***Matrix comparisons***

SPSS macros by Kirill Orlov

kior@comtv.ru, ttnphns@gmail.com

<https://www.spsstools.net/en/KO-spssmacros>

All rights reserved

*Differences inside or between matrices.* Macros compute a matrix of distances between matrices of proximity coefficients (rather than between variables or cases), - such as correlation or distance matrices; or between columns inside such matrices. These comparisons can help a researcher: for example, before a cluster or a factor analysis.

# MACRO !KO\_DMATRIX: DISTANCES BETWEEN COLUMNS OF A PROXIMITY MATRIX

Version 1 (Aug 2002). Tested on SPSS 11, 11.5, 13, 14.

!KO\_dmatrix cols= *v1 v3* to *v7* /\*Variables (matrix columns), all or needed; may write via “to”

/trans= NONE /\*Transform values in matrix columns: NONE - no (defaulf), CENTER - center, ZSCORE - standardize,

/\*MEANSS – bring mean SS to 1, RANGE – bring range to 1, RESCALE – bring range to 0 - 1

/measure= BLOCK /\*Distance measure to calculate between columns: EUCLID (euclidean), SEUCLID (square euclidean), BLOCK (manhattan, default).

Minimal specification COLS.

Макрос вычисляет попарно расстояния между столбцами какой-л. квадратной матрицы близостей, – вообще какой-л. диагонализованной квадратной матрицы, – т.е. матрицы, столбцы и ряды которой задает одна и та же последовательность «предметов» (под словом «предметы» подразумеваются или респонденты, или объекты, или признаки). Матрица коэффициентов взаимокорреляции переменных или матрица сходств или различий между респондентами – типичные примеры подобной матрицы.

Математическое расстояние между столбцами матрицы это величина различия профилей столбцов. Сравнение профилей коэффициентов в корреляционной или в иной «матрице мер близости» (чем и занимается данный макрос) может быть исследователю существенно важно. К примеру, если две переменные показали между собой высокую корреляцию, то и профили их корреляционных коэффициентов с остальными переменными, скорее всего, будут похожи, однако если они показали слабую корреляцию между собой, то их профили корреляции с прочими переменными могут как сильно различаться, так и быть довольно сходными.

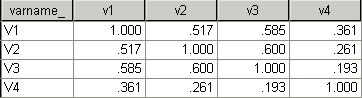
При вычислении расстояния между любыми двумя столбцами i и j макрос *игнорирует элементы* (i,i), (j,j), (i,j), (j,i). Другими словами, диагональные коэффициенты и коэффициенты близости между этими двумя столбцами не участвуют в определении расстояния между ними. Указанные элементы не участвуют и при предварительной стандартизации столбцов, которую вы можете заказать.

Входящая матрица в общем не обязана быть симметрической (бывают матрицы близости несимметрического строения). Если она не симметрическая, то результирующие расстояния будут неодинаковы для матрицы и ее транспоната (т.е. если матрицу вы предварительно транспонируете, превратив столбцы в ряды и наоборот, результат будет уже другой). Результат работы макроса – новый рабочий файл, являющийся оформленной по SPSS-правилам квадратной симметрической матрицей расстояний. Вы можете использовать эту матрицу в таких видах анализа как кластерный или многомерное шкалирование.

# Подкоманды

## COLS

Укажите все или нужные столбцы входящей квадратной матрицы: минимум три переменные рабочего файла данных. Вы можете писать поименно и/или через «to»: *var* to *var*. Имена рекомендуется иметь не длиннее 8 символов (иначе они будут обрезаны на выходе). Рабочий файл *должен* иметь текстовую переменную VARNAME\_, содержащую имена переменных в том порядке, в каком они идут в файле (т.е. матрица диагонального строения). Вот пример симметрической матрицы диагонального строения с единицами на диагонали:



## TRANS

В этой подкоманде можно затребовать стандартизацию элементов столбцов матрицы перед вычислением расстояний между столбцами, т.е. создать нужную исследователю стандартизацию профилей столбцов. Выберите:

NONE – не стандартизовать; тж. по умолчанию или незаданию подкоманды.

CENTER - центрация (приведение средней элементов каждого столбца к 0).

ZSCORE - z-стандартизация (приведение в столбце средней к 0 и дисперсии к 1). Дисперсия вычисляется на df="N" (не "N-1").

MEANSS - стандартизация без центрации (приведение в столбце средней суммы квадратов сырых значений к 1).

RANGE - приведение в столбце ширины диапазона значений к 1.

RESCALE - приведение в столбце диапазона значений к диапазону 0 - 1.

Стандартизация делается перед сравнением каждой пары столбцов заново, т.к. необходимо исключить из стандартизации и сравнения два элемента, принадлежащие диагонали, и два элемента, характеризующие частное взаимоотношение пары столбцов (см. выше).

## MEASURE

Укажите вид вычисляемого расстояния[[1]](#footnote-1):

BLOCK – манхеттенское расстояние (тж. по умолчанию или незаданию подкоманды).

EUCLID – евклидово расстояние.

SEUCLID – квадратное евклидово расстояние (квадрат предыдущего).

### Особые режимы

Макрос не рассчитан на какие-л. особые режимы (расщепление, отбор, взвешивание, временные преобразования).

# MACRO !KO\_CMPRMX1: COMPARING PROXIMITY MATRICES

Version 1 (Dec 2003). Tested on SPSS 11, 11.5, 13, 14.

**MACRO !KO\_CMPRMX2: COMPARING PROXIMITY MATRICES**

Version 1 (May 2005). Tested on SPSS 11, 11.5, 13, 14.

!KO\_cmprmx1 or !KO\_cmprmx2 matr= *'d:\exercise\mx1.sav' 'd:\exercise\mx2.sav' 'd:\exercise\mx3.sav'* /\*Two or more square matrices: quoted paths/names of external files

/cols= *v1* to *v8* /\*Names of variables (matrix columns) of interest (may use “to”), or ALL – take all matrix columns (default)

/triang= BOTH /\*What triangle to compare: lower (LOWER), upper (UPPER) or both at once (BOTH, default)

/elem= ASIS /\*Compare elements: as they are (ASIS,default), ignore sign (ABS) or square elements (SQUARE)

/measure= BLOCK /\* Distance measure to calculate between matrices: EUCLID (euclidean), SEUCLID (square euclidean), BLOCK (manhattan, default)

/norm= YES /\*Normalize computed distances by the overall magnitude of off-diagonal elements of all matrices - YES or NO (default)

/output= YES /\*Produce results output - YES or NO (default).

Minimal specification MATR.

Оба макроса тождественны по синтаксису их задания и по результату, и предназначены для сравнения квадратных матриц близостей и вообще диагонализованных квадратных матриц, т.е. таких где столбцы и ряды представлены одними и теми же «предметами». Корреляционные, дистанционные матрицы – типичные примеры таких матриц. Макрос сравнивает каждые две матрицы (матрицы должны быть одного размера) соответствующими по положению внедиагональными элементами и суммирует разницы в расстояние. Диагональные элементы не сравниваются. Макрос выводит квадратную симметрическую матрицу расстояний между матрицами как новый рабочий файл. Эту матрицу можно затем использовать, например, в кластерном анализе или многомерном шкалировании. Опция нормирования расстояний делает макрос также полезным в эвристическом подборе «лучшего» коэффициента близости.

Предпочтительность одного или другого из двух макросов выявляется только в ситуациях больших массивов данных. !KO\_CMPRMX1 работает медленнее, но бережет оперативную память; !KO\_CMPRMX2 – наоборот. !KO\_CMPRMX1 допускает *максимум 20 входящих матриц[[2]](#footnote-2)*; !KO\_CMPRMX2 избавлен от такого ограничения. !KO\_CMPRMX1 может спасти в случае, когда матрицы настолько большие, что памяти для работы !KO\_CMPRMX2 не хватает[[3]](#footnote-3). Время исполнения обоих макросов зависит также от задания некоторых подкоманд: COLS=ALL и OUTPUT=NO экономнее по времени.

***Подкоманды***

## MATR

Список сравниваемых между собой попарно квадратных матриц, которые должны быть внешними SPSS-файлами данных матричного типа. Имя каждого файла вместе с указанием пути к нему должно быть взято в кавычки или апострофы. Наличие переменной ROWTYPE\_ в матрицах необязательно. Кроме ROWTYPE\_ и VARNAME\_[[4]](#footnote-4) в матрицах могут быть только столбцы и ряды, составляющие тело квадратной матрицы. См. тж. подкоманду COLS.

Если в матрицах есть пропущенные значения (system- или user-missing), макрос заменит их перед сравнением на нуль, чтобы сравнение матриц было осуществимо. Однако корректна ли такая замена и вообще сравнение недозаполненных матриц – этот вопрос в ведении исследователя.

## COLS

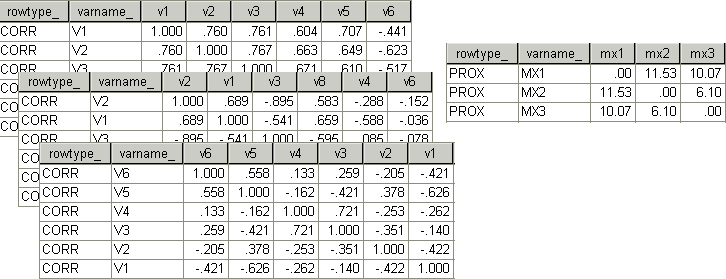
Эта подкоманда указывает, брать ли матрицы целиком или только столбцы/ряды с нужными именами.

ALL – сравнивать матрицы целиком (тж. по умолчанию или незаданию подкоманды). Все матрицы должны быть одного размера. Имена столбцов (т.е. переменных, образующих тело матрицы) могут быть разными в разных матрицах («одноименными» будут считаться столбцы, одинаковые по положению в матрице). Присутствие переменной VARNAME\_ в файлах необязательно.

*Список имен переменных* – в качестве сравниваемых матриц взять квадратные подматрицы, образуемые столбцами и рядами с этими именами. Матрицы MATR не обязаны быть одного размера, но должны включать в себя все перечисленные столбцы, как переменные файла, и одноименные им ряды, как значения переменной VARNAME\_; наличие этой последней в файлах обязательно[[5]](#footnote-5). В разных файлах расположение этих столбцов/рядов может быть неодинаково. Список можно задать поименно и/или диапазоном *var* to *var*. Не указывайте ROWTYPE\_ или VARNAME\_ в списке.

ПРИМЕР 1.

!KO\_cmprmx1 matr= 'd:\exercise\mx1.sav' 'd:\exercise\mx2.sav' 'd:\exercise\mx3.sav' /cols= v1 v2 v3 v4.



* Три входящие матрицы (в данном случае корреляционные, на рис. слева) указаны как внешние файлы. Задано сравнивать их не целиком, а выбранные столбцы/ряды V1, V2, V3, V4. Матрицы содержат все 4 упомянутых столбца/ряда, хотя и в разном порядке. Обязательно наличие VARNAME\_.
* Макрос вычислил манхеттенские расстояния между тремя 4х4 матрицами и вывел их в рабочий файл как матрицу (на рис. справа).

## TRIANG

Какой треугольник матриц сравнивать: нижний (LOWER), верхний (UPPER) или оба вместе (BOTH, т.е. все элементы кроме диагонали). LOWER или UPPER имеет смысл, если матрицы несимметрические. По умолчанию или незаданию подкоманды – BOTH.

## ELEM

Этой подкомандой можно предварительно отменить знак элементов матрицы (ABS) или возвести элементы в квадрат (SQUARE). Например, если матрицы – корреляционные, и знак корреляции для вас не является основанием для различия (но только – величина корреляции), то отмените знак. Или, если вас интересует скорее коэффициент детерминации, нежели собственно корреляции, то возведите в квадрат. Умолчание/незадание подкоманды – то же, что ELEM=ASIS, то есть элементы будут сравниваться такими, какие они есть.

## MEASURE

Укажите вид вычисляемого расстояния:

BLOCK – манхэттенское расстояние (тж. по умолчанию или незаданию подкоманды).

EUCLID – евклидово расстояние.

SEUCLID – квадратное евклидово расстояние (квадрат предыдущего).

## NORM

Если NORM=NO (тж. по умолчанию/незаданию подкоманды), вычисленные попарно между матрицами расстояния дополнительно не преобразуются. При NORM=YES они после вычисления нормируются совокупной величиной внедиагональных элементов[[6]](#footnote-6) всех матриц. Конкретно, при MEASURE=BLOCK расстояние делится на сумму абсолютных величин внедиагональных элементов всех матриц, деленную на число матриц. При MEASURE=SEUCLID расстояние делится на сумму квадратов внедиагональных элементов всех матриц, деленную на число матриц. При MEASURE=EUCLID расстояние делится на корень квадратный из того же, на что делится при SEUCLID.

Нормирование устраняет зависимость результатов – величины вычисленных расстояний – от уровня величины элементов набора матриц, что делает возможным сравнение результатов, полученных на одном наборе матриц, с результатами, полученными на другом наборе матриц, где средняя абсолютная величина элементов может быть иной, - по разным причинам, и прежде всего, например, потому, что один набор матриц был какой-то одной мерой близости, а другой набор – уже другая мера близости. В подобном случае нормирование позволяет выяснить, какая из двух мер сильнее или слабее дифференцирует матрицы (матрицы же могут представлять собой, скажем, разные группы респондентов). После нормирования каждое отдельное расстояние, и следовательно средняя арифметическая всех расстояний, становятся сравнимыми между пусками макроса, несмотря на то что входящие могли быть неродственны.

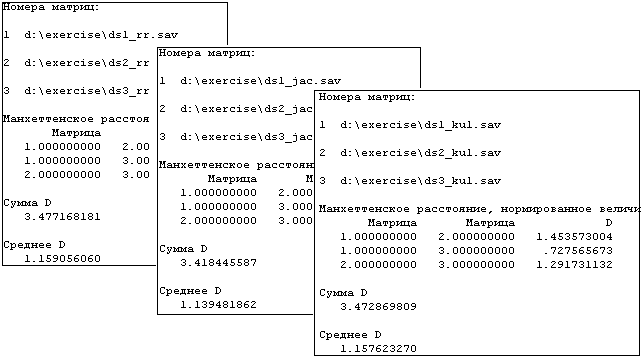
ПРИМЕР 2.

Исследователь решил продемонстрировать различия между клиническими картинами 3-х заболеваний, доказав неодинаковость сочетания симптомов у них с помощью иерархического кластерного анализа симптомов, проделанного, соответственно, 3 раза – на подвыборках, соответветствующих 3-м заболеваниям. Но он колеблется, какую меру близости употребить в кластерном анализе. К его двоичным (симптом присутствует/отсутствует) данным подходят разные меры, а теоретических доводов в пользу одной из них сейчас не нашлось. Тогда исследователь решил практически: выбрать ту меру близости, которая максимизировала бы различия между 3-мя заболеваниями, т.е. между 3-мя матрицами близостей, вводимыми в кластерный анализ.

!KO\_cmprmx1 matr= 'd:\exercise\ds1\_rr.sav' 'd:\exercise\ds2\_rr.sav' 'd:\exercise\ds3\_rr.sav' /norm=YES /output=YES.

!KO\_cmprmx1 matr= 'd:\exercise\ds1\_jac.sav' 'd:\exercise\ds2\_jac.sav' 'd:\exercise\ds3\_jac.sav' /norm=YES /output=YES.

!KO\_cmprmx1 matr= 'd:\exercise\ds1\_kul.sav' 'd:\exercise\ds2\_kul.sav' 'd:\exercise\ds3\_kul.sav' /norm=YES /output=YES.



* В каждом из трех пусков макроса сравнивались три матрицы, соответствующие трем заболеваниям (трем группам пациентов). В 1-м пуске матрицы содержали коэффициент Рассела–Рао, во 2-м пуске матрицы содержали коэффициент Жаккара, а в 3-м – коэффициент Кульчинского. (Все 9 матриц близостей исследователь предварительно получил в SPSS-процедуре Cluster или Proximities и сохранил как внешние файлы. Столбцами и рядами тех матриц были симптомы.)
* Затребовано нормировать расстояния и распечатать результаты.
* Благодаря нормированию результаты трех пусков можно сравнивать несмотря на факт, что три типа коэффициентов, имея разную формулу, слонны в общем к разной величине и изменчивости. Сумма расстояний («Сумма D») свидетельствует, что коэффициенты примерно одинаково дифференцируют клинические картины, но все же Рассел–Рао делает это чуть сильнее. Если бы в трех пусках число входящих матриц было разным, Сумма D была бы непоказательна, но оставалось бы показательно Среднее D.

## OUTPUT

По умолчанию/незаданию подкоманды и при OUTPUT=NO макрос не печатает результатов, а только сохраняет матрицу межматричных расстояний в рабочий файл. Укажите YES, чтобы видеть расстояния, а также их сумму и среднюю, в окне результатов.

### Особые режимы

Макрос не рассчитан на какие-л. особые режимы (расщепление, отбор, взвешивание, временные преобразования).

1. Формулы см. в руководстве по синтаксису, команда Proximities. [↑](#footnote-ref-1)
2. Этого ограничения уже нет в SPSS версии 14 и выше. (Однако в версии 14, из-за дефекта в ней, макрос !KO\_CMPRMX1 вызывает ошибку при повторных пусках.) [↑](#footnote-ref-2)
3. Надо добавить, что вы всегда можете разрешить SPSS занимать одномоментно больше оперативной памяти компьютера, чем разрешено по умолчанию, командой SET WORKSPACE (см. руководство по синтаксису). Используйте команду SHOW WORKSPACE для уяснения, сколько разрешено занимать памяти. [↑](#footnote-ref-3)
4. В SPSS 12 и выше пусть имя этих переменных будут в файле целиком строчными или заглавными буквами [↑](#footnote-ref-4)
5. В SPSS ниже 12-й версии имена рядов в этой переменной пишите целиком заглавными буквами. В 12 и выше версиях имена должны точно совпадать регистром букв с именами переменных-столбцов; и имена переменных – максимум 8 символов. [↑](#footnote-ref-5)
6. Напомним, что если ELEM=SQUARE, элементы матриц возведены в квадрат. [↑](#footnote-ref-6)