)})(x^{(j)} - a^{(j)}), \quad i, j = 1, 2, \dots, p. \quad (4.1)$$

Здесь $a^{(i)}$ компоненты вектора $\mathbf{a}$ средних значений признаков $x^{(i)}$. Будем считать, что вектор средних значений $\mathbf{a} = \mathbf{0}$, чего всегда можно добиться, рассматривая в качестве исходных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ не сами измерения, а их отклонения от своих выборочных средних значений.

Назовем *первой главной компонентой* исследуемой генеральной совокупности наблюдений такую нормированную линейную комбинацию p исходных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$,

$$y^{(1)} = l_{11}x^{(1)} + l_{12}x^{(2)} + \dots + l_{1p}x^{(p)} = \mathbf{L}_1^T \mathbf{X}$$

(здесь $\mathbf{L}_1^T = [l_{11} \ l_{12} \ \dots \ l_{1p}]$, причем $l_{11}^2 + l_{12}^2 + \dots + l_{1p}^2 = 1$), которая среди всех прочих нормированных линейных комбинаций $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ обладает *наибольшей* дисперсией.

Вообще, i -й *главной компонентой* исследуемой генеральной совокупности ($i = 2, 3, \dots, p$) будем называть такую нормированную линейную комбинацию p исходных признаков $x^{(1)}, x^{(2)}, \dots, x^{(p)}$

$$y^{(i)} = l_{i1}x^{(1)} + l_{i2}x^{(2)} + \dots + l_{ip}x^{(p)} = \mathbf{L}_i^T \mathbf{X},$$

которая среди всех прочих линейных нормированных ($l_{i1}^2 + l_{i2}^2 + \dots + l_{ip}^2 = 1$) комбинаций, некоррелированных со всеми предшествующими главными компонентами $y^{(1)}, \dots, y^{(i-1)}$ (т. е. $\text{cov}(y^{(i)}, y^{(j)}) = \mathbf{M}(y^{(i)} y^{(j)}) = 0$ для $j < i$), обладает наибольшей дисперсией.

Из определения следует, что, во-первых, главные компоненты $y^{(1)}, y^{(2)}, \dots, y^{(p)}$ занумерованы в порядке убывания их дисперсий, т. е. $\mathbf{D}y^{(1)} \geq \mathbf{D}y^{(2)} \geq \dots \geq \mathbf{D}y^{(p)}$, причем легко подсчитать, что, во-первых,

$$\mathbf{D}y^{(i)} = \mathbf{M}(\mathbf{L}_i^T \mathbf{X})^2 = \mathbf{M}(\mathbf{L}_i^T \mathbf{X} \mathbf{X}^T \mathbf{L}_i) = \mathbf{L}_i^T \sum \mathbf{L}_i \quad (4.2)$$

и, во-вторых, вектор $\mathbf{L}_i$ определяющий преобразование перехода от $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ к $y^{(i)}$ является так называемым i -м собственным вектором ковариационной матрицы $\sum$, т. е.

его компоненты $l_{i1}, l_{i2}, \dots, l_{ip}$, определяются как нормированное ($\sum_{j=1}^p l_{ij}^2 = 1$) решение системы уравнений

$$(\sum - \lambda_i \mathbf{I}) \mathbf{L}_i = 0, \quad (4.3)$$

где λ_i - i -й по величине корень уравнения

$$|\sum - \lambda \mathbf{I}| = 0. \quad (4.4)$$

Под $|\mathbf{M}|$ подразумевается определитель матрицы $\mathbf{M}$, под $\mathbf{I}$ — единичная матрица, а под λ - неизвестное число.

Из сопоставления выражений (4.2), (4.3) и (4.4) вытекает, что

$$\mathbf{D}y^{(i)} = \lambda_i.$$

Таким образом, ковариационная матрица $\sum_Y$ главных компонент $y^{(1)}, y^{(2)}, \dots, y^{(p)}$ будет иметь вид

$$\sum_Y = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix}.$$

Опираясь на то, что преобразование, с помощью которого осуществляется переход от исходных компонент $\mathbf{X}$ к главным компонентам $\mathbf{Y}$, ($\mathbf{Y} = \mathbf{L}\mathbf{X}$), является *ортгональным*, нетрудно выразить исходные переменные $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ через главные компоненты $x^{(i)} = l_{i1}y^{(1)} + l_{i2}y^{(2)} + \dots + l_{ip}y^{(p)}$ (в матричной записи $\mathbf{X} = \mathbf{L}^T \mathbf{Y}$), а также показать, что сумма дисперсий $(\mathbf{D}y^{(1)} + \mathbf{D}y^{(2)} + \dots + \mathbf{D}y^{(p)})$ главных компонент равны сумме дисперсий $(\mathbf{D}x^{(1)} + \mathbf{D}x^{(2)} + \dots + \mathbf{D}x^{(p)})$ исходных признаков.

Это дает основу для принятия решения о том, сколько последних главных компонент можно без особого ущерба изъять из рассмотрения, сократив тем самым размерность исследуемого пространства. Действительно, анализируя изменение относительной доли дисперсии

$$q(p') = \frac{\mathbf{D}y^{(1)} + \mathbf{D}y^{(2)} + \dots + \mathbf{D}y^{(p')}}{\mathbf{D}y^{(1)} + \mathbf{D}y^{(2)} + \dots + \mathbf{D}y^{(p)}} = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_{p'}}{\lambda_1 + \lambda_2 + \dots + \lambda_p} \quad (4.5)$$

($1 \leq p' \leq p$), вносимой первыми p' главными компонентами, в зависимости от числа этих компонент, можно разумно определить число компонент, которое целесообразно оставить в рассмотрении. Так, при изменении $q(p')$, изображенном на рис.4.1, очевидно целесообразно 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%.

Использование главных компонент оказывается наиболее естественным и полезным в том случае, когда все компоненты $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ исследуемого вектора $\mathbf{X}$ имеют общую физическую природу и соответственно измерены в одних и тех же единицах. Если же признаки $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ измеряются в различных единицах, то результаты исследования с помощью главных компонент будут существенно зависеть от размерности единиц измерения и выбора масштаба. Поэтому в этих случаях необходимо предварительно перейти к вспомогательным безразмерным признакам $x^{*(i)}$, например, с помощью нормирующего преобразования вида

$$x_v^{*(i)} = \frac{x_v^{(i)}}{\sqrt{\sigma_{ii}}}, \quad (i = 1, 2, \dots, p, \quad v = 1, 2, \dots, n),$$

где σ_{ii} соответствует (3.1), а затем строить главные компоненты относительно этих признаков $x^{*(1)}, \dots, x^{*(p)}$ и их ковариационной матрицы.

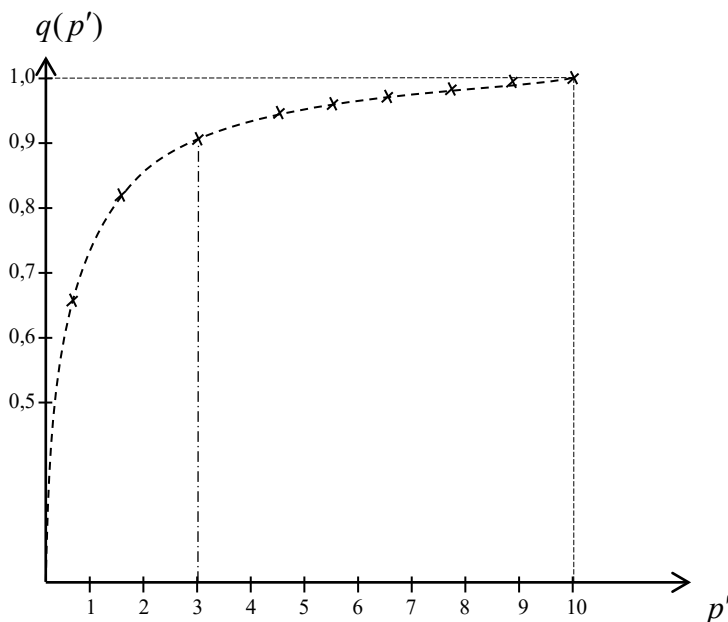


Рис. 4.1. Изменение относительной доли суммарной дисперсии исследуемых признаков, обусловленной первыми p' главными компонентами, в зависимости от p' (случай $p = 10$)

Пример. Исходные данные представляют собой результат описания множества реализаций ЭКГ (ритмограмм) набором признаков $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}$. Эти признаки получены в процессе вычисления и анализа параметров энтропии Колмогорова, которая

отражает степень сложности (хаотичности) сигналов. Выборка данных включает несколько классов ЭКГ: мерцательная аритмия (МА), нормальный ритм (НР), частая экстрасистолия (ЧЭ). Каждый из классов представлен 25 объектами.

Требуется осуществить переход от исходных признаков к главным компонентам, оценить долю дисперсии, которая обеспечивается первыми двумя главными компонентами $(y^{(1)}, y^{(2)})$ и отобразить объекты в пространстве новых признаков.

По данным измерений в соответствии с формулой (4.1) определена выборочная ковариационная матрица

$$\begin{bmatrix} 0,327 & 0,177 & -4,359 \cdot 10^{-4} & -0,061 & -0,085 \\ 0,177 & 0,112 & 0,018 & -0,023 & -0,052 \\ -4,359 \cdot 10^{-4} & 0,018 & 0,026 & 0,015 & -1,636 \cdot 10^{-3} \\ -0,061 & -0,023 & 0,015 & 0,022 & 0,016 \\ -0,085 & -0,052 & -1,636 \cdot 10^{-3} & 0,016 & 0,064 \end{bmatrix}.$$

Решая, в соответствии с (4.4), уравнение пятой степени относительно λ

$$\begin{vmatrix} 0,327 - \lambda & 0,177 & -4,359 \cdot 10^{-4} & -0,061 & -0,085 \\ 0,177 & 0,112 - \lambda & 0,018 & -0,023 & -0,052 \\ -4,359 \cdot 10^{-4} & 0,018 & 0,026 - \lambda & 0,015 & -1,636 \cdot 10^{-3} \\ -0,061 & -0,023 & 0,015 & 0,022 - \lambda & 0,016 \\ -0,085 & -0,052 & -1,636 \cdot 10^{-3} & 0,016 & 0,064 - \lambda \end{vmatrix} = 0,$$

находим значения $\lambda_1 = 0,463$; $\lambda_2 = 0,047$; $\lambda_3 = 0,038$, $\lambda_4 = 2,667 \cdot 10^{-3}$; $\lambda_5 = 8,135 \cdot 10^{-4}$.

Подставляя последовательно численные значения $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ в систему уравнений (4.3) и решая ее относительно неизвестных $\mathbf{L}_i = [l_{i1}, l_{i2}, l_{i3}, l_{i4}, l_{i5}]^T$ ($i = 1, 2, 3, 4, 5$), находим

$$\mathbf{L}_1 = \begin{bmatrix} -0,835 \\ -0,469 \\ -0,015 \\ 0,148 \\ 0,246 \end{bmatrix}, \mathbf{L}_2 = \begin{bmatrix} -0,228 \\ 0,431 \\ 0,724 \\ 0,453 \\ -0,182 \end{bmatrix}, \mathbf{L}_3 = \begin{bmatrix} 0,217 \\ 0,132 \\ 0,172 \\ 0,089 \\ 0,947 \end{bmatrix}, \mathbf{L}_4 = \begin{bmatrix} 0,421 \\ -0,575 \\ 0,018 \\ 0,696 \\ -0,085 \end{bmatrix}, \mathbf{L}_5 = \begin{bmatrix} 0,162 \\ -0,469 \\ 0,667 \\ -0,53 \\ -0,039 \end{bmatrix}.$$

В качестве главных компонент получаем

$$\begin{aligned} y^{(1)} &= -0,84x^{(1)} - 0,47x^{(2)} - 0,02x^{(3)} + 0,15x^{(4)} + 0,25x^{(5)}, \\ y^{(2)} &= -0,23x^{(1)} + 0,43x^{(2)} + 0,72x^{(3)} + 0,45x^{(4)} - 0,18x^{(5)}, \\ y^{(3)} &= 0,22x^{(1)} + 0,13x^{(2)} + 0,17x^{(3)} + 0,09x^{(4)} + 0,95x^{(5)}, \\ y^{(4)} &= 0,42x^{(1)} - 0,57x^{(2)} + 0,02x^{(3)} + 0,70x^{(4)} - 0,08x^{(5)}, \\ y^{(5)} &= 0,16x^{(1)} - 0,47x^{(2)} + 0,67x^{(3)} - 0,53x^{(4)} - 0,04x^{(5)}. \end{aligned}$$

Здесь под $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}$ подразумеваются отклонения параметров энтропии от своих средних значений.

Вычисление относительной доли суммарной дисперсии, обусловленной пятью главными компонентами, в соответствии с формулой (4.5) дает следующие результаты

$$q(1) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5} = 0,84,$$

$$q(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5} = 0,925,$$

$$q(3) = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5} = 0,994,$$

$$q(4) = \frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5} = 0,999,$$

$$q(5) = 1.$$

По результатам вычислений можно сделать вывод о том, что почти вся информация (92,5%) об изменчивости энтропийных характеристик, наблюдаемой на анализируемой выборке ЭКГ-данных (75 объектов), содержится в первых двух главных компонентах.

На рис. 4.2 приведено изображение в пространстве двух первых главных компонент множества ритмограмм для трех классов ЭКГ (МА, НР, ЧЭ).

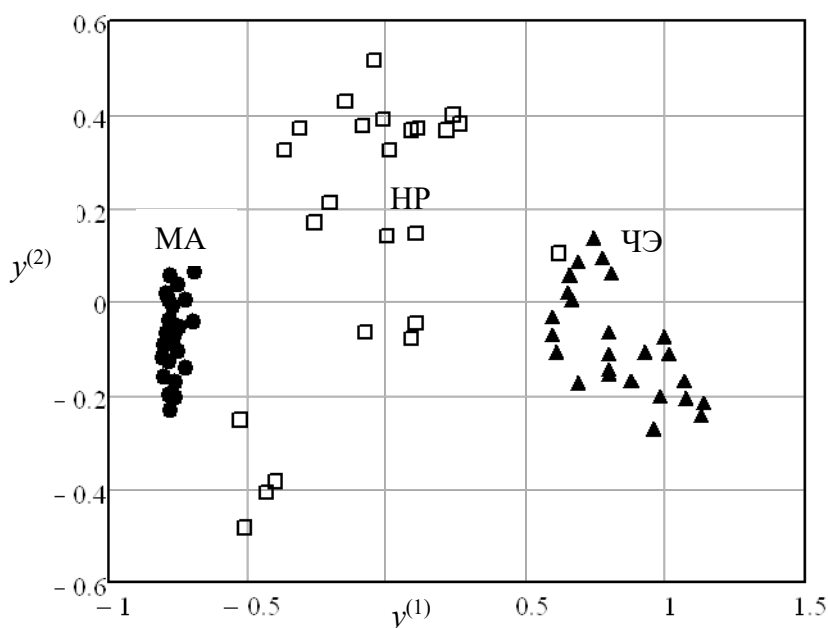


Рис. 4.2. Представление ритмограмм классов мерцательная аритмия (МА), нормальный ритм (НР) и частая экстрасистолия (ЧЭ) в пространстве двух первых главных компонент

Как видно и рисунка, описание объектов параметрами энтропии позволяет получить группировку объектов, соответствующую трем классам рассматриваемых ЭКГ сигналов. Наиболее плотная группировка объектов наблюдается в классе МА. Возможно линейное разделение классов, причем, используя энтропийные характеристики ритмограмм, можно эффективно решить задачу распознавания опасного нарушения (МА) на фоне альтернативных групп аритмий (ЧЭ и НР с выраженной дыхательной аритмией).