# О собственных числах в интеллектуальном анализе данных

**Двоенко С.Д.**
**Тульский государственный университет**

**O&ML: Семинар по оптимизации, машинному обучению, искусственному интеллекту (СПбГУ)**
**20.11.2025**

## ПЛАН

# 1. Свойства матрицы линейного преобразования

• Дано линейное преобразование $A\mathbf{x} = \mathbf{y}$, $\mathbf{x}, \mathbf{y} \in R^m$. Свойства преобразования определяются квадратной матрицей $A(m,m)$. В частности, существуют такие направления в $R^m$, что $A\mathbf{x} = \lambda\mathbf{x}$.

• Нормированный вектор $\mathbf{a} = \mathbf{x}/\|\mathbf{x}\|$ вдоль $\mathbf{x} \neq 0$ в положительном направлении называют собственным вектором, а коэффициент $\lambda -$ собственным числом. Собственные числа и соответствующие им собственные векторы при $\mathbf{x} \neq 0$ находятся из нетривиального решения векторно-матричного уравнения $(A - \lambda I) = 0$, где $I(m,m) = diag(1,...1)$ – единичная матрица. Раскрытие определителя $\det(A - \lambda I) = 0$ приводит к характеристическому полиному степени $m$ относительно $\lambda$: $(-1)^m \lambda^m + (-1)^{m-1} p_1 \lambda^{m-1} + ... + (-1)\lambda p_{m-1} + p_m = 0$.

• Уравнение имеет $m$ корней $\lambda_i$, $i = 1,... m$, которые в общем случае могут быть различными положительными и отрицательными, а также кратными и комплексными для квадратной матрицы $A$ общего вида.

• Нас интересует частный случай, когда, с одной стороны, матрица $A$ состоит из действительных чисел, является симметричной, имеет максимальные и неотрицательные элементы на главной диагонали. С другой стороны, нас интересует случай, когда все ее собственные числа неотрицательны $\lambda_i \geq 0,\ i = 1,... m$. Ненулевые собственные числа упорядочим по убыванию и проиндексируем: $\lambda_{\max} = \lambda_1 > ... > \lambda_m = \lambda_{\min} > 0$.

# 1. Свойства матрицы линейного преобразования

• При выполнении этих условий одновременно введем для матрицы $A(m,m)$ обозначение $S(m,m)$, рассматривая ее как матрицу скалярных произведений $s_{ij} = (\omega_i \circ \omega_j)$, построенную для некоторого множества элементов $\Omega = \{\omega_1, \ldots \omega_m\}$ в некотором гипотетическом многомерном метрическом (евклидовом) пространстве. Множество $\Omega$ часто удобно называть *обучающим*.

• Для ненормированной матрицы $S(m,m)$ должны быть выполнены строгие неравенства $|s_{ij}| < \sqrt{s_{ii} s_{jj}}$, если в решении характеристического полинома отсутствуют нулевые собственные числа. Преобразование $s'_{ij} = s_{ij} / \sqrt{s_{ii} s_{jj}}$ дает нормированную матрицу $S(m,m)$ полного ранга $rank\ S = m$. Для нестрогих неравенств $|s_{ij}| \le \sqrt{s_{ii} s_{jj}}$ следует рассматривать полуметрики, что, в частности, будет означать неполный ранг матрицы *S*.

*Случай частного решения* $\lambda_{\max} = \lambda_1 > \ldots > \lambda_m = \lambda_{\min} > 0$ *характеристического полинома особенно востребован в области анализа данных. Такое решение означает, что элементы множества* $\Omega$ *корректно (без нарушений) погружены в многомерное метрическое пространство, которое оказывается аналогичным по свойствам обычному трехмерному пространству, в котором мы живем.*

# 1. Свойства матрицы линейного преобразования

*Замечание 1*. Задача построения оптимальной разделяющей гиперплоскости (SVM – задача обучения Вапника-Червоненкиса) в одной из своих формулировок сводится к решению задачи квадратичного программирования в двойственной постановке, где решение опирается на матрицу скалярных произведений элементов обучающей выборки. Поэтому оказывается, что для поиска решающего правила исходные векторы наблюдений $\mathbf{x} = \mathbf{x}(\omega),\ \omega \in \Omega$ уже не нужны.

*Задачи, приводящие к матрицам $S(m,m)$, имеющим смысл скалярных произведений, потенциально оказываются задачами разделения множества $\Omega$ на подмножества (кластеризация, построение оптимальной гиперплоскости и т.п.).*

*Замечание 2*. В современном анализе данных часто изучаются сложные объекты (структуры). Для них может оказаться затруднительным сформировать подходящий набор количественных характеристик (признаков, атрибутов и т.п.). Может оказаться, что проще выполнить непосредственное сравнение таких объектов попарно, оценив количественно их *различие* или *сходство*. Такие оценки обычно получают специально разработанными алгоритмами парного сравнения. Поэтому результат стремятся получить в виде матрицы $S(m,m)$ неотрицательных парных близостей или матрицы $D(m,m)$ неотрицательных различий, минуя представление векторами $\mathbf{x} = \mathbf{x}(\omega),\ \omega \in \Omega$.

*Естественно предположить, что на практике мы будем сталкиваться с т.н. метрическими нарушениями. Выясним, в чем они будут заключаться.*

# 2. Метрические нарушения

• Рассмотрим матрицу $S(m,m)$ как результат скалярных произведений $s_{ij} = (\mathbf{x}_i \circ \mathbf{x}_j)$ векторов $\mathbf{x} \in R^m$, где $\mathbf{x} = \mathbf{x}(\omega),\ \omega \in \Omega$. Нормированная матрица *S* соответствует расположению концов этих векторов на гиперсфере единичного радиуса в $R^m$. Ненормированная матрица *S* сохраняет конфигурацию векторов различной длины в $R^m$. В частности, это означает, что их расстояния до начала координат *различны*.

• Выражение $\mathbf{x} = \mathbf{x}(\omega),\ \omega \in \Omega$ означает погружение множества $\Omega$ в метрическое пространство характеристик (атрибутов, признаков) элементов множества. Обычно это происходит в результате измерений.

• Также существуют аналитические методы, когда атрибуты вычисляются на основе подсчетов или специальными алгоритмами. В задачах обработки текстов применяется т.н. *эмбеддинг*, когда словам или текстам сопоставляются соответствующие векторы чисел.

• Рассмотрим матрицу евклидовых расстояний $D(m,m)$ с элементами $d_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|$ между теми же векторами $\mathbf{x} \in R^m$. Связь между скалярными произведениями и расстояниями определяется теоремой косинусов. Если были заданы только скалярные произведения $S(m,m)$, то евклидовы расстояния $D(m,m)$ вычисляются как $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$.

# 2. Метрические нарушения

• Если были заданы только расстояния $D(m,m)$, то предварительно нужно определить начало координат (н.к.), относительно которого будут взяты скалярные произведения. Представим н.к. как новый элемент $\omega_0$ в центре множества и определим его расстояния до всех остальных элементов (формула Торгерсона):

$$d_{0i}^2 = \frac{1}{m}\sum_{p=1}^{m} d_{ip}^2 - \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2 .$$

• Второе слагаемое определяет разброс элементов множества $\Omega$ относительно $\omega_0$ и является хорошо известной дисперсией: $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} d_{0i}^2 = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{m}\sum_{p=1}^{m} d_{ip}^2 - \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2\right) = \frac{1}{2m^2}\sum_{p=1}^{m}\sum_{q=1}^{m} d_{pq}^2$ .

• Тройка $<\omega_0, \omega_i, \omega_j>$ элементов представляет треугольник со сторонами $d_{0i}, d_{0j}$ и $d_{ij}$. По теореме косинусов $d_{ij}^2 = d_{0i}^2 + d_{0j}^2 - 2d_{0i}d_{0j}s_{ij}$. Тогда нормированные скалярные произведения $S(m,m)$ относительно н.к. $\omega_0$ вычисляются как $s_{ij} = \frac{1}{2d_{0i}d_{0j}}\left(d_{0i}^2 + d_{0j}^2 - d_{ij}^2\right)$, где $s_{ii} = 1$. Ненормированные скалярные произведения $s_{ij} = \left(d_{0i}^2 + d_{0j}^2 - d_{ij}^2\right)/2$ сохраняют конфигурацию множества элементов в пространстве, т.к. $s_{ii} = d_{0i}^2$.

# 2. Метрические нарушения

• Разнообразные нарушения в структуре матрицы скалярных произведений приводят к появлению отрицательных собственных чисел (и нулевых, т.к. полуметрики не рассматриваем).

• Заметим, что нарушение неравенства треугольника обычно довольно трудно осуществить. Нарушить теорему косинусов проще, т.к. это более жесткое условие.

• Нарушение теоремы косинусов хотя бы на одной тройке элементов приводит к появлению хотя бы одного отрицательного собственного числа для соответствующей подматрицы скалярных произведений и, следовательно, для всей матрицы.

*Тем не менее, даже при отсутствии нарушений теоремы косинусов на всех тройках элементов множества, нарушения могут возникать на конфигурациях, содержащих более трех элементов.*

*Утверждение.* Метрическое нарушение нарушает конфигурацию гиперсфер в специальном координатном пространстве для соответствующего множества элементов.

# 2. Метрические нарушения

*Доказательство.* Просмотрим множество $\Omega = \{\omega_1, \dots \omega_m\}$. Пусть $k$ элементов просмотрено. На шаге $k = 1, \dots m$ они представлены главным минором $S(k,k)$ матрицы $S(m,m)$, где $S(1,1) = s_{11} = 1$, $S(2,2) = \begin{pmatrix} 1 & s_{12} \\ s_{21} & 1 \end{pmatrix}$ и т.д. Обозначим значения главных миноров $S(k,k)$ как $S_k = \det S(k,k)$, где $S_1 = 1$, $S_2 = 1 - s_{12}^2$ и т.д. Доказательство заключается в определении гиперсфер в координатном пространстве.

Концы векторов в $m$–мерном пространстве сосредоточены на гиперсфере единичного радиуса. Будем добавлять каждый новый элемент в виде вектора в таком пространстве. Направим первый вектор вдоль первой оси этой гиперсферы и представим координатами $(u_1, \dots u_m) = (1, 0, \dots 0)$.

По теореме косинусов положение второго вектора относительно первого определяется скалярным произведением $s_{12}$, где расстояние между их концами $d_{12} = \sqrt{2 - 2s_{12}}$. С одной стороны, его конец расположен на *первой* $m$–мерной гиперсфере единичного радиуса. С другой стороны, возможные положения конца второго вектора относительно конца первого вектора также определяют $m$–мерную гиперсферу радиуса $d_{12}$. Положение конца второго вектора удовлетворяет условиям

$$\begin{cases} u_1^2 + \dots + u_m^2 = 1 \\ (u_1 - 1)^2 + \dots + u_m^2 = 2 - 2s_{12} \end{cases}.$$

Вычтем второе уравнение из первого и получим $u_1 = s_{12}$. После подстановки получим уравнение *второй* гиперсферы $u_2^2 + \dots + u_m^2 = 1 - s_{12}^2 = S_2 / S_1$. Возможные позиции конца второго вектора расположены

на *второй* $(m-1)-$мерной гиперсфере с центром в начале системы координат $(u_2,...u_m)$ и радиусом $\sqrt{S_2/S_1}$. Так как $u_1=s_{12}$, то в *m*–мерной системе координат $(u_1,...u_m)$ зафиксируем положение второго вектора и представим координатами $(s_{12},\sqrt{S_2/S_1},0,...0)$.

Положение *третьего* вектора относительно первого вектора определяется скалярным произведением $s_{13}$, где расстояние между их концами $d_{13}=\sqrt{2-2s_{13}}$. Как и ранее, это дает значение первой координаты $u_1=s_{13}$. С другой стороны, возможные положения конца третьего вектора относительно конца второго вектора определяют уже *третью* $(m-2)-$мерную гиперсферу радиуса $d_{23}=\sqrt{2-2s_{23}}$. Положение конца третьего вектора удовлетворяет условиям

$$\begin{cases} u_1=s_{13} \\ u_2^2+...+u_m^2=1-s_{13}^2 \\ (u_1-s_{12})^2+\left(u_2-\sqrt{1-s_{12}^2}\right)^2+u_3^2+...+u_m^2=2-2s_{23}\,. \end{cases}$$

Вычитая третье уравнение из второго с учетом первого равенства, получим

$$u_2=(s_{23}-s_{12}s_{13})/\sqrt{1-s_{12}^2}=(S_3)_3^2/\sqrt{S_1S_2}\,,$$

где $(S_k)_j^i=\det(S(k,k))_j^i$ представляет значение дополнительного минора $(S(k,k))_j^i$ главного минора $S(k,k)$ после вычеркивания *i*–й строки и *j*–го столбца. После подстановки во второе равенство получим

# 2. Метрические нарушения

*третью* гиперсферу: $u_3^2+\ldots+u_m^2=(1+2s_{12}s_{13}s_{23}-s_{12}^2-s_{13}^2-s_{23}^2)/(1-s_{12}^2)=S_3/S_2$.

Следовательно, все возможные позиции конца третьего вектора расположены на *третьей* $(m-2)-$мерной гиперсфере с центром в начале системы координат $(u_3,\ldots u_m)$ и радиусом $\sqrt{S_3/S_2}$. Так как $u_1=s_{13}$ и $u_2=(S_3)_3^2/\sqrt{S_1S_2}$, то в *m*–мерной системе координат $(u_1,\ldots u_m)$ зафиксируем положение третьего вектора и представим координатами $(s_{13},\,(S_3)_3^2/\sqrt{S_1S_2},\,\sqrt{S_3/S_2},0,\ldots 0)$.

Продолжая для *четвертого* вектора и далее, получим компоненты $u_{k-1}=(S_k)_k^{k-1}/\sqrt{S_{k-2}S_{k-1}}$ и $(m-k+1)$–мерные гиперсферы радиусов $\sqrt{S_k/S_{k-1}}$ в виде уравнений $u_k^2+\ldots+u_m^2=S_k/S_{k-1}$. Тогда очередной *k*–й вектор в *m*–мерной системе координат представлен координатами

$$\begin{cases} u_1=s_{1k} \\ u_t=(S_k)_{t+1}^t/\sqrt{S_{t-1}S_t},\ t=2,\ldots k-1 \\ u_k=\sqrt{S_k/S_{k-1}} \\ u_{k+1}=\ldots=u_m=0 \end{cases},$$

где $S_0=1$, а координаты $u_t$ с индексами вне диапазона $1\le i\le m$ не существуют. Для нормированной матрицы $S(m,m)$ последовательность ее главных миноров определяет убывающие значения их детерминантов: $(S_1=1)>(S_2=1-s_{12}^2)>\ldots>S_m=\det S(m,m)$.

# 2. Метрические нарушения

*Заключение.* Нарушение возникает, когда текущий главный минор оказывается отрицательным. В этом случае квадрат радиуса соответствующей гиперсферы оказывается отрицательным. Тогда сам радиус оказывается комплексной величиной. Конец соответствующего вектора не располагается на соответствующей гиперсфере. ■

*Следствие 1.* Пусть множество $\Omega = \{\omega_1, \ldots \omega_m\}$ представлено матрицей $S(m,m)$ скалярных произведений, и теорема косинусов не нарушена ни на одной тройке элементов из $\Omega$. Пусть найдется элемент, представленный вектором, чей конец не расположен на соответствующей гиперсфере с радиусом, определенным относительно всех предыдущих векторов. Тогда матрица $S(m,m)$ имеет хотя бы одно отрицательное собственное число.

*Следствие 2.* Значения главных миноров положительно определенной нормированной матрицы скалярных произведений $S(m,m)$ убывают, начиная с единицы, оставаясь положительными. Если на множестве $\Omega = \{\omega_1, \ldots \omega_m\}$ возникают метрические нарушения, то значения главных миноров убывают по абсолютной величине. Число смен знака их детерминантов определяется числом отрицательных собственных чисел, согласно закону инерции квадратичных форм Сильвестра.

# 2. Метрические нарушения

*Следствие 3*. Пусть после добавления к уже просмотренным элементам очередного $k$–го элемента из множества $\Omega = \{\omega_1, \ldots \omega_m\}$ возникло метрическое нарушение. Тогда в последовательности главных миноров текущий главный минор $S(k,k)$ изменит свой знак на противоположный по сравнению со знаком предыдущего главного минора $S(k-1,k-1)$.

*Следствие 4*. Пусть метрические нарушения на множестве $\Omega$ устраняются поочередно по мере их возникновения. Тогда при возникновении метрического нарушения текущий главный минор $S(k,k)$ становится отрицательным. Для устранения данного метрического нарушения нужно восстановить положительную определенность данного главного минора. Для этого необходимо скорректировать парные сравнения текущего $k$–го элемента из множества $\Omega$, который представлен последней строкой и последним столбцом данного главного минора.

*Следствие 5*. Пусть эмпирические парные сравнения неотрицательной похожести элементов из $\Omega$ представлены положительно определенной матрицей $S(m,m)$. Тогда данная матрица рассматривается как матрица скалярных произведений векторов, представляющих элементы множества $\Omega$ в одном и том же квадранте метрического пространства размерности не выше $m$.

*Следствие 6*. Пусть эмпирические парные сравнения различий элементов из $\Omega$ представлены матрицей $D(m,m)$. Если ответствующая ей по теореме косинусов матрица скалярных произведений или неотрицательных близостей $S(m,m)$ положительно определена, то данная матрица $D(m,m)$ рассматривается как матрица расстояний.

# 3. Оптимальная коррекция парных сравнений

• В последовательности главных миноров $S(k,k), k=1,...m$ нормированной матрицы $S(m,m)$ их значения $S_k = \det S(k,k)$ определены как $S_1 = 1$, $S_2 = 1 - s_{12}^2$ и т.д. Согласно доказанному выше утверждению, последовательность главных миноров определяет последовательность убывающих значений их детерминантов $(S_1 = 1) > (S_2 = 1 - s_{12}^2) > ... > S_m = \det S(m,m)$.

• Чем медленнее убывают значения детерминантов в этой последовательности, тем меньше возможностей, что следующий главный минор окажется нулевым или отрицательным, т.е. возникнет метрическое нарушение.

• В случае метрического нарушения нужно изменить отрицательное значение детерминанта текущего главного минора на положительное, обеспечив наименьшее отклонение от значения детерминанта предыдущего главного минора.

• Представим разложение главного минора $S(k,k)$ по элементам *k*–ой строки

$$S_k = \sum_{p=1}^{k} (-1)^{k+p} s_{kp} (S_k)_p^k = \sum_{p=1}^{k-1} (-1)^{k+p} s_{kp} (S_k)_p^k + S_{k-1},$$

где $(S_k)_p^k$ – это значение дополнительного минора $(S(k,k))_p^k$ после исключения *k*–ой строки и *p*–го столбца, и $S_{k-1} = (-1)^{k+k} s_{kk} (S_k)_k^k = (S_k)_k^k$.

# 3. Оптимальная коррекция парных сравнений

• Далее представим разложения миноров $(S(k,k))_p^k$ по элементам *k*–го столбца, сохранив исходную индексацию минора $S(k,k)$:

$$S_k = S_{k-1} + \sum_{p=1}^{k-1}(-1)^{k+p} s_{kp}\left(\sum_{q=1}^{k-1}(-1)^{(q+k)-1} s_{qk}((S_k)_p^k)_k^q\right) =$$

$$= S_{k-1} + \sum_{p=1}^{k-1}\sum_{q=1}^{k-1}(-1)^{(2k-1)+p+q} s_{kp}s_{qk}(S_{k-1})_p^q = S_{k-1} - \sum_{p=1}^{k-1}\sum_{q=1}^{k-1}(-1)^{p+q} s_{kp}s_{qk}(S_{k-1})_p^q .$$

• Рассмотрим обратную матрицу $R(k-1,k-1) = S(k-1,k-1)^{-1}$ с элементами $r_{pq} = (-1)^{p+q}(S_{k-1})_p^q / S_{k-1}$. Декомпозиция минора $S(k,k)$ по элементам *k*–й строки и *k*–го столбца, рассмотренная выше, теперь имеет вид

$$S_k = S_{k-1} - \sum_{p=1}^{k-1}\sum_{q=1}^{k-1} s_{kp}s_{qk}r_{pq}S_{k-1} = S_{k-1}\left(1 - \sum_{p=1}^{k-1}\sum_{q=1}^{k-1} s_{kp}s_{qk}r_{pq}\right).$$

• Если $S_k < 0$, то найдем новое значение $0 \le c \le S_{k-1}$. Обозначим новые значения изменяемых элементов как неизвестные переменные $x_p = s_{kp} = s_{pk}$, $p = 1,\ldots k-1$ в скорректированном миноре $S(k,k)$.

# 3. Оптимальная коррекция парных сравнений

• Его значение удовлетворяет условию $c = S_{k-1}\left(1 - \sum_{p=1}^{k-1}\sum_{q=1}^{k-1} s_{kp}s_{qk}r_{pq}\right) = S_{k-1}(1-C)$, где $C = 1 - c/S_{k-1}$ для предварительно заданной величины $0 \le c \le S_{k-1}$ детерминанта $S_k$. Так как $S_k \le S_{k-1}$, то удобно определить $c = \tau S_{k-1}$ как часть величины $S_{k-1}$, где $0 \le \tau \le 1$.

• Рассмотрим *задачу оптимальной коррекции* с возможностью корректировать не все парные сравнения элемента $\omega_k \in \Omega$, вызвавшего метрическое нарушение. Обозначим $P = \{1,...k-1\}$ множество индексов всех элементов. Обозначим $I \subseteq P$ подмножество индексов модифицированных элементов из $P$. Тогда индексы неизмененных элементов образуют подмножество $P \backslash I$.

• Задача условной оптимизации: $\sum_{p\in I}(s_{pk} - x_p)^2 \to \min$, $\sum_{p\in P}\sum_{q\in P} s_{kp}s_{qk}r_{pq} = C$.

• Решение методом множителей Лагранжа (обозначим их $\xi$) дает систему уравнений, где число уравнений определено индексами $p \in I$ модифицированных элементов

$$\begin{cases} \xi\sum_{i\in I} x_i r_{ip} + \sum_{i\in P\backslash I} s_{ki}r_{ip} = s_{kp} - x_p,\ p \in I \\ \sum_{i\in I}\sum_{j\in I} x_i x_j r_{ij} + \sum_{i\in I}\sum_{j\in P\backslash I} x_i s_{jk} r_{ij} + \sum_{i\in P\backslash I}\sum_{j\in I} s_{ki} x_j r_{ij} + \sum_{i\in P\backslash I}\sum_{j\in P\backslash I} s_{ki}s_{jk}r_{ij} = C \end{cases}.$$

# 3. Оптимальная коррекция парных сравнений

*Пример 1*. Дана матрица $S=\begin{pmatrix}1 & 0.5 & 0.5\\ & 1 & -0.9\\ & & 1\end{pmatrix}$. Детерминанты ее главных миноров: $S_1=1$, $S_2=\det\begin{pmatrix}1 & 0.5\\ & 1\end{pmatrix}=0.75$, $S_3=\det S=-0.76$ и собственные числа: 1.9, 1.3882, –0.2882.

Скорректируем последние строку и столбец полностью так, чтобы последний минор имел величину $c=0.1$. Определим обратную матрицу

$$R=\begin{pmatrix}1 & 0.5\\ 0.5 & 1\end{pmatrix}^{-1}=\frac{1}{0.75}\begin{pmatrix}1 & -0.5\\ -0.5 & 1\end{pmatrix}=\begin{pmatrix}4/3 & -2/3\\ -2/3 & 4/3\end{pmatrix}$$

и вычислим $C=1-0.1/0.75=13/15$. Получим систему уравнений:

$$\begin{cases}\xi(x_1r_{11}+x_2r_{21})=s_{31}-x_1\\ \xi(x_1r_{12}+x_2r_{22})=s_{32}-x_2\\ x_1^2r_{11}+x_1x_2r_{12}+x_2x_1r_{21}+x_2^2r_{22}=C\end{cases}=\begin{cases}\xi(\frac{4}{3}x_1-\frac{2}{3}x_2)=0.5-x_1\\ \xi(-\frac{2}{3}x_1+\frac{4}{3}x_2)=-0.9-x_2\\ \frac{4}{3}x_1^2-\frac{4}{3}x_1x_2+\frac{4}{3}x_2^2=\frac{13}{15}\end{cases}.$$

Решение: $x_1=0.285487$, $x_2=-0.624637$, $\xi=0.269126$. После подстановки в новой матрице $\tilde{S}=\begin{pmatrix}1 & 0.5 & 0.285487\\ & 1 & -0.624637\\ & & 1\end{pmatrix}$ все детерминанты: $\tilde{S}_1=1$, $\tilde{S}_2=0.75$, $\tilde{S}_3=0.1$ и собственные числа: 1.6769, 1.2763, 0.0467 положительны.

# 3. Оптимальная коррекция парных сравнений

*Пример 2*. Скорректируем только первый элемент последних строки и столбца так, чтобы последний минор снова имел величину $c = 0.1$. Получим систему уравнений:

$$\begin{cases} \xi x_1 r_{11} + s_{32} r_{21} = s_{31} - x_1 \\ x_1^2 r_{11} + x_1 s_{23} r_{12} + s_{32} x_1 r_{21} + s_{32}^2 r_{22} = C \end{cases} = \begin{cases} \frac{4}{3}\xi x_1 - \frac{2}{3}(-0.9) = 0.5 - x_1 \\ \frac{4}{3}x_1^2 - \frac{4}{3}x_1(-0.9) + \frac{4}{3}(-0.9)^2 = \frac{13}{15} \end{cases}.$$

Решение: $x_1 = -0.243845$, $\xi = -0.442427$. После подстановки в новой матрице $\tilde{S} = \begin{pmatrix} 1 & 0.5 & -0.243845 \\ & 1 & -0.9 \\ & & 1 \end{pmatrix}$

снова все детерминанты: $\tilde{S}_1 = 1$, $\tilde{S}_2 = 0.75$, $\tilde{S}_3 = 0.1$ и собственные числа: 2.145, 0.7964, 0.0585 положительны.

*Пример 3*. Скорректируем только второй элемент последних строки и столбца так, чтобы, как и ранее, последний минор имел величину $c = 0.1$. Получим систему уравнений:

$$\begin{cases} \xi x_2 r_{22} + s_{31} r_{12} = s_{32} - x_2 \\ x_2^2 r_{22} + x_2 s_{13} r_{21} + s_{31} x_2 r_{12} + s_{31}^2 r_{11} = C \end{cases} = \begin{cases} \frac{4}{3}\xi x_2 - \frac{2}{3}0.5 = -0.9 - x_2 \\ \frac{4}{3}x_2^2 - \frac{4}{3}0.5 x_2 + \frac{4}{3}0.5^2 = \frac{13}{15} \end{cases}.$$

Решение: $x_2 = -0.430074$, $\xi = 0.238203$. После подстановки в новой матрице $\tilde{S} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ & 1 & -0.430074 \\ & & 1 \end{pmatrix}$

снова все детерминанты: $\tilde{S}_1 = 1$, $\tilde{S}_2 = 0.75$, $\tilde{S}_3 = 0.1$ и собственные числа: 1.524, 1.4301, 0.0459 положительны.

# 3. Оптимальная коррекция парных сравнений

- Численные примеры показывают, что корректировка сразу всех парных сравнений (корректировка вектором) дает значения, которые более похожи на исходные значения.

- Коррекция отдельных парных сравнений дает значения, которые могут сильно отличаться от исходных значений.

- В общем случае вообще не идет речи о восстановлении «исходных» значений, так как восстанавливаются только такие значения, которые обеспечивают корректность конфигурации.

# 4. Локализация отрицательных собственных чисел

• Пусть в конфигурации множества $\Omega = \{\omega_1, \ldots \omega_m\}$ возникли нарушения. Заметим, что элементы этого множества можно просмотреть и в другом порядке. Может оказаться так, что в другой последовательности главных миноров потребуется скорректировать отрицательные значения других главных миноров. В итоге, может оказаться, что потребуется скорректировать меньшее их число и внести меньше изменений в исходную матрицу парных сравнений.

• Тогда среди всех возможных последовательностей главных миноров найдется такая последовательность, что обеспечит минимум возможных коррекций, как по числу миноров, так и по величине самих изменений. Задача поиска такой последовательности представляет собой комбинаторную проблему в общем случае.

• Известно, что порядок просмотра элементов множества $\Omega$ не влияет на свойства матрицы парных сравнений $S(m,m)$. В частности, одновременная перестановка строк и столбцов квадратной матрицы не изменяет ее собственных чисел.

• В соответствии с *законом инерции* Сильвестра число смен знаков значений детерминантов при просмотре последовательности главных миноров совпадает с числом отрицательных собственных чисел матрицы $S(m,m)$.

*Следовательно, число изменений знаков детерминантов в последовательности главных миноров определяет число элементов множества $\Omega$, которые вносят метрические нарушения в конфигурацию.*

# 4. Локализация отрицательных собственных чисел

• Заметим, что закон инерции Сильвестра определяет число изменений знаков детерминантов главных миноров, но *не места их расположения* в последовательности.

• Определим порядок просмотра элементов множества $\Omega = \{\omega_1, \ldots \omega_m\}$ так, что в последовательности главных миноров $S(k,k), k = 1, \ldots m$ значения их детерминантов $S_k, k = 1, \ldots m$ меняют знаки только в конце последовательности.

• Пусть матрица $S(m,m)$ имеет $v$ отрицательных собственных чисел. В идеальном случае соответствующий порядок просмотра определяет последовательность главных миноров, где в первый раз становится отрицательным значение детерминанта $S_{m-v+1} < 0$, а знаки последующих $v-1$ детерминантов чередуются. Следовательно, не более, чем $v$ элементов множества $\Omega$ нарушают конфигурацию.

• Введем термин «локализация отрицательных собственных чисел» в неположительно определенной матрице и рассмотрим процедуру локализации.

# 4. Локализация отрицательных собственных чисел

*Процедура локализации*. Матрица $S(m,m)$ определяет последовательность ее главных миноров. Значение детерминанта $S_m$ матрицы $S(m,m)$ определяется произведением ее собственных чисел, где при нечетном числе отрицательных собственных чисел $S_m < 0$, а при четном – $S_m > 0$.

Просмотрим главные миноры и их детерминанты в обратном порядке $S_k, k = m,...1$. На текущем шаге $k$ вычисляются детерминанты $(S_k)_q^q, q = 1,...k$ всех дополнительных миноров $(S(k,k))_q^q$ текущего главного минора $S(k,k)$. Найдем среди них минор $(S(k,k))_{q_k}^{q_k}$, значение детерминанта $(S_k)_{q_k}^{q_k}$ которого изменит знак относительно детерминанта $S_k$ и окажется максимальным по абсолютной величине среди всех таких. Переставим строку и столбец с индексом $q_k$ на последнее место $k$ в миноре $S(k,k)$.

• Полученная перестановка строк и столбцов матрицы $S(m,m)$ определяет локально-оптимальную последовательность ее главных миноров, в которой их детерминанты убывают по абсолютной величине наиболее медленно. При этом чередование их знаков сосредоточено в конце последовательности главных миноров.

# 4. Локализация отрицательных собственных чисел

• Часто оказывается, что чередование знаков в конце оптимальной последовательности происходит на более широком интервале длины больше $v$, так как бывает невозможно получить смену знаков детерминантов на каждом шаге в конце последовательности.

• Пусть $u$ – дополнительное число детерминантов без смены знака для всех $v$ отрицательных собственных чисел. Рассмотренная процедура определяет не более, чем $v+u$ последних миноров в конце оптимальной последовательности на интервале чередования их знаков.

• В итоге, число метрических нарушений часто удается значительно снизить по сравнению с неоптимальной последовательностью главных миноров. Дело в том, что часто очередное метрическое нарушение вызывает *шлейф дополнительных нарушений*, число которых может быть значительным по сравнению с числом отрицательных собственных чисел. В оптимальной последовательности главных миноров число таких шлейфов и их длины удается свести к минимуму.

• Отсюда следует, что метрические нарушения оказались *связанными* с конкретными элементами множества $\Omega$. Это – *новое понимание роли отрицательных собственных чисел* в отличие от классического на основе разложения по ортогональному базису собственных векторов, где матрица скалярных произведений расслаивается на вклады собственных векторов пропорционально соответствующим собственным числам. Традиционно метрическое нарушение не связывается с конкретным элементом множества, а распределяется по всему множеству.

# 4. Локализация отрицательных собственных чисел

• В итоге, для метрической коррекции требуется сначала определить оптимальную последовательность главных миноров матрицы скалярных произведений, а потом выполнить коррекцию, например, методом, представленным выше.

*Пример. Коррекция близостей между белковыми последовательностями.* Одно из основных утверждений молекулярной биологии заключается в том, что первичная структура белка (последовательность аминокислотных остатков) содержит достаточную информацию о его пространственной структуре. Известно, что конфигурации белковых макромолекул, как правило, похожи в больших группах эволюционно близких белков.

Поэтому множество существенно различающихся пространственных структур белков значительно меньше множества известных белков. Как следствие, проблема выявления пространственной структуры может решаться как задача отнесения белковой макромолекулы к одному из ограниченного набора классов, т.е. как *задача распознавания*.

Из базы данных SCOP по принципу *наименьшей похожести* было выбрано 420 белковых последовательностей, образующих 51 класс белков. Похожесть белков друг на друга (score) в каждом из классов определялась по результатам парного выравнивания последовательностей аминокислотных остатков программой Fasta.

Таким образом, классы оказались малонаполненными и состоящими из не слишком похожих друг на друга белков с минимальным порогом похожести 27%.

# 4. Локализация отрицательных собственных чисел

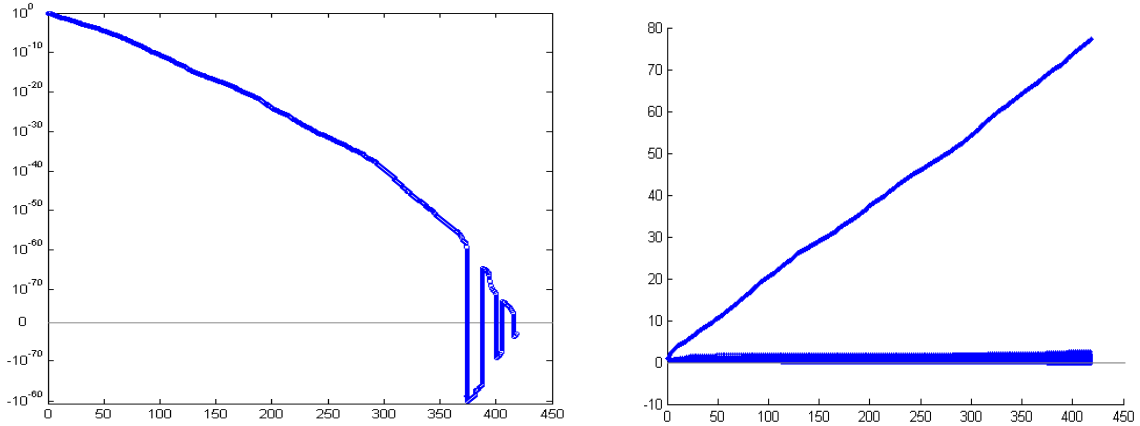
Это множество исследовалось в Национальной лаборатории Лоуренса (Беркли, США) на разделимость классов белковых макромолекул в наихудших условиях [I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.-H. Kim, "Recognition of a protein fold in the context of the SCOP classification", *Proteins: Structure, Function, and Genetics*, 1999, 35, 401–407.].

Здесь это множество было уменьшено до 418 белков (удалены белки с номерами 47 и 373), т.к. две пары белков (пара с номерами 45 и 47, пара с номерами 371 и 373) были представлены идентичными сравнениями в парах.

Парные сравнения представлены нормированной матрицей близости (из-за большого размера не представлена), содержащей 413 положительных собственных чисел в диапазоне (округлено) от 77.228337 до 0.010171 и пять отрицательных собственных чисел: –0.002455, –0.007835, –0.015917, –0.052960, –0.076052, где их общая сумма равна 418.

# 4. Локализация отрицательных собственных чисел

Таким образом, в последовательности главных миноров их детерминанты пять раз меняют знак, где последний детерминант отрицателен (Рис. 1). В исходной последовательности главных миноров детерминант 375-го главного минора впервые оказался отрицательным. В исходной последовательности скорректировано 43 главных минора. Это означает, что каждая из пяти корректировок вызвала шлейф из 8–9 дополнительных корректировок последующих главных миноров.




а) б)

Рис. 1. Характеристики исходной последовательности 418 главных миноров: а) значения детерминантов в логарифмическом масштабе, б) собственные числа

# 4. Локализация отрицательных собственных чисел

В оптимальной последовательности главных миноров (Рис. 2) первым отрицательным оказался детерминант 411-го главного минора. Далее смены знаков произошли у детерминантов 413, 416, 417 и 418-го главных миноров. Всего было скорректировано только 8 миноров, т.к. трижды не удалось найти элемент, изменяющий знак детерминанта очередного минора.

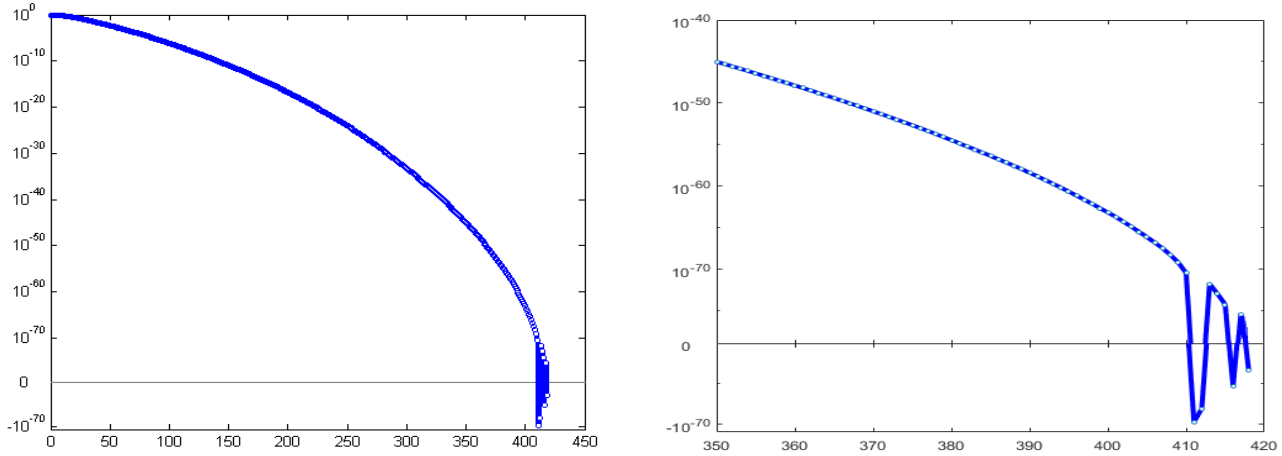



Рис. 2. Значения детерминантов оптимальной последовательности 418 главных миноров

# 4. Локализация отрицательных собственных чисел

• В итоге, все собственные числа скорректированной матрицы близостей оказались положительными в диапазоне от $\lambda_1 = 77.228$ до $\lambda_{418} = 6.927 \cdot 10^{-8}$. Заметим, что отношение $\lambda_1 / \lambda_{418} = 1.1148 \cdot 10^9$ является большим.

• Ниже рассматривается проблема обусловленности скорректированных матриц парных сравнений, где обусловленность матрицы скалярных произведений оценивается числом $Cond\,(S(m,m)) = \lambda_{\max} / \lambda_{\min} = \lambda_1 / \lambda_m$. Очевидно, что полученное сейчас число обусловленности 1 114 800 000 слишком велико (округлено).

# 5. Прямая коррекция собственных чисел

• Проблема обусловленности матриц известна. Традиционным источником матриц коэффициентов обычно являются системы уравнений и неравенств.

• Число обусловленности $Cond(A)$ некоторой матрицы $A$ показывает степень ее вырожденности. Если, например, в системе линейных уравнений $A\mathbf{x} = \mathbf{b}$ матрица $A$ плохо обусловлена, то малые изменения $A$ или $\mathbf{b}$ вызовут большие изменения в решении $\mathbf{x}$. Если же $A$ хорошо обусловлена, то малые изменения $A$ или $\mathbf{b}$ повлекут только соразмерные малые изменения в решении $\mathbf{x}$.

• Хорошо обусловленная матрица характеризуется небольшим числом обусловленности. Например, единичная матрица $E(m,m) = diag(1,...1)$ имеет наилучшую обусловленность $Cond(E) = 1$, где $m$ – размерность матрицы.

• Методы минимизации числа обусловленности матриц известны. Например, решение заключается в представлении исходной матрицы (в общем случае прямоугольной) в составе разложения, которое включает в себя левую и правую невырожденные диагональные матрицы. В итоге, задача сводится к проблеме обобщенных собственных чисел (GEVP).

# 5. Прямая коррекция собственных чисел

• Рассмотрим положительно определенную квадратную матрицу $S(m,m)$ парных сравнений. Число обусловленности определяется как произведение норм ее и обратной ей матриц $Cond(S) = \| S \| \cdot \| S^{-1} \|$. Норму матрицы можно определить разными способами, например, как максимальное по модулю собственное число $\| S \| = \max | \lambda |$. Тогда норма обратной матрицы окажется $\| S^{-1} \| = 1/\min | \lambda |$, т.к. собственные числа обратной матрицы обратны собственным числам исходной матрицы. Рассмотрим для положительно определенной матрицы $S(m,m)$ ее число обусловленности как $Cond(S) = \lambda_1 / \lambda_m$, где $\lambda_1 = \lambda_{\max}$, $\lambda_m = \lambda_{\min}$.

• Отрицательное собственное число нас не устраивает, т.к. в этом случае обусловленность будет отрицательна. В развиваемом здесь подходе это означает, что предварительно нужно выполнить метрическую коррекцию, чтобы обеспечить положительность $\lambda_{\min} > 0$.

• Минимизация обусловленности неизбежно связана с изменением матрицы $S(m,m)$. Это накладывает ограничения изменения ее элементов. Может оказаться, что требуемая обусловленность при ее минимизации противоречит допустимой величине изменения элементов матрицы *S*. Например, при сильных искажениях, с которыми нельзя согласиться.

# 5. Прямая коррекция собственных чисел

• Решение задачи $Cond(S) = \lambda_1 / \lambda_m \to \min$ напрямую не связано с изменяемыми элементами матрицы $S(m,m)$. Нам лишь известно, что $\det S = \prod_{i=1}^{m} \lambda_i$ и $\mathrm{tr}\, S = \sum_{i=1}^{m} \lambda_i$, где $\mathrm{tr}\, S = m$ для нормированной матрицы *S*. Можно лишь утверждать, что изменение элементов $s_{ij}$ при сохранении $\mathrm{tr}\, S = m$ должно привести к такому перераспределению значений собственных чисел, что $\lambda_{\max} = \lambda_1 > 0$ уменьшится, а $\lambda_{\min} = \lambda_m > 0$ возрастет, уменьшая значение $Cond(S)$.

• Дополнительное требование заключается в том, что степень перераспределения значений собственных чисел не должна приводить к недопустимо большим изменениям исходной матрицы парных сравнений.

• Пусть исходная матрица данных $X(m,n)$, представляющая множество $\Omega = \{\omega_1, \ldots \omega_m\}$ в некотором *n*–мерном признаковом пространстве, утрачена. Если это не так, то не представляет проблемы вычислить скалярные произведения $S(m,m) = (1/n) XX^T$.

• Пусть множество $\Omega = \{\omega_1, \ldots \omega_m\}$ представлено нормированной матрицей $S(m,m)$ парных сравнений. Известное разложение Карунена-Лоэва является спектральным разложением квадратной матрицы по системе собственных векторов.

# 5. Прямая коррекция собственных чисел

• Спектральное разложение невырожденной матрицы $S(m,m)$ имеет вид $S = A\mathrm{L}A^T$, где $A(m,m) = (\mathbf{a}_1,\dots \mathbf{a}_m)$ – ортогональная матрица собственных векторов-столбцов $\mathbf{a}_i = (a_{1i},\dots a_{mi})^T$, $\|\mathbf{a}_i\| = 1$, $A^T A = AA^T = E$, $E(m,m) = diag(1,\dots 1)$ – единичная матрица, $\mathrm{L}(m,m) = diag(\lambda_1,\dots \lambda_m)$ – диагональная матрица собственных чисел, $\lambda_1 \ge \dots \ge \lambda_m$. На основе разложения матрицы *S* получим $\mathrm{L} = A^T SA$, где $\mathrm{tr}\,\mathrm{L} = \mathrm{tr}\,S = m$.

• Обычно в анализе данных декомпозиция по системе ортогональных векторов применяется к корреляционной матрице признаков $R(n,n) = (1/m)X^T X$ с собственными числами $\mu_1 \ge \dots \ge \mu_n$ для снижения размерности пространства собственных векторов матрицы $R(n,n)$ до величины $n' < n$. В частности, некорректная матрица $R(n,n)$ имеет отрицательные собственные числа. Поэтому объекты исходной матрицы данных $X(m,n)$ проецируются в новое пространство размерности $n' < n$ с ортогональными осями координат. Новая размерность определяется только положительными собственными числами $\mu_1 \ge \dots \ge \mu_{n'} > 0$ ортогонального разложения (т.н. дискретное разложение Карунена-Лоэва).

• Обратим внимание, что устранение отрицательных собственных чисел можно понимать как их *прямое изменение* на нулевые значения.

# 5. Прямая коррекция собственных чисел

• Новая корреляционная матрица диагональна $R(n',n') = diag(\mu_1,\dots\mu_{n'})$, $\operatorname{tr} R' = \sum_{i=1}^{n'} \mu_i = n' < n$. Для проецирования объектов в новое пространство из исходной матрицы $X(m,n)$ нужно получить новую матрицу $X'(m,n')$. В этом случае общая вариация нормированных данных снижается до величины $n' < n$.

*В отличие от традиционного подхода, здесь спектральное разложение применяется для множества, элементы которого представлены только парными сравнениями (скалярными произведениями) в матрице* $S(m,m)$.

• В случае метрических нарушений в конфигурации элементов спектральное разложение $\mathrm{L} = A^T S A$ матрицы $S(m,m)$ имеет отрицательные собственные числа.

*Для устранения нарушений предлагается не редуцировать исходные парные сравнения, но прямо заменить отрицательные собственные числа подходящими положительными значениями.*

• В итоге, получим новую матрицу $\tilde{\mathrm{L}}(m,m)$ той же размерности и того же ранга. После этого восстановим новую матрицу парных сравнений $\tilde{S}(m,m)$ преобразованием $\tilde{S} = A\tilde{\mathrm{L}}A^T$.

• Новая матрица $\tilde{S}(m,m)$ окажется ненормированной с диагональными элементами больше единицы, т.к. $\operatorname{tr}\tilde{\mathrm{L}} = \operatorname{tr}\tilde{S} > m$. После нормировки $\hat{s}_{ij} = \tilde{s}_{ij} / \sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}$ получим окончательное разложение матрицы $\hat{S}(m,m)$, где $\hat{S} = \hat{A}\hat{\mathrm{L}}\hat{A}^T$, $\operatorname{tr}\hat{\mathrm{L}} = \operatorname{tr}\hat{S} = m$.

# 5. Прямая коррекция собственных чисел

• При таком подходе можно модифицировать *любые* собственные числа, а не только отрицательные. *Но что мы получим в результате*? Чтобы избежать бесконтрольных изменений, нужны *ограничения*. Например, следует обеспечить подходящую *обусловленность* скорректированной матрицы парных сравнений.

• Рассмотрим эту задачу. Известно, что матрица $S(m,m)$ скалярных произведений расслаивается на вклады от собственных векторов пропорционально соответствующим собственным числам. Обычно распределение значений упорядоченных по убыванию собственных чисел сильно неравномерно: на небольшую часть больших значений приходится основная доля дисперсии нормированных данных, которая равна размерности метрического пространства $m$. Очевидно, что вклады от малых значений собственных чисел в этих условиях малы. Поэтому их коррекция не приводит к сильному изменению исходной матрицы парных сравнений.

• Вместо того, чтобы искать представление исходной матрицы на основе спектрального разложения и оптимизировать ее обусловленность путем минимизации максимального собственного числа, предлагается, наоборот, *максимизировать минимальное* собственное число.

• Если собственные числа упорядочены по убыванию, то минимальное собственное число не может быть больше предыдущего. Поэтому будем непосредственно изменять значения меньших собственных чисел.

# 5. Прямая коррекция собственных чисел

• Для положительно определенной нормированной матрицы $S(m,m)$ скалярных произведений увеличим значения малых собственных чисел, увеличивая вклады соответствующих им собственных векторов в дисперсию нормированных данных. Т.к. сумма собственных чисел постоянна и равна $m$, то вклады больших собственных чисел уменьшатся. Следовательно, уменьшится и величина $\lambda_1$, что приведет к уменьшению $Cond(S) = \lambda_1 / \lambda_m$.

*Оптимизации обусловленности*. Определим для $S(m,m)$ последовательность собственных чисел $\lambda_1 > ... > \lambda_m > 0$ и обусловленность $Cond(S) = \lambda_1 / \lambda_m$. Просмотрим поочередно с конца собственные числа $\lambda_i, i = m-1,...1$. Определим для собственных чисел $\lambda_j, j = i+1,...m$ новые значения: $\lambda_j = \lambda_i,\ j = i+1,...m,\ i = m-1,...1$. Как и ранее $\mathrm{L}(m,m) = diag(\lambda_1,...\lambda_m)$ – матрица собственных чисел, $A(m,m) = (\mathbf{a}_1,...\mathbf{a}_m)$ – матрица нормированных собственных векторов $\mathbf{a}_i = (a_{i1},...a_{im})^T$, $\mathbf{a}_i^T\mathbf{a}_i = 1$, $\mathbf{a}_i^T\mathbf{a}_j = 0, i \neq i$.

На очередном шаге $i$ после определения новых *одинаковых* значений соответствующих собственных чисел получим матрицу $\tilde{\mathrm{L}}(m,m) = diag(\tilde{\lambda}_1,...\tilde{\lambda}_m)$ и восстановим матрицу $\tilde{S} = A\tilde{\mathrm{L}}A^T$. После такого восстановления $\mathrm{tr}\,\tilde{S} = \mathrm{tr}\,\tilde{\mathrm{L}} > m$, где элементы $\tilde{s}_{ii} > 1$. После преобразования $\hat{s}_{ij} = \tilde{s}_{ij} / \sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}$ получим нормированную матрицу $\hat{S}(m,m)$, где $\mathrm{tr}\,\hat{S} = m$.

# 5. Прямая коррекция собственных чисел

Снова выполним разложение и получим $\hat{\mathrm{L}} = \hat{A}^T \hat{S} \hat{A}$, где $\mathrm{tr}\,\hat{S} = \mathrm{tr}\,\hat{\mathrm{L}} = m$. Такое разложение вновь дает упорядоченную последовательность *различных* собственных чисел $\hat{\lambda}_1 > ... > \hat{\lambda}_m > 0$, где $\hat{\mathrm{L}}(m,m) = diag(\hat{\lambda}_1, ... \hat{\lambda}_m)$. Определим новую обусловленность $Cond(\hat{S}) = \hat{\lambda}_1 / \hat{\lambda}_m < Cond(S)$. Повторим процесс на следующем шаге $i$ снова для исходной матрицы $S(m,m)$.

На последнем шаге $i = 1$ все значения собственных чисел будут одинаковы $\tilde{\lambda}_1 = \tilde{\lambda}_2 = ... = \tilde{\lambda}_m$. После восстановления $\tilde{S}(m,m)$ и нормировки $\hat{S}(m,m)$ получим единичную матрицу $\hat{S}(m,m) = E(m,m)$, где $Cond(\hat{S}) = 1$. В этом случае матрица $\hat{S}(m,m)$ обладает минимальным значением обусловленности, но радикально отличается от исходной матрицы $S(m,m)$.

• Нужно определить условие остановки. Можно контролировать изменения элементов матрицы парных сравнений на основе *проверки статистических гипотез* об их значимости, рассматривая их как коэффициенты корреляции.

*Условие остановки*. Рассмотрим критерий Стьюдента проверки значимости корреляционной связи. В матрице *S* некоторые связи $s_{ij}$ окажутся значимыми. Минимизация числа обусловленности матрицы *S* приводит к единичной матрице, поэтому все скалярные произведения окажутся незначимыми. Процедура оптимизации обусловленности уменьшает число значимых связей. Критерием остановки является ситуация, когда число значимых связей резко падает.

# 6. Сохранение ранга матрицы парных сравнений

• Множество элементов $\Omega = \{\omega_1, ... \omega_m\}$ представлено измерениями $n$ признаков и рассматривается как погруженное в $n$–мерное евклидово пространство. В этом случае предпочтительна ситуация, когда $m > n$, т.к. мы не встретимся с т.н. проблемой «проклятия размерности». Это известная проблема в анализе данных и машинном обучении, возникающая при недостаточном числе актов измерений (объектов) в многомерном пространстве, когда $m \le n$. Такая ситуация приводит к излишней мощности линейного решающего правила.

• Считается, что реальные данные не должны быть разделимы всеми возможными способами в координатном пространстве. Именно при таком условии, когда некоторые разбиения невозможны, оставшиеся разбиения более адекватно характеризуют в координатном пространстве конфигурации взаимного расположения элементов множества. С другой стороны, известно, что парные сравнения множества из $m$ элементов можно погрузить в метрическое пространство размерности не более $m$ – мощности этого множества.

• Если представлены только парные сравнения, то в общем случае неизвестно, из какого координатного пространства были извлечены эти данные. Мы потенциально попадаем в «плохую» ситуацию, когда размерность $m$ метрического пространства всегда совпадает с кардинальностью множества, вложенного в него и представленного только парными сравнениями его элементов.

# 6. Сохранение ранга матрицы парных сравнений

• Конечно, обычно мы достаточно легко можем распознать размерность недоступного нам пространства на основе анализа значений собственных чисел матрицы близостей $S(m,m)$.

• Если действительно $m > n$, то матрица скалярных произведений $S(m,m)$ окажется положительно полуопределенной, так как множество элементов будет размещено в координатном пространстве размерности меньше, чем $m$. Ранг такой матрицы окажется меньше $m$, так как некоторые собственные числа окажутся нулевыми или «достаточно» малыми, что позволит рассматривать их как нулевые. Какие значения считать «достаточно» малыми, обычно решается, исходя из формулировки соответствующей оптимизационной задачи, как, например, в задаче *многомерного шкалирования*.

• В отличие от задачи *многомерного шкалирования*, мы рассматриваем ситуацию, когда не нужно восстанавливать в явном виде т.н. «пространство стимулов». Мы полагаем, что только парных сравнений уже достаточно для исследования взаимного расположения элементов множества в метрически корректных конфигурациях. Но в этом случае возникает следующее *противоречие*.

• С одной стороны, условие $rank\, S \le m$ позволяет предположить, что для реконструирования пространства стимулов его размерность достаточна. Напомним, что цель задачи многомерного шкалирования заключается в снижении размерности восстановленного пространства в наибольшей степени.

# 6. Сохранение ранга матрицы парных сравнений

• С другой стороны, необходимо вложить множество, представленное матрицей $S(m,m)$, в метрическое пространство без нарушений. Но тогда матрица $S(m,m)$ считается некорректной из-за нулевых собственных чисел.

• Обусловленность такой матрицы бесконечна, если рассмотреть отношение максимального и минимального собственных чисел.

• Так как задачу восстановления координатного пространства стимулов мы не рассматриваем, то следует восстановить полный ранг матрицы $S(m,m)$. В итоге, для разрешения противоречия нужно решить обратную проблему и *максимизировать размерность данных*.

• Очевидно, что в этом случае формально у нас нет достаточного числа измерений. Известно, что в такой ситуации обычно применяют «регуляризацию» задачи.

• Тогда требование оптимизации обусловленности при метрической коррекции следует рассматривать специфическую регуляризацию на уровне самих данных.

# 6. Сохранение ранга матрицы парных сравнений

*Пример. Сохранение ранга матрицы парных сравнений*. Рассмотрим известные данные Р. Фишера по ирисам [R.A. Fisher, “The use of multiple measurements in taxonomic problems”, *Annals of Eugenics*, **7**, 1936, 179–188.]. Они представляют измерения четырех признаков (длина и ширина чашелистика, длина и ширина лепестка) по 50 экземпляров ирисов трех видов (Iris Setosa, Iris Versicolor, Iris Virginica), всего 150 экземпляров, идущих подряд. Известно, что первый класс (Iris Setosa) хорошо отделен от двух остальных, которые частично пересекаются. В опубликованных исходных данных значения признаков экземпляров 102 и 143 совпадают. Оба они относятся к классу Iris Virginica. Поэтому, чтобы не терять один объект, мы просто слегка изменили значения признаков у 143 экземпляра. В обоих случаях матрицы корреляций признаков:

$$\begin{pmatrix} 1 & -0.1176 & 0.8718 & 0.8179 \\ -0.1176 & 1 & -0.4284 & -0.3661 \\ 0.8718 & -0.4284 & 1 & 0.9629 \\ 0.8179 & -0.3661 & 0.9629 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -0.1178 & 0.8715 & 0.8186 \\ -0.1178 & 1 & -0.4278 & -0.3649 \\ 0.8715 & -0.4278 & 1 & 0.9629 \\ 0.8186 & -0.3649 & 0.9629 & 1 \end{pmatrix}$$

и их собственные числа до: 2.9185, 0.9140, 0.1468, 0.0207 и после: 2.9182, 0.9141, 0.1467, 0.0209 изменений практически одинаковы. Отметим, что дисперсия нормированных данных равна 4, где первые три собственных числа объясняют 99.48% дисперсии данных. Это означает, что данные практически трехмерны.

# 6. Сохранение ранга матрицы парных сравнений

Матрица скалярных произведений объектов размером 150x150 содержит только положительные значения, которые в этом случае можно рассматривать как близости (из-за размера здесь не показана). Теоретически эта матрица содержит четыре положительных собственных числа, сумма которых равна 150, а все оставшиеся 146 собственных чисел равны нулю. Дисперсия данных равна 150.

При вычислении собственных чисел данной матрицы различные вычислительные методы дают различающиеся результаты для малых значений собственных чисел. В итоге, матрица близостей содержит три больших собственных числа (округлено): 128.51048, 21.12805 и 0.36148. Все остальные значения собственных чисел не превышают значений от $10^{-5}$ до $10^{-4}$, где четвертое собственное число не превышает $10^{-6}$. Среди всех собственных чисел с 4–го по 150–е встречаются как положительные, так и отрицательные значения. Таким образом, мы имеем *вычислительный мусор*, т.е. ошибки вычислительных методов.

Очевидно, что данная матрица близостей должна быть скорректирована для помещения в 150-мерное пространство. Метод оптимальной коррекции здесь совершенно не подходит, так как требуется исправить нарушения от 146 объектов, то есть почти от всех объектов. Лучше сразу скорректировать все парные сравнения.

Применим коррекцию на основе прямого изменения собственных чисел. Легко увидеть, что из-за большой разницы между третьим и четвертым собственными числами практически на 5 порядков нужно сразу же обеспечить при замене значительно меньшее различие.

# 6. Сохранение ранга матрицы парных сравнений

Избавиться от очень маленьких положительных и отрицательных собственных чисел удается только, начиная с коррекции собственных чисел с 5–го по 150–е значением четвертого собственного числа $6.8 \cdot 10^{-7}$. Согласно процедуре оптимизации обусловленности, которая была рассмотрена выше, восстановленная и скорректированная нормированная матрица близостей окончательно содержит следующие собственные числа (округлено): 128.51038998, 21.12803187, 0.36147827, $1.3885 \cdot 10^{-6}$ и остальные с 5–го по 150–е в диапазоне от $1.3881 \cdot 10^{-6}$ до $5.979 \cdot 10^{-9}$.

В итоге, исходные парные сравнения изменились совершенно незначительно, так как незначительно изменились первые три собственных числа. Тем не менее, достигнутая обусловленность (округлено) 21 493 798 795.15 оказывается недопустимо большой.

Тогда коррекция собственных чисел с 4–го по 150–е значением не более $10^{-3}$, позволяет получить число обусловленности уже на 5 порядков меньше: его величина составляет 128599.2.

В этом случае первые четыре собственных числа получают значения (округлено): 128.3846, 21.1074, 0.3611 и $\approx 10^{-3}$, где первые три собственных числа также лишь немного изменились.

По сравнению с исходной матрицей парных сравнений, у которой теоретически бесконечное число обусловленности, сумма квадратов отклонений значений элементов измененной матрицы от элементов исходной матрицы составляет величину $D = 0.016418$. Поэтому относительно дисперсии данных $\sigma^2 = 150$ взвешенное суммарное отклонение $(100\% \cdot D / \sigma^2)$ от нее составляет

примерно 0.01%. Таким образом, парные сравнения оказались корректно вложенными в 150–мерное пространство практически неизменными.

Заметим, что результаты кластеризации в обоих случаях (некорректная и корректная матрицы близостей ирисов) оказываются одинаковыми. В обоих случаях первый класс (Iris Setosa) полностью отделим от остальных. Из-за того, что второй и третий классы частично пересекаются, они разделяются с 6 ошибками. При этом три объекта из второго класса (69, 74 и 84) отнесены к третьему классу, а три объекта из третьего класса (111, 139 и 142) отнесены ко второму классу.

- Показано, что модификация парных сравнений для их корректного вложения в многомерное метрическое пространство не вызвала ухудшения результата обработки.
- Для корректного вложения парных сравнений в метрическое пространство могут потребоваться изменения некоторых из них. Если представить, что они были ранее намеренно искажены, то их восстановление совсем не означает, что будут получены исходные «истинные» значения.
- Будут восстановлены лишь такие значения, которые позволят исключить метрические нарушения.

# 6. Сохранение ранга матрицы парных сравнений

• Следует отметить, что результат кластеризации в 150–мерном пространстве несколько отличается от результата кластеризации в 4–мерном пространстве стандартизированных признаков. Там второй и третий классы разделены с 25 ошибками, где 11 объектов из второго класса (51, 52, 53, 57, 66, 71, 76, 77, 78, 86, 81) отнесены к третьему классу, а 14 объектов из третьего класса (102, 107, 114, 115, 120, 122, 124, 127, 134, 135, 139, 143, 147, 150) отнесены ко второму классу.

• Отметим, что множество объектов, представленное исправленной матрицей их скалярных произведений размером 150x150, оказывается, в итоге, корректно погруженным в 150–мерное метрическое пространство.

• Преодолеть «проклятие размерности» можно при проецировании этих нормированных парных сравнений в пространство нескольких первых главных компонент, объясняющих, например, 80% дисперсии парных сравнений. Т.е. следует рассматривать аналог известного *дискретного разложения Карунена-Лоэва*, которое применяется по отношению к традиционной матрице данных.

# Заключение

• В задаче анализа данных формируется пространство признаков (атрибутов, характеристик) , в котором реальные объекты представлены числовыми векторами. Часто содержательно важно и удобно, чтобы это пространство воспринималось как интуитивный аналог реального пространства. Для этого необходимо, чтобы матрицы парных сравнений похожести оказались положительно определенными. Тогда они могут рассматриваться как матрицы скалярных произведений в одном и том же квадранте метрического пространства.

• При погружении элементов множества в метрическое пространство могут возникать нарушения конфигураций, что приводит к появлению отрицательных собственных чисел. Возникает необходимость их коррекции (изменения собственных чисел). Другие задачи погружения элементов множества в пространство признаков могут также рассматриваться как задачи изменения собственных чисел (проецирование в подпространство собственных векторов – дискретное разложение Карунена-Лоэва).

• Так как матрицы скалярных произведений потенциально могут рассматриваться как данные для решения задачи разделения на подмножества (например, построение разделяющей гиперплоскости), то важно иметь их хорошую обусловленность. Эта задача также решается изменением собственных чисел. Также рассмотрена задача сохранения ранга матрицы.

• В общем случае, по-видимому, следует рассматривать различные задачи, которые требуют соответствующей прямой коррекции собственных чисел таких матриц.

# Заключение

В докладе показано:

1. нарушения метрики возникают не только на тройках элементов множества, но и на подмножествах, содержащих более трех элементов;

2. отрицательное собственное число матрицы скалярных произведений можно непосредственно связать с элементом множества, который нарушил метрику;

3. собственные числа в ортогональном разложении матрицы скалярных произведений можно менять совершенно произвольно, но лишь наличие ограничений позволяет сформулировать задачу оптимизации;

4. задача оптимизации обусловленности и задача сохранения ранга матрицы рассмотрены как задачи прямой коррекции собственных чисел.