Применение t-распределения Стьюдента при метрической коррекции матриц парных сравнений

ИОИ-2018
8-12.10.2018
гаэта
италия

Двоенко С.Д.
Пшеничный Д.О.

Тульский государственный университет

Положительная определённость

- Пусть парные сравнения похожести (близости) n элементов множества представлены матрицей $S(n,n)$.
- Если $S(n,n)$ положительно определена, то элементы множества могут быть погружены в n-мерное евклидово пространство в виде векторов, различие между которыми рассматривается как расстояние, а похожесть – как скалярное произведение.
- Согласно теореме косинусов, расстояния можно преобразовать в скалярные произведения и наоборот.

Неположительная определённость

- В общем случае экспериментальная матрица парных сравнений похожести (близости) не будет положительно определенной. Это не позволяет считать элементы множества погруженными в неизвестное нам признаковое пространство.
- Для математически корректной обработки таких наблюдений в виде близостей необходимо скорректировать некоторые парные близости. Только в этом случае мы можем быть уверены, что при «внезапном» появлении признакового пространства результаты обработки не изменятся.
- Решим задачу так, чтобы элементы матрицы $S(n,n)$ были изменены в минимальной степени.

Новый подход к коррекции

- **Традиционной подход** основан на т.н. дискретном разложении Карунена-Лоэва, когда из квадратной матрицы нормированных скалярных произведений (корреляций в статистике) «послойно» устраняются вклады собственных векторов $\mathbf{a}\mathbf{a}^T$, соответствующих отрицательным собственным числам (о.с.ч.).
- **Новый подход** основан на разработанной нами ранее технологии, позволяющей корректировать как все, так и некоторые парные сравнения отдельных элементов в любой квадратной матрице сходства или различий.
- **Новизна** заключается в том, что о.с.ч. «локализуются», т.е. их появление связывается с метрическими нарушениями, вносимыми конкретными элементами множества, а не «всеми сразу».

Новый подход к коррекции

- Естественно потребовать минимизации отклонений скорректированных парных сравнений от исходных.
- Но такое естественное требование приводит к нулевому детерминанту скорректированной матрицы, что, как минимум, говорит о некорректной вложенности множества в гипотетическое пространство избыточной размерности.
- Поэтому необходимо обеспечить некоторое небольшое, но положительное значение детерминанта скорректированной матрицы парных близостей (сходства).
- Выбор такого значения немедленно приводит к **проблеме обусловленности матрицы** парных сравнений сходства.

Обусловленность матрицы парных сравнений

- Число обусловленности квадратной матрицы показывает степень ее вырожденности.
- Если квадратная матрица A коэффициентов системы линейных уравнений $A\mathbf{x} = \mathbf{b}$ почти вырождена, то малые изменения A и $\mathbf{b}$ вызовут большие изменения в решении $\mathbf{x}$.
- Если матрица коэффициентов невырождена (например, близка к единичной), то малые изменения A и $\mathbf{b}$ повлекут малые изменения в решении $\mathbf{x}$.
- Невырожденная квадратная матрица S характеризуется небольшим числом обусловленности $Cond(S)$.

Число обусловленности

- Число обусловленности матрицы S определяется как произведение норм ее и обратной ей матриц

$$Cond(S) = \| S \| \cdot \| S^{-1} \|, \ \| S \| = \max |\lambda|, \ \| S^{-1} \| = 1 / \min |\lambda| .$$

- Для положительно определенной матрицы $S(n,n)$ ее число обусловленности имеет вид

$$Cond\left(S(n,n)\right) = \lambda_1 / \lambda_n , \text{ где } \lambda_{\max} = \lambda_1 > \ldots > \lambda_n = \lambda_{\min} > 0 .$$

- Метрические нарушения приводят к появлению отрицательных собственных чисел, а коррекция их устраняет.
- Такое определение числа обусловленности является приемлемым в задаче коррекции метрических нарушений.

Число обусловленности и детерминант

- Но число обусловленности $Cond(S(n,n))$ матрицы $S(n,n)$ не связано явно со значением ее детерминанта $S_n = \det S(n,n)$, хотя известно, что $S_n = \prod_{i=1}^{n} \lambda_i$.

- Для нормированной матрицы $S(n,n)$ из условия $n = \sum_{i=1}^{n} \lambda_i$ можно лишь утверждать, что $\lambda_{\max} = \lambda_1 > 0$ уменьшится, если остальные собственные числа не изменятся при возрастании $\lambda_{\min}$, когда оно станет положительным $\lambda_{\min} = \lambda_n > 0$.

Конфликт требований

- Наилучшее решение задачи коррекции, доставляющее минимум отклонения от $S_n < 0$, определяется значением $S_n = 0$.
- Очевидно, что такое решение **неприемлемо**, т.к. $\lambda_{\min} = \lambda_n = 0$, а $Cond\left(S(n,n)\right) = \lambda_1 / \lambda_n = \infty$.

- Указанный конфликт необходимо разрешить, позволив сильнее скорректировать элементы последней строки и столбца матрицы $S(n,n)$.

Конфликт требований

- С другой стороны, значение детерминанта скорректированной матрицы не может превысить значения предыдущего минора $S_n \le S_{n-1}$.
- Если после коррекции $S_n = S_{n-1}$, то по правилу вычисления детерминанта как разложения по элементам последней строки следует, что эта строка должна быть нулевой, как и последний столбец $s_{ni} = s_{in} = 0, i = 1,\dots n$.
- Тогда число обусловленности для данной матрицы минимально $Cond\left(S(n,n)\right) = Cond\left(S(n-1,n-1)\right)$.
- Такая коррекция **неприемлема** из-за недопустимого расхождения между исходными и новыми значениями элементов скорректированной матрицы $S(n,n)$.

Критерий Сильвестра

- Рассмотрим матрицу парных сравнений $S(n,n)$, где $s_{ij} = s_{ji}, -1 < s_{ij} < 1, s_{ii} = 1.$
- Пусть S – матрица некоторой квадратичной формы, $S(1,1) = 1, S(2,2), \ldots S(k,k), \ldots S(n,n)$ - последовательность главных миноров.
- **Критерий Сильвестра**: квадратичная форма (и её матрица S) является положительно определённой, тогда и только тогда, когда все главные миноры её матрицы положительны $S_k = \det S(k,k) > 0, k = 1,\ldots n.$
- **Следствие:** число о.с.ч. матрицы S в точности совпадает с числом знакоперемен в последовательности главных миноров.

Оптимальная коррекция

- Просматриваются последовательно все главные миноры. Их детерминанты убывают
$$S_1 = 1 > \ldots > S_k = \det S(k,k) > \ldots > S_n = \det S(n,n).$$
- Если значение очередного минора $S_k < 0$, то он корректируется, чтобы $S_k > 0$.
- Значение очередного минора вычисляется разложением по элементам последних k-ой строки и k-го столбца

$$S_k = S_{k-1}\left(1 - \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} s_{ki} s_{jk} r_{ij}\right),$$

где $r_{ij} = (-1)^{i+j}(S_{k-1})_i^j / S_{k-1}$ - элементы обратной матрицы $R = S^{-1}(k-1,k-1)$, $(S_{k-1})_i^j$ – значение минора $S(k-1,k-1)$ без строки j и без столбца i.

Оптимальная коррекция

- В общем случае мы корректируем лишь некоторые элементы последних k-ой строки и k-го столбца текущего минора с индексами из множества $I \subseteq \{1,\dots k-1\}$.
- После коррекции получим $S_k = C,\ 0 \le C \le S_{k-1}$.
- Обозначим элементы $s_{ki} = s_{ik}$ корректируемого минора как переменные x_i, $i = 1,\dots k-1$ и получим ограничение

$$\sum_{i=1}^{k-1}\sum_{j=1}^{k-1} x_i x_j r_{ij} = 1 - C / S_{k-1} = 1 - S_k / S_{k-1} = 1 - \tau .$$

- Задача минимизации отклонений при ограничениях:

$$\sum_{i=1}^{k-1} (s_{ik} - x_i)^2 \to \min;\ \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} x_i x_j r_{ij} = 1 - \tau = c .$$

Оптимальная коррекция

- Решение методом множителей Лагранжа приводит к системе уравнений

$$\begin{cases} \lambda \sum_{i \in I} x_i r_{ip} + \sum_{i \notin I} s_{ki} r_{ip} = s_{kp} - x_p, \quad p \in I \\ \sum_{i \in I} \sum_{j \in I} x_i x_j r_{ij} + \sum_{i \in I} \sum_{j \notin I} x_i s_{jk} r_{ij} + \sum_{i \notin I} \sum_{j \in I} s_{ki} x_j r_{ij} + \sum_{i \notin I} \sum_{j \notin I} s_{ki} s_{jk} r_{ij} = c, \end{cases}$$

  которая решается численным методом.
- Число уравнений зависит от индексов $p \in I$.
- Параметр оптимизации c неявно связан с числом обусловленности $Cond(S(k,k))$.

Разрешение конфликтов

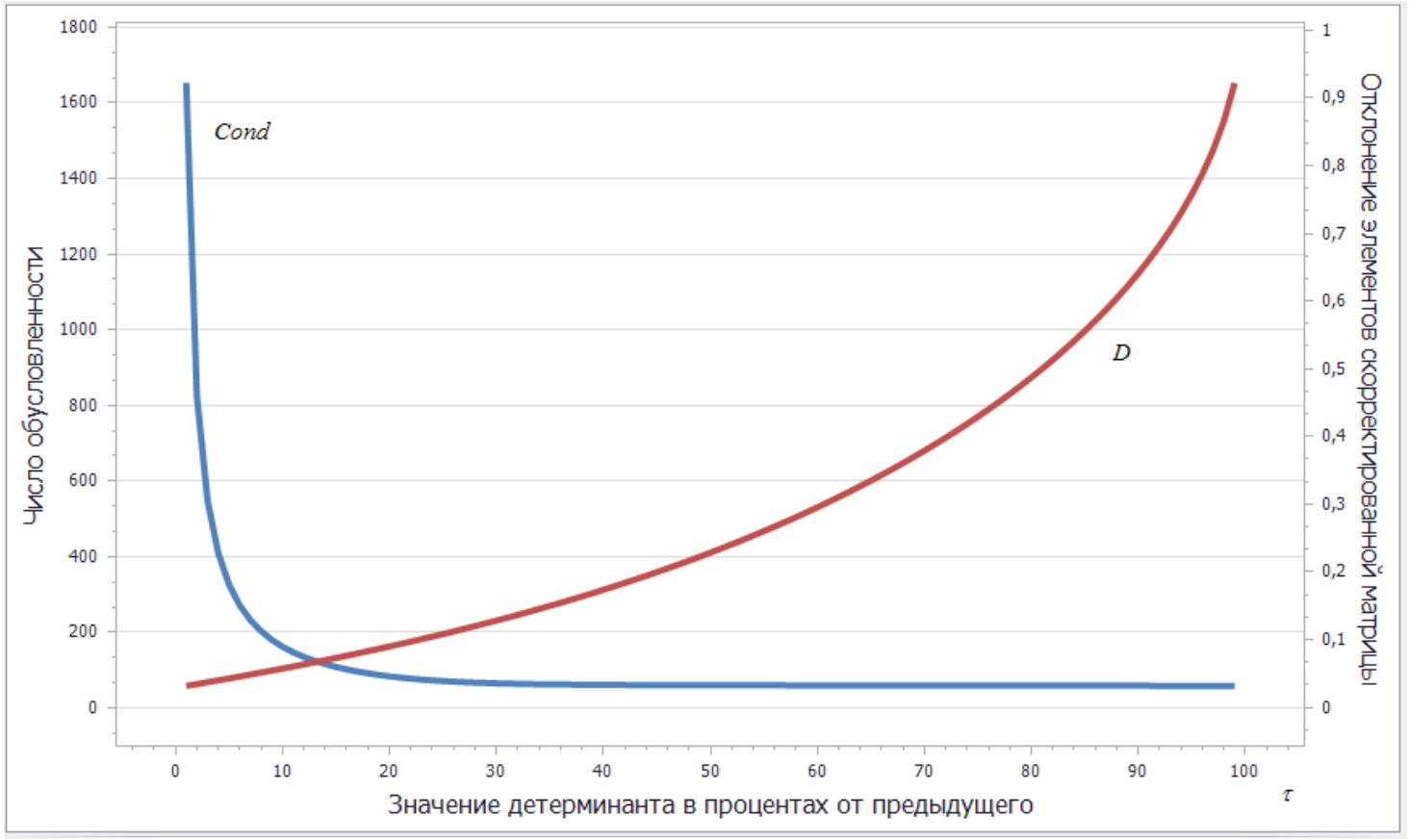
- Пусть $\tau = S_k / S_{k-1}$ - доля от S_{k-1} минора $S(k-1,k-1)$, где $c = 1-\tau$ - параметр оптимизации в задаче минимизации отклонений

$$\sum_{i=1}^{k-1}(s_{ik} - x_i)^2 \to \min; \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} x_i x_j r_{ij} = c\,.$$

- Построим с некоторым шагом из интервала $0 \le \tau \le 1$ $Cond_\tau\left(S(k,k)\right)$ - убывающий график числа обусловленности и $D_\tau = \sum_{i=1}^{k-1}(s_{ik} - x_i)^2$ - возрастающий график суммарного отклонения .
- Точка пересечения графиков такого вида часто рассматривается как оптимальная в некотором эвристическом смысле.

Разрешение конфликтов

- Эксперименты показывают, что найденный по точке пересечения уровень коррекции τ часто еще можно повысить.



Проверка гипотез об уровне коррекции

- Более строгий принцип заключается в выдвижении соответствующих статистических гипотез.
- Рассмотрим нормированные близости $S(n,n)$ как выборочные коэффициенты корреляции, считая, что это – результаты парных сравнений вариационных рядов наблюдений.
- Для сравнения на заданном уровне значимости α выдвигается нулевая гипотеза об отсутствии корреляции.
- Критерий проверки $T = x_p \sqrt{(v-2)/(1-x_p^2)}$, $p = 1,\ldots k$ имеет t-распределение Стьюдента с $v-2$ степенями свободы для выборки размера v для двусторонней критической области.

Проверка гипотез об уровне коррекции

- Для критической точки $t = t(\alpha, v-2)$ значимый уровень парной близости определяется как величина

$$x(\alpha, v) = \sqrt{t^2 / (t^2 + v - 2)} \ .$$

- Поэтому для каждого уровня коррекции $0 \le \tau \le 1$ можно определить число m значимых значений парных близостей $x_p > x(\alpha, v),\ p = 1, \ldots k$.
- Вследствие плавного изменения оптимальных отклонений и числа обусловленности из-за плавного изменения *max* и *min* собственных чисел, скорректированные элементы также изменяются плавно, возрастая или убывая.

Проверка гипотез об уровне коррекции

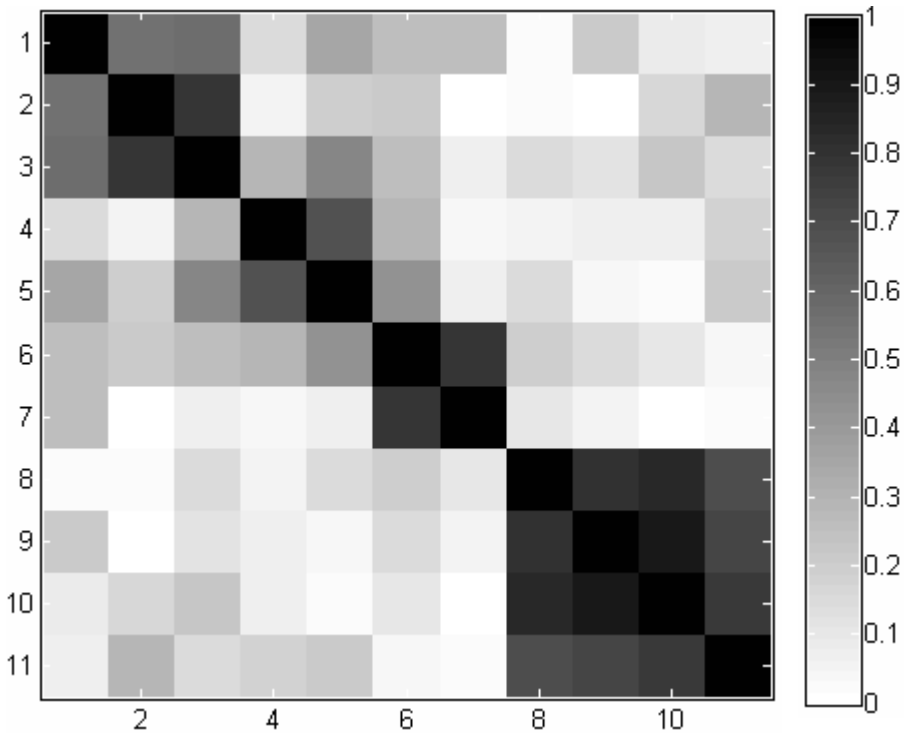
- Тем не менее, при повышении уровня коррекции до $\tau = 1$, все скорректированные значения, в итоге, начнут уменьшаться и окажутся нулевыми $x_p = 0,\ p = 1,\dots k$.
- Таким образом, оптимальное число обусловленности должно соответствовать пороговому уровню коррекции τ, начиная с которого число m значимых значений парных близостей $x_p > x(\alpha, \nu),\ p = 1,\dots k$ резко падает.
- Проблема заключается в том, что гипотетические «выборки», якобы использованные для вычисления корреляций вариационных рядов , и их размер ν - неизвестны.
- Этот размер нужно определить.

Проверка гипотез об уровне коррекции

- При заданном уровне значимости, например $\alpha = 0.01$, значимая парная близость определяется критической точкой *t*-распределения Стьюдента:
  - при статистически достаточном размере выборки $\nu \geq 122$ как $x(\alpha, \nu) = 0.208$,
  - при среднем размере выборки $\nu \geq 62$ как $x(\alpha, \nu) = 0.2948$,
  - при небольшом размере выборки $\nu \geq 32$ как $x(\alpha, \nu) = 0.4097$.
- Пороговый уровень коррекции τ определяется резким и необратимым падением числа значимых парных близостей относительно этих критических значений.

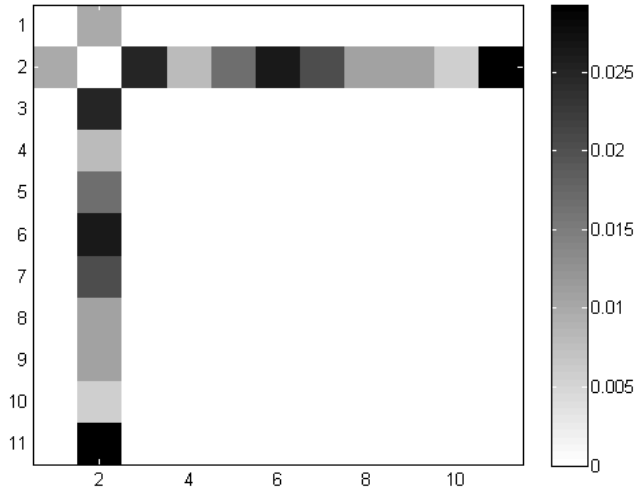
ЭЭГ биоритмов головного мозга (11x11)

- Корреляционная матрица статистических взаимосвязей между энергетическими свойствами биоритмов головного мозга для 11 частот навязанных ритмов (Небылицын В.Д., 1966 г.)



ЭЭГ биоритмов головного мозга (11x11)

- При просмотре главных миноров их приходится переупорядочивать, чтобы элементы, вносящие нарушения, оказались в конце. Можно показать, что неоптимальная последовательность миноров вызывает при их корректировке дополнительный шлейф искажений, которые тоже нужно корректировать. Поэтому общее число коррекций может оказаться значительно больше числа о.с.ч.
- Перестановка : 7 4 8 1 3 11 5 9 6 10 **2**



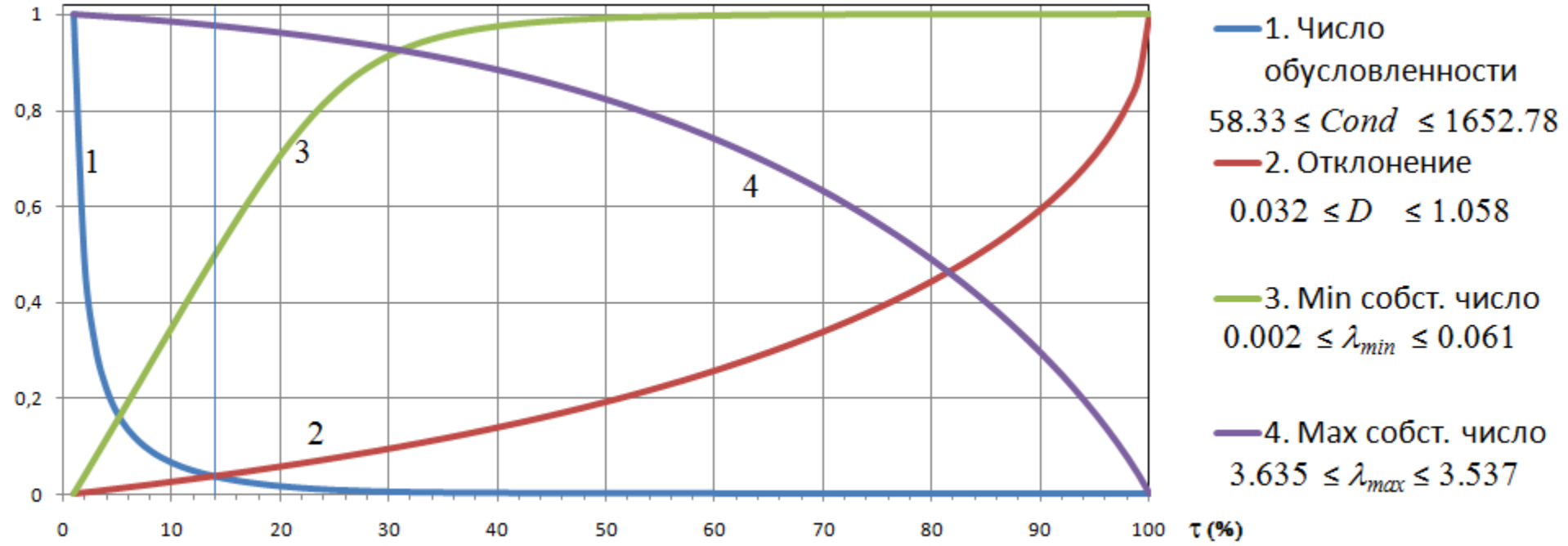
КОРРЕКЦИЯ ВТОРОГО ЭЛЕМЕНТА

(расположение элементов показано до перестановки)

- Собств.числа: 3.64, 2.83,1.61, 1.36, 0.52, 0.41, 0.28, 0.16, 0.15, 0.07, **-0.024**
- Детерминанты: 1, 0.999, 0.985, 0.893, 0.539, 0.256, 0.109, 0.025, 0.00475, 0.00048, **-0.000057**

ЭЭГ биоритмов головного мозга (11x11)

Корректировка строки



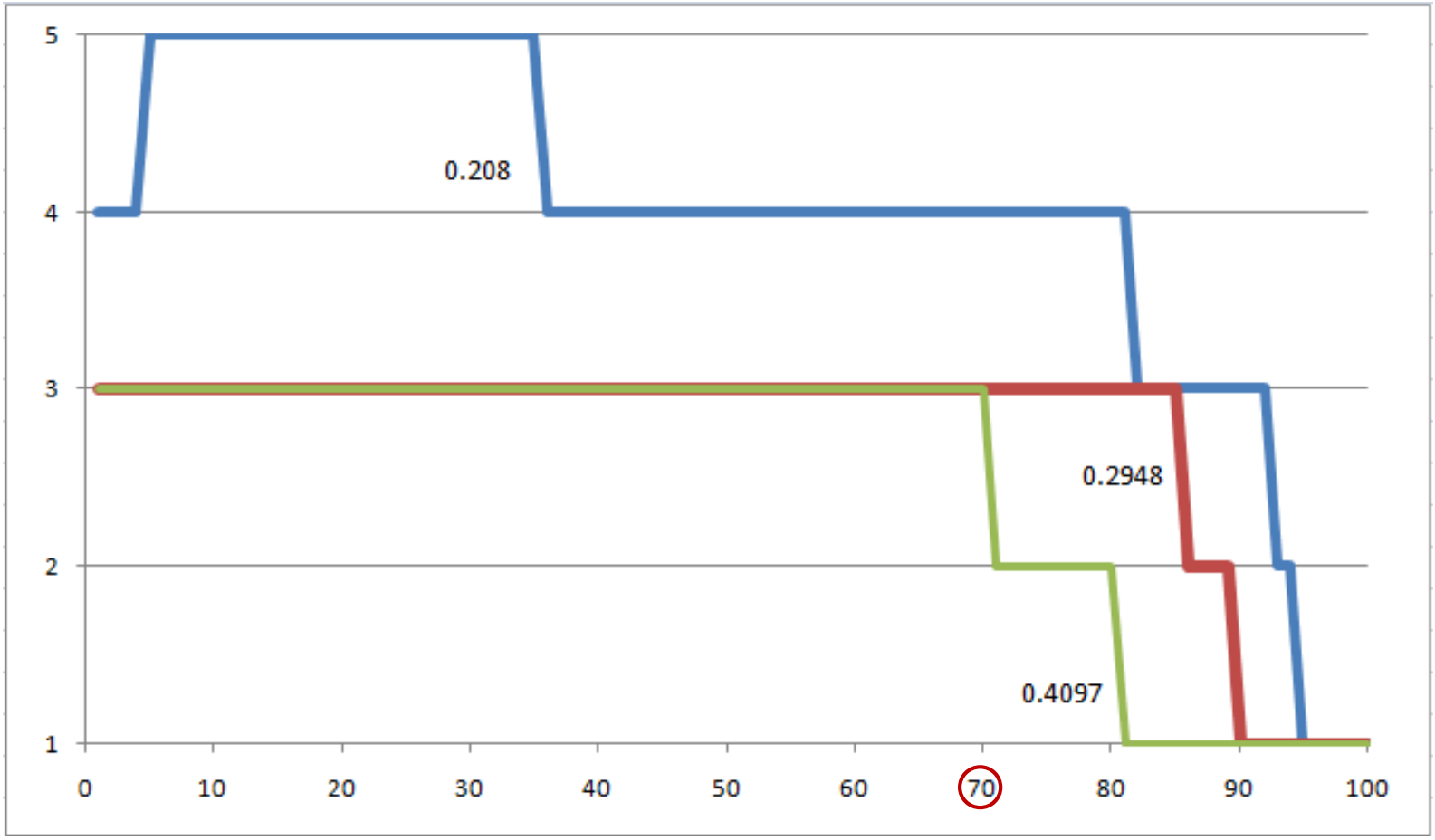
τ=0.01 *Cond* = 1652.784 почти теоретический минимум отклонений

τ=0.14 *Cond* = 116.027 эвристика - уменьшено в 14.24 раза

τ=0.7 *Cond* = 59.407 статистически - еще уменьшено в 1.95 раз, итого в 27.8 раза

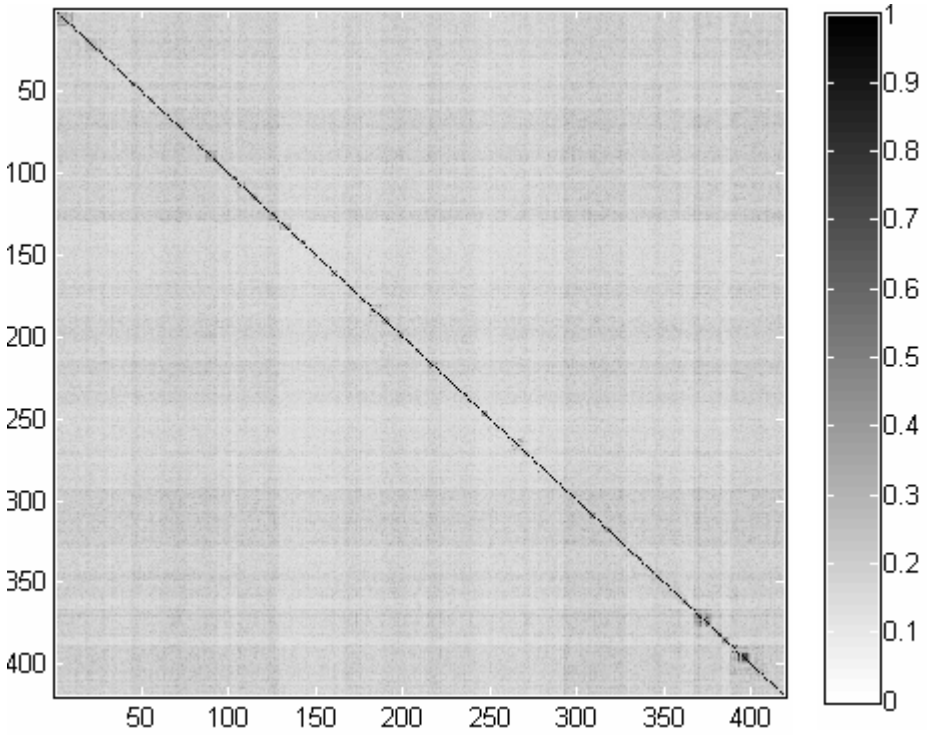
ЭЭГ биоритмов головного мозга (11x11)

- Изменение числа значимых элементов матрицы S (11,11)

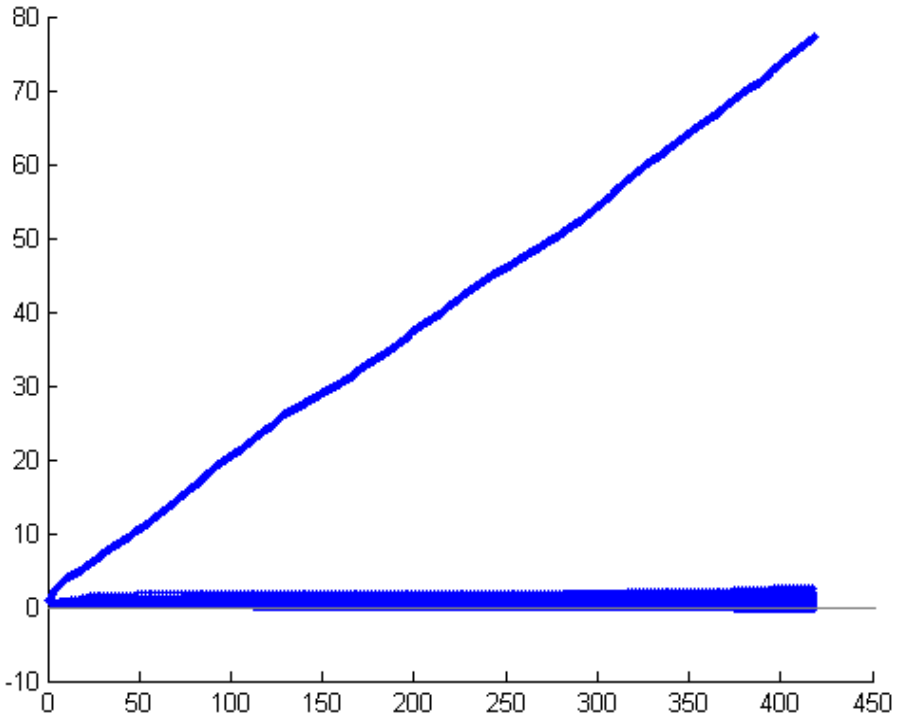


Близости белковых последовательностей (418x418)

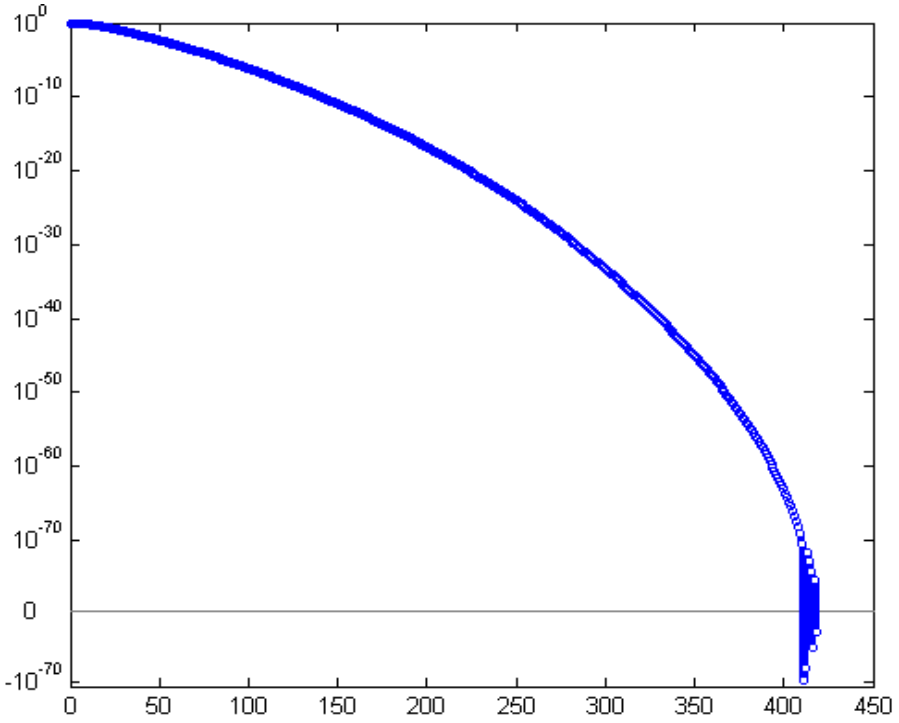
- Матрица нормированных близостей 418 белковых последовательностей (S.-H. Kim, 1999 г.)
- 415 с.ч. в диапазоне:
  77.767 ÷ 0.0103
- Пять о.с.ч.:
  -0.002479,
  -0.008042,
  -0.016319,
  -0.053135,
  -0.077528



Близости белковых последовательностей (418x418)

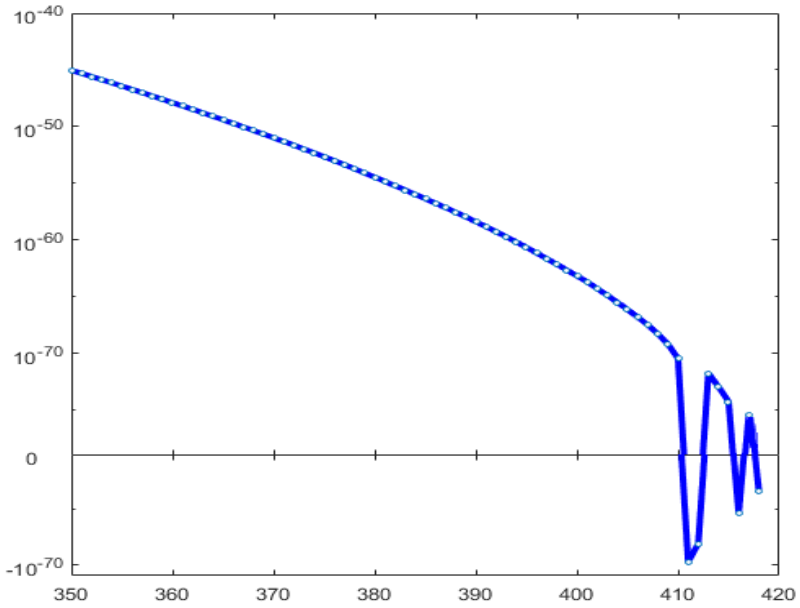


Собственные числа

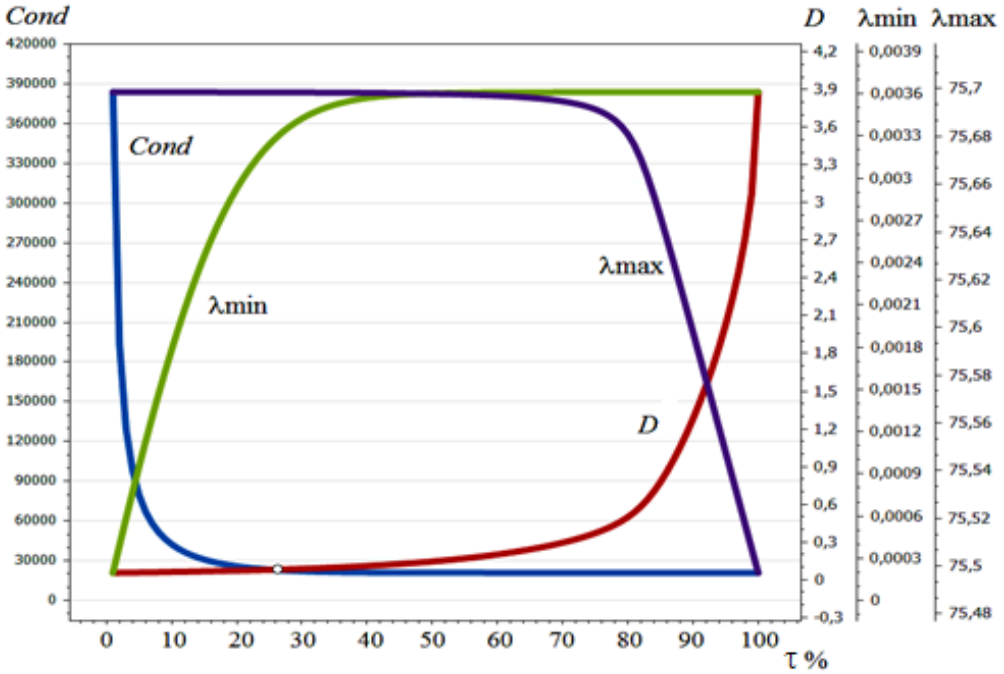


Значения главных миноров в оптимальной последовательности

Близости белковых последовательностей (418x418)



Знакоперемены в последних минорах с 411 по 418



Корректировка строки минора $S(411,411)$

τ=0.01 *Cond*=383759.99

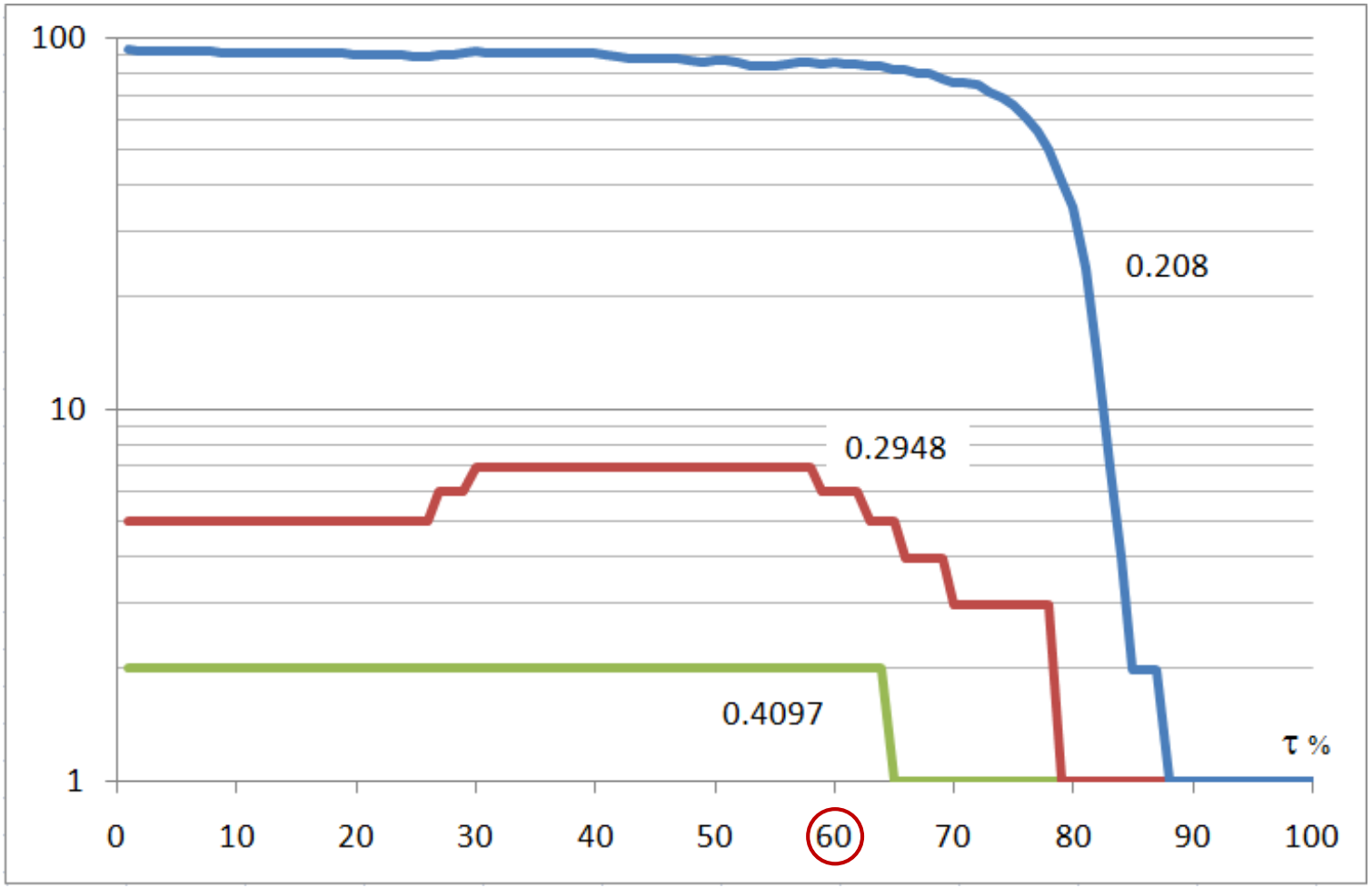
τ=0.26 *Cond*=23036.74 – меньше в 16.66 раз

τ=0.3 *Cond*=22110.24 – меньше в 17.36 раз

τ=0.6 *Cond*=20972.27 – меньше в 18.30 раз

Близости белковых последовательностей (418x418)

- Изменение числа значимых элементов минора $S(411,411)$



Близости белковых последовательностей (418x418)

Числа обусловленности всех скорректированных миноров

Size of minors	$\tau = 0.01$	$\tau = 0.3$		$\tau = 0.6$				
	Cond	*Cond*	*Cond* ratio	*Cond*	*D*	λ_{min}	λ_{max}	*Cond* ratio
411	383759.99	22110.24	17.36	20972.27	0.201	0.0036	75.697	18.30
412	83425.31	25926.10	3.22	21025.49	0.372	0.0036	75.889	3.97
413	279173.09	26320.00	10.61	21086.91	0.234	0.0036	76.100	13.24
414	113028.51	27741.61	4.07	21144.72	0.399	0.0036	76.308	5.35
416	49729.21	44612.87	1.11	30039.22	0.489	0.0026	76.767	1.66
417	295262.92	44746.38	6.60	30123.56	0.144	0.0026	76.982	9.80
418	79583.51	44914.97	1.77	30120.85	0.297	0.0026	77.204	2.63

Выводы

- Детерминант скорректированной матрицы парных сравнений остается положительным. Его значение можно задать заранее, регулируя степень коррекции.
- Чем меньше число скорректированных элементов, тем сильнее их отклонения от исходных значений. Поэтому не всегда удается скорректировать одиночные элементы.
- В общем случае можно корректировать только некоторые парные сравнения объектов, вносящих метрические искажения. Это позволяет выбирать, какие именно парные сравнения следует скорректировать.
- Требование минимизации отклонений при коррекции приводит к (почти) нулевому детерминанту. Скорректированная матрица оказывается плохо определенной с (почти) бесконечным числом обусловленности.

Выводы

- Оптимальная коррекция позволяет обеспечить приемлемое число обусловленности после коррекции.
- Тем не менее, значение скорректированного детерминанта лишь неявно связано с числом обусловленности.
- Предложен **статистически обоснованный** способ определения оптимального числа обусловленности.
- Предложены рекомендации для применения практической технологии коррекции произвольных матриц парных сравнений.