***Euclidean space tools***

SPSS macros by Kirill Orlov

kior@comtv.ru, ttnphns@gmail.com

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

All rights reserved

*Euclidean corrections and convertions.* Macros for matrices of proximities that must be layed in euclidean or metric space. You can convert similarities (of a covariance/correlation type or interpretation) geometrically correctly into distances or vice versa; correct similarities or dissimilarities not fully satisfying space to ones satisfying it.

* If you have matrix of “sum of squares and cross-products” – SSCP (or its kind: correlational, covariance or cosin matrix), or matrix of any similarities to which you attach analogous meaning (for example take values as correlations), you can convert it to matrix of euclidean distances with macro [!KO\_SPTOD](#_МАКРОС_!SPTOD:_ПЕРЕВОД_СКАЛЯРНЫХ ПР).
* In order to convert euclidean distances (or dissimilarities interpreted by you as euclidean distances) into SSCP matrix, use [!KO\_DTOSP](#_МАКРОС_!DTOSP:_ПЕРЕВОД_РАССТОЯНИЙ В).
* To test/correct satisfiability of SSCP or other similarity measures matrix for the requirement of space euclidness use [!KO\_SPTOEUC](#_МАКРОС_!SPTOEUC:_ПРОВЕРКА/ПРАВКА_СК).
* To test/correct satisfiability of matrix od distances for the requirement of space euclidness (i.e. whether they are euclidean), use [!KO\_DTOEUC](#_МАКРОС_!DTOEUC:_ПРОВЕРКА/ПРАВКА_РАС).

*Attention*, for all these macros it is recommended that names of input variables (columns of matrix) be no longer than 8 characters, otherwise they will be truncated at output.

# MACRO !KO\_SPTOD: CONVERTION OF SCALAR PRODUCTS INTO EUCLIDEAN DISTANCES

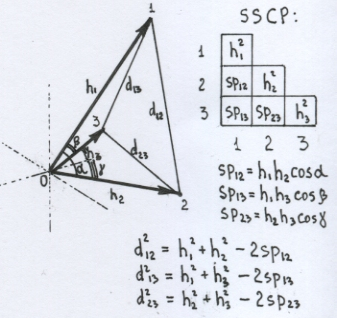
Version 1, July 1999. Tested on SPSS 11.5, 13, 15.

!KO\_sptod cols= *v1 v2 v3* /\*Variables–columns of the input matrix: all or just needed (may write via “to”)

/dist= EUCLID /\*Which distances you want - euclidean (EUCLID, default) or square euclidean (SEUCLID).

Minimal specification COLS.

Матрица сумм квадратов и кросс-произведений (сокращенно SSCP-матрица) играет центральную роль в линейной многомерной статистике. Достаточно сказать, что ковариационная и корреляционная матрицы – это просто частные случаи ее[[1]](#footnote-1). SSCP-матрица по-другому называется матрицей скалярных произведений, SP-матрицей. Это второе ее название передает ее геометрический смысл: если столбцы матрицы данных изобразить в виде точек в пространстве, заданном осями – рядами этой матрицы данных[[2]](#footnote-2), то произведение длин векторов, соединяющих центр системы координат с точками, умноженное на косинус угла между векторами, есть скалярное произведение, *sp*, оно же внедиагональный элемент SSCP-матрицы, а квадрат длины отрезка *h* есть диагональный ее элемент, т.е. в статистико-алгебраических терминах сумма квадратов, SS. См. рисунок:

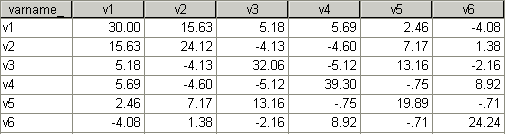


Согласно тригонометрической «теореме косинусов», квадрат евклидового расстояния между точками, например, 1 и 2 равен сумме диагональных элементов h1^2 и h2^2 минус удвоенное скалярное произведение sp12 (см. рисунок). Это то самое евклидово расстояние, которое можно получить между столбцами матрицы данных в меню Distances в SPSS. Таким образом, SSCP-матрицу (в том числе ковариационную или корреляционную матрицу, или матрицу сходств, понимаемую в этом смысле) как матрицу «евклидовых сходств» можно всегда пересчитать *геометрически корректно* в соответствующую ей матрицу евклидовых расстояний. Это и делает данный макрос. Выходящая матрица расстояний выводится в новое окно данных. В окно результатов распечатываются собственные числа входящей матрицы – для справки о том, укладывается ли входящая конфигурация точек в евклидово пространство безошибочно (если все значения неотрицательны, то да).

***Подкоманды***

**COLS**

Укажите все или нужные вам столбцы входящей SSCP-матрицы (матрицы скалярных произведений), поименно и/или ч-з “to”. Матрица должна являться рабочим массивом данных и быть квадратной симметрической. Столбец VARNAME\_ с именами столбцов обязателен. Столбец ROWTYPE\_ не обязателен. На рисунке показан пример входящей матрицы:



**DIST**

Укажите, какие расстояния вы хотите иметь в выходящей матрице: евклидовы (EUCLID, тж. по умочанию/незаданию подкоманды) или квадратные евклидовы (SEUCLID).

***Особые режимы***

Макрос не слушается взвешивания или расщепления данных.

# 

# MACRO !KO\_DTOSP: CONVERTION OF DISTANCES INTO SCALAR PRODUCTS

Version 1, July 1999. Tested on SPSS 11.5, 13, 15.

!KO\_dtosp cols= *v1 to v50* /\*Variables–columns of the input matrix: all or just needed (may write via “to”)

/square= /\*Whether to square the input distances (YES, default), or they are already squared (NO)

/diag= GIVEN /\*Source of diagonal elements for output matrix: define by "double centering" (CENTER, default),

/\*or they are directly specified (GIVEN, they should be at the input matrix).

Minimal specification COLS.

Макрос берет матрицу расстояний (которые трактует как евклидовые) и переводит ее в матрицу скалярных произведений, или SSCP-матрицу, т.е. делает операцию, обратную той что делает макрос !KO\_SPTOD. Диагональные элементы для SSCP-матрицы пользователь может предложить, а может попросить макрос определить их самостоятельно методом так наз. двойной центрации. Выходящая SSCP-матрица выводится в новое рабочее окно данных. В окно результатов распечатываются собственные числа выходящей матрицы – для справки о том, укладывается ли выходящая конфигурация точек в евклидово пространство безошибочно. Если все эти значения неотрицательны, то да. Если же есть отрицательные значения и перевод делался двойной центрацией, то значит входящие расстояния не удовлетворяли свойству евклидовых расстояний: они не «сходились» без зазоров в евклидовом пространстве. Если же есть отрицательные значения и перевод делался с привлечением заданных пользователем диагональных элементов, то причиной неевклидовости могла оказаться как возможная предыдущая причина, так и возможная неадекватность заданных диагональных элементов: «несхождение» *h* в точке O начала координат.

***Подкоманды***

**COLS**

Укажите все или нужные вам столбцы входящей матрицы расстояний, поименно и/или ч-з “to”. Матрица должна являться рабочим массивом данных и быть квадратной симметрической; на диагонали должны быть нули. Столбец VARNAME\_ с именами столбцов обязателен. Столбец ROWTYPE\_ не обязателен.

**SQUARE**

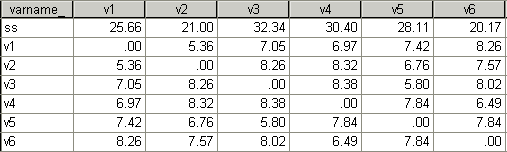
Макрос относится к входящим расстояниям как к неквадратным евклидовым и перед калькуляцией будет возводить их в квадрат. Если ваша матрица содержит уже квадратные евклидовы расстояния (или меры, которые вы трактуете как квадратные евклидовы расстояния), то укажите SQUARE=NO.

**DIAG**

Укажите источник диагональных элементов для будущей SSCP-матрицы:

CENTER (тж по умолчанию/незаданию п/к) Матрица будет переведена в SSCP-матрицу методом «двойной центрации», при котором точка O, начало системы координат, находится как геометрический центр конфигурации точек, расстояния между которыми вы вводите[[3]](#footnote-3). Таким образом, диагональные элементы SSCP-матрицы – h^2, они же SS – определяются автоматически.

GIVEN Диагональные элементы задаются пользователем. В этом случае они должны быть приложены к входящей матрице расстояний как ряд значений, лежащий над матрицей и обозначенный в столбце VARNAME\_ словом “SS” (регистр букв неважен) – см. рисунок:



При DIAG=GIVEN никакие другие столбцы/ряды, кроме ряда SS, не могут иметь имя “SS”. При DIAG=CENTER ряд SS, если приложен к матрице, игнорируется.

***Особые режимы***

Макрос не слушается взвешивания или расщепления данных.

# 

# MACRO !KO\_SPTOEUC: TEST/CORRECTION OF SCALAR PRODUCTS OR SIMILARITIES FOR EUCLIDNESS

Version 2, Sep 2004 (Version 1, Feb 2000). Tested on SPSS 11.5, 13, 15.

!KO\_sptoeuc cols= *v1 v2 v3* /\*Variables–columns of the input matrix: all or just needed (may write via “to”)

/k= *.01* /\*Constant to add (default = mean diag element / 100)

/elong= H /\*What to elongate by a constant: squares of h-vectors (HSQ, default) or h-vectors (H)

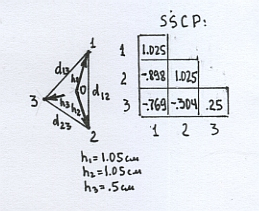
/mxiter= *500* /\*Maximum number of iterations (default=500)

/return= /\*Optional: bring back diagonal to its original sum (YES) or not (NO, default).

Minimum specification COLS.

Макрос берет матрицу сходств, воспринимая ее как SSCP-матрицу (на диагонали должны стоять положительные числа; если они все равны 1, следовательно макрос воспринимет матрицу как корреляционную), и проверяет, укладывается ли она в евклидовое пространство, и, если нет, пытается довести ее до евклидовости. Он выдает подправленную матрицу в новое окно данных; в результаты распечатываются собственные числа, говорящие об успехе или неуспехе.

Матрица скалярных произведений, она же SSCP-матрица, может иметь одно или несколько отрицательных собственных чисел. Это признак того, что облако данных (точек), связи между которыми она показывает, не укладывается в евклидово пространство безошибочно[[4]](#footnote-4). Часто причиной этого бывает несоразмерность длин векторов *h*, идущих из точки O (центр системы координат) к точкам конфигурации. Другими словами, несколько *h* неспособны сойтись в одной точке O. (Напомним, что h^2 это диагональный элемент SSCP матрицы – см. выше в описании макроса !KO\_SPTOD.) Например, на рисунке ниже h3 слишком короток, чтобы связать O с точкой 3. Нужно, чтобы он был минимум примерно вдвое длиннее или чтобы h1 и h2 были длиннее, чтобы им троим сойтись в точке O.



Макрос как раз и занимается удлинением *h*, т.е. увеличением диагональных элементов матрицы повторительным прибавлением константы, пока матрица не лишится отрицательных собственных чисел и не станет, следовательно, геометрически евклидовой. Расстояния (*d*) между точками при этом удерживаются неизменными, следовательно, удлинение векторов *h* должно сопровождаться уменьшением угла между ними около точки O: поэтому константа прибавляется не только к диагональным, а ко всем элементам матрицы скалярных произведений. Таким путем, данный макрос может исправить «подпорченную» матрицу SSCP (в том числе матрицу ковариаций или корреляций) или же сделать геометрически евклидовой матрицу заведомо неевклидовых сходств, чтобы вы могли ввести ее, скажем, в классический факторный анализ (который требует геометрической евклидовости).

***Подкоманды***

**COLS**

Укажите все или нужные вам столбцы входящей SSCP-матрицы, поименно и/или ч-з “to”. Матрица должна являться рабочим массивом данных и быть квадратной симметрической. Столбец VARNAME\_ с именами столбцов обязателен. Столбец ROWTYPE\_ не обязателен.

**K**

Константа для прибавления, укажите положительное число. По умолчанию/незаданию подкоманды берется как средний в матрице диагональный эл-т / 100. Если итог правки успешен, но последнее собственное число слишком близко к 0 (т.е. матрица близка к вырожденности), что вам может не понравиться, увеличьте K.

**ELONG**

Что удлинять на константу, или к чему прибавлять константу. Укажите одно из двух.

HSQ (тж по умолчанию/незаданию) На константу увеличивается длина вектора *h*, взятая в квадрат; т.е. короткие *h* будут удлиняться быстрее, чем длинные h. Алгебраически это есть прибавление константы к элементу SSCP-матрицы по формуле *sp+k*.

H На константу увеличивается длина вектора *h*; т.е. все *h* будут расти с одной скоростью, равномерно. Алгебраически это есть прибавление константы к элементу SSCP-матрицы по формуле *sp+k\*(h1+h2+k)*. Этот вариант в некоторых случаях неэффективен, когда предыдущий эффективен.

Если евклидовость не достигается (т.е. отрицательные собственные числа сохраняются) даже при ELONG=HSQ и большом числе итераций (или большом K), причиной неевклидовости являются не *h*, а сами расстояния *d* между точками. В таком случае следует: (1) перевести SSCP-матрицу в матрицу расстояний макросом !KO\_SPTOD, (2) которую довести до евклидовости макросом !KO\_DTOEUC и (3) обратно перевести в SSCP-матрицу макросом !KO\_DTOSP с заданием ей тех диагональных элементов, что были изначально. После чего следует матрицу снова проверить макросом !KO\_SPTOEUC.

**MXITER**

Предельное число итераций. По умолчанию подкоманды = 500.

**RETURN**

Вы можете затребовать (RETURN=YES) по окончании подгонки матрицы к евклидовости восстановить диагональным элементам их начальную сумму (если же на диагонали исходно стояло одно число, как напр. в корреляционной матрице, оно вернется на диагональ). Внедиагональные элементы при этом пересчитаются: изменятся в одинаковое число раз, что значит что евклидовые расстояния *d* между точками конфигурации тоже изменятся по сравнению с исходными в одно и то же число раз.

***Особые режимы***

Макрос не слушается взвешивания или расщепления данных.

# MACRO !KO\_DTOEUC: TEST/CORRECTION OF DISTANCES FOR EUCLIDNESS

Version 1, Oct 1999. Tested on SPSS 11.5, 13, 15.

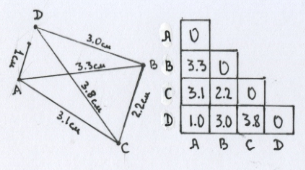
!KO\_dtoeuc cols= *v1 to v10* /\*Variables–columns of the input matrix: all or just needed (may write via “to”)

/k= /\*Constant to add to nondiagonal distances (default = mean distance / 100)

/mxiter= *500* /\*Maximum number of iterations (default=500).

Minimum specification COLS.

Расстояния между точками облака могут не сходиться в евклидовом пространстве, что делает эти расстояния не евклидовыми, не удовлетворяющими евклидовости. Например, на рисунке ниже расстояние AD в 1 см слишком коротко, чтобы состыковаться. Нужна длина в 1.5 см, чтобы отрезок достиг точки D. Или больше чем 1.5 см (тогда точка D окажется приподнята отрезком AD над плоскостью, вращаясь вокруг оси BC).



Если имеется матрица расстояний между точками, то признаком подобной неевклидовости конфигурации будет наличие отрицательных собственных чисел у матрицы-двойного центрата матрицы расстояний. (Матрица-двойной центрат это такая полученная из матрицы расстояний матрица скалярных произведений, начало системы координат которой (точка O) находится в геометрическом центре конфигурации точек.) Макрос повторительно прибавляет к элементам входящей матрицы расстояний константу до тех пор, пока матрица-двойной центрат не лишится отрицательных собственных чисел[[5]](#footnote-5), т.е. расстояния не начнут удовлетворять свойству евклидовости. Таким путем вы можете переправить «подпорченные» евклидовы расстояния или заведомо неевклидовые расстояния в геометрически корректные евклидовые, чтобы использовать их в методах анализа, требующих таковых (к примеру, некоторые методы иерархической кластеризации).

Макрос относится к элементам входящей матрицы как к *неквадратным* расстояниям; поэтому если ваши элементы это расстояния, возведенные в квадрат, то извлеките из них кв. корень, прежде чем вводить матрицу в макрос. Макрос выдает подправленную матрицу расстояний (неквадратных) в новое окно данных; в результаты распечатываются собственные числа матрицы-двойного центрата.

***Подкоманды***

**COLS**

Укажите все или нужные вам столбцы входящей матрицы расстояний, поименно и/или ч-з “to”. Матрица должна являться рабочим массивом данных и быть квадратной симметрической; на диагонали должны быть нули. Столбец VARNAME\_ с именами столбцов обязателен. Столбец ROWTYPE\_ не обязателен.

**K**

Константа для прибавления к расстояниям, укажите положительное число. По умолчанию/незаданию подкоманды берется как среднее в матрице расстояние / 100.

**MXITER**

Предельное число итераций. По умолчанию подкоманды = 500.

***Особые режимы***

Макрос не слушается взвешивания или расщепления данных.

1. Если данные (их столбцы) центрованы, то полученная из них SSCP матрица, деленная на N-1, называется ковариационной; а если они стандартизованы – то корреляционной. Поскольку делением на константу можно пренебречь (или, в конце концов, вы можете всегда умножить ковариационную или корреляционную матрицу на N-1), то можно сказать, что ковариационная и корреляционная матрица это SSCP матрица, полученная из, соответственно, центрованных и стандартизованных данных. [↑](#footnote-ref-1)
2. Это пространство, «перевернутое с ног на голову» (потому что точки в нем – переменные, столбцы данных, а оси – индивиды, ряды данных, а не как привычно – наоборот), называется «пространством индивидов». [↑](#footnote-ref-2)
3. Это привносит одно нулевое собственное число в набор собственных чисел SSCP-матрицы. [↑](#footnote-ref-3)
4. Вообще-то, если отрицательные собственные числа очень малы по абсолютной величине, на них не обращают внимания как на нарушение евклидовости, даже если их несколько; отрицательность таких очень малых значений может оказаться следствием просто калькуляторных округлений при вычислении собственных чисел, а не следствием неевклидовости. [↑](#footnote-ref-4)
5. Минимум одно нулевое собственное число всегда присутствует в матрице-двойном центрате, оно происходит от того что точка O, будучи центром облака, принадлежит пространству, задаваемому точками облака. [↑](#footnote-ref-5)