Ciencias Matemáticas, Vol. 32, No. 2, Pag. 125-129, 2018
Recibido 02-2018

Tabla de Evaluación de Stress para Escalamiento Multidimensional
Stress Evaluation Table for Metric Multidimensional Scaling

Lic. Javier Alejandro Quintero Roba [1], Lic. María Esther Reyes Calzado [2*], Dra. Elina Miret Barroso [3], Alejandro Javier Quintero Roba [4]

Resumen El Escalamiento Multidimensional es una técnica de exploración multivariada que permite, mediante la minimización de una función de pérdida STRESS, la reducción de la dimensión de los datos para la búsqueda de patrones en la estructura de los mismos. Para medir la bondad de ajuste de la representación a partir de su valor de STRESS se han planteado numerosos criterios empíricos, teniendo en cuenta que, cuando se estudia un número grande de objetos, los valores de STRESS tienden a crecer junto a la dimensión del problema. En esta investigación se presenta una tabla probabilística de evaluación de STRESS con la intención de contar en la práctica de cotas superiores para dicha magnitud. Esta tabla se obtuvo a partir generación aleatoria de matrices de disimilitud y su procesamiento por varios métodos de Escalamiento Multidimensional.

Abstract Multidimensional Scaling is a multivariate exploration technique that allows to reduce dimension on data by minimizing a loss function called STRESS so that we can find patterns on the data structure. To measure the goodness of the adjusted representation from its STRESS value it has been studied a number of empirical criteria, knowing that STRESS tends to increase along with the dimension of the problem when the number of objects is large. This investigation presents a probabilistic table for STRESS evaluation, obtained from study of random dissimilarity matrices, intended to find heights for that magnitude.

Palabras Clave
Escalamiento Multidimensional — STRESS — Tabla de Evaluación de Stress

[1] *Departamento de Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba, jjquinte5693@gmail.com*
[3] *Departamento de Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba, m.reyes@matcom.uh.cu*
[2] *Departamento de Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba, elina@matcom.uh.cu*
[4] *Estudiante de Licenciatura en Matemática, Facultad de Matemática y Computación, Universidad de la Habana, Cuba*
[*] **Autor para Correspondencia**

Introducción

El Escalamiento Multidimensional (Multidimensional Scaling, MDS por sus siglas en inglés) es una técnica exploratoria para el análisis de similitudes y disimilitudes entre objetos; en general, medidas de relación entre varios individuos de estudio. La técnica intenta modelar los datos como distancias entre puntos en un espacio geométrico, usualmente que sea reconocible de manera visual.

Para realizar MDS es necesario un algoritmo computacional que a partir de los datos iniciales sea capaz de encontrar una representación fiel a la realidad. Para medir la bondad de ajuste de una solución se utiliza la función STRESS o alguna de sus variantes.

En la evaluación de la calidad de la representación del MDS a partir de su valor de STRESS se han planteado numerosos criterios empíricos teniendo en cuenta que, cuando se estudia un número grande de objetos, los valores de STRESS tienden a crecer junto a la dimensión del problema.

En esta investigación se presentan tablas probabilísticas de evaluación de STRESS con la intención de contar en la práctica de cotas superiores para dicha magnitud. Esta tabla se obtuvo a partir generación aleatoria de matrices de disimilitud y su procesamiento mediante métodos de Escalamiento Métrico y No Métrico.

1. Perspectiva de Escalamiento Multidimensional

En el MDS se representan medidas de similitud o disimilitud entre pares de objetos (individuos), como distancias entre espacios de baja dimensión, usualmente 2 ó 3. La representación gráfica que provee el Escalamiento Multidimensional permite literalmente explorar su estructura visualmente, lo que revela regularidades que permanecen ocultas cuando se estudian los números indistintamente. [9]

En dependencia de la transformación utilizada para el escalamiento de las proximidades iniciales, los modelos de MDS se clasifican en Métrico y No Métrico. El modelo más utilizado en la literatura es el no métrico u ordinal que se basa en la premisa de que las disparidades están en escala ordinal, lo que significa que para la construcción de las distancias finales solo importa el orden inicial de las proximidades y no se tiene en cuenta su valor. El modelo no métrico posee varias ventajas respecto al métrico, ya que los valores de bondad de ajuste son mejores que su contraparte, también el proceso de optimización es más flexible, lo que permite mover con mayor libertad las disparidades. Las desventajas de este modelo radican en la aparición de soluciones degeneradas y que no se tiene en cuenta como información relevante las nociones de cercanía o lejanía. [5] [6]

Se trabaja con MDS Métrico cuando la función de disparidades es continua, paramétrica y monótona. En este caso se desea conservar, además de los rangos ordinales entre las proximidades, la noción de magnitud entre ellas. Estos modelos poseen la ventaja de que a partir de las características de la función de disparidad es posible buscar propiedades matemáticas deseables para los algoritmos. Además, el modelo métrico evita la degeneración al imponer una estructura menos flexible que el modelo ordinal. Su desventaja radica en la bondad de ajuste, pues la libertad de movimiento de las disparidades es controlada por una función continua y para lograr una buena representación se requiere datos bastante precisos. Por lo anterior, los valores de STRESS en el caso del modelo métrico serán generalmente mayores, aun cuando las representaciones sean aproximadamente equivalentes. [3] [4]

A partir del planteamiento matemático del problema del MDS, los algoritmos pueden dividirse en dos ramas: técnicas algebraicas y algoritmos iterativos. Dentro de las primeras se encuentra el Escalamiento Clásico, cuya función de ajuste es el STRAIN y su optimización se basa en la teoría de la descomposición espectral. [8] [5] Dentro de los iterativos están, por ejemplo: los cuasi-Newton, máximo descenso, ALSCAL, SMACOF, entre otros; así también como métodos heurísticos aproximados.

Los algoritmos iterativos son un poco más flexibles que el MDS Clásico, pues permiten el re-escalamiento óptimo de los datos. Usualmente estos algoritmos parten de una configuración inicial, mueven los puntos para reducir el STRESS, y posteriormente reescalan las disimilitudes iniciales de forma óptima para generar nuevas disparidades, dentro de los límites de los datos. Este proceso de modificar la configuración del MDS (manteniendo fijas las disparidades), y re-escalar las disparidades (manteniendo fijas las distancias), se repite hasta lograr convergencia.[3] [4]

Sin embargo, no en todos los casos los resultados han sido del todo satisfactorios especialmente cuando la dimensión 1 (i.e cantidad de individuos) es alta. Además, la implementación de una herramienta a partir del material disponible resulta una tarea que en ocasiones se hace imposible debido a que las explicaciones en la literatura no son del todo claras y resultan insuficientes.[11]

1.1 STRESS y calidad de la representación

Para medir la bondad de ajuste de una solución se utiliza la función STRESS o alguna de sus variantes. Sea $X_{(n\times p)}$ una configuración en $\mathbb{R}^p$, la expresión de X como solución del MDS en p dimensiones se halla generalmente minimizando la función de pérdida $Stress-1$ ó de Kruskal, dada por:

$$Stress-1=\sqrt{\frac{\sum_{i<j}(d_{ij}-\widehat{\delta}_{ij})^2}{\sum_{i<j}d_{ij}^2}}$$

donde los d_{ij} son las nuevas distancias euclídeas, obtenidas de las nuevas coordenadas para los puntos de Ω.[8] [6]

Esta función toma valores en el intervalo $[0,1]$ y expresa si la representación obtenida refleja correctamente las relaciones de distancias originales. Otras variantes de la función de *Stress* son: el *Stress-normado* y el *S-Stress*.[3] [4]

Para evaluar la calidad de la representación del MDS a partir del STRESS existen varias opiniones. Esta investigación sigue dos líneas de esta teoría: los criterios empíricos de Kruskal de 1964, y el de bondad de ajuste bajo selección aleatorizada planteado por Borg, Groenen, & Mair en 2013. [10] [4]

El criterio empírico de Kruskal se basa en la experiencia del investigador, y establece las siguientes clasificaciones para una representación de puntos:

Stress	Ajuste
0.20	Pobre
0.10	Justo
0.05	Bueno
0.025	Excelente
0	"Perfecto"

Borg, Groenen, & Mair indican que una solución de MDS perfecta es aquella que tiene STRESS = 0, pues en este caso las distancias de la configuración representan los datos de manera precisa (en el sentido deseado). Esto conlleva a la pregunta de cuándo un valor de STRESS es suficientemente bueno.

Primeramente ha de analizarse que el algoritmo utilizado debe ser capaz de reconocer patrones ocultos en los datos, i.e. captar la topología de estos, reconocer que los datos poseen cierta estructura, y que no son solo un conjunto de datos aleatorios. Según los autores, la evaluación de un valor de STRESS específico es una cuestión compleja, que involucra un gran número de parámetros y consideraciones.

En general, un valor de STRESS es suficientemente bueno si es menor que el valor esperado de STRESS a partir de aleatorización de las matrices de disimilitud. Si esto no se cumple, es imposible interpretar significativamente en ningún sentido las distancias que se obtienen mediante el algoritmo, ya que estas no están realmente relacionadas con los datos. [12]

En [12] se establecen límites de aleatorización del STRESS, o sea el límite que indica que un valor por debajo de él tiene probabilidad 1 % de ser obtenido a partir de configuraciones aleatorias, o sea el percentil de la distribución empírica bajo simulaciones.

Otras medidas, como el STRESS por punto, y los diagramas de Shepard se reportan que pueden ser útiles. No obstante, en la práctica, los valores de STRESS obtenidos con datos reales son menores que los que provienen de selección aleatoria, esto se debe a la estructura propia de dichos datos.

2. Evaluación del STRESS para matrices aleatorias

La evaluación del STRESS es costosa cuando el número de individuos es elevado, pero cuando la dimensión crece es importante obtener indicadores de la calidad de la representación.

Teniendo en cuenta que, una matriz de disimilitudes que representa datos reales contiene una estructura subyacente al escalamiento, pero una matriz de disimilitudes generada aleatoriamente carece de estructura predefinida, es acertado pensar que los valores de STRESS que se obtienen de escalar matrices aleatorias son mayores que aquellos que se obtienen de matrices reales. Por lo cual, los valores de ajuste obtenidos del escalamiento de matrices aleatorias proveen una cota superior en la práctica para la función de STRESS.

En 1969, Klahr muestra que para 8 objetos los valores de STRESS obtenidos del escalamiento de matrices aleatorias eran superiores a los de matrices con estructura real. Posteriormente, en 1973, Spence and Ogilvie reportan experimentación con matrices aleatorias de 12-48 objetos y de 1-5 dimensiones.

En 1978, Levine produce una tabla que refleja la probabilidad de encontrar un valor específico de STRESS para una configuración dada. Esta tabla muestra valores de STRESS de 1–5 dimensiones para 6, 8, 10, 12, 16, 20 y 24 objetos. Esta idea fue retomada por Sturrock & Rocha, en 2000, quienes ofrecen una tabla de valores aleatorios de STRESS para 5-100 objetos escalados por métodos no métricos.[12]

Borg & Groenen plantean la necesidad de obtener el valor esperado bajo selección aleatoria de la función de STRESS como posible cota superior. Sin embargo, no se cuenta en la literatura con cifras de este tipo para el modelo métrico. [4]

El objetivo de esta investigación es elaborar una tabla que recopile los percentiles de nivel 1 de una muestra de valores de STRESS obtenidos a partir del escalamiento de matrices de disimilitud generadas aleatoriamente para contar en la práctica con cotas superiores de bondad de ajuste.

Para ello se generaron matrices de disimilitud de 5-25 objetos a partir de una distribución uniforme [0,1] (800 matrices en cada caso) que fueron escaladas mediante MDS No métrico con Evolución Diferencial y MDS Métrico con CMA-ES absoluto.

La selección de estos algoritmos se basa en trabajos previos de los autores donde, al aplicar Escalamiento Multidimensional empleando diferentes Metaheurísticas a ejemplos de la literatura y reales, se obtuvieron los mejores resultados para los métodos antes mencionados. [11] [10]

2.1 Evolución Diferencial en MDS No Métrico

Los algoritmos evolutivos son métodos de búsqueda dirigida basada en probabilidad. Estos algoritmos establecen una analogía entre el conjunto de soluciones de un problema y el conjunto de individuos de una población natural, codificando la información de cada solución en un string a modo de cromosoma. A tal efecto se introduce una función de evaluación de los cromosomas, que llamaremos calidad ("fitness") y que está basada en la función objetivo del problema. Igualmente se introduce un mecanismo de selección de manera que los cromosomas con mejor evaluación sean escogidos para "reproducirse" más a menudo que los que la tienen peor.[13]

La Evolución Diferencial se caracteriza por el uso de vectores de prueba, los cuales compiten con los individuos de la población actual a fin de sobrevivir. El algoritmo asume que las variables del problema a optimizar están codificadas como un vector de números reales y que el dominio de las variables del problema está restringido por ciertas cotas definidas para cada variable. Dado que se opera con una población en cada iteración, se espera que el método converja de modo que al final del proceso la población sea muy similar, y en el infinito se reduzca a un solo individuo. [7]

2.2 CMA-ES en MDS Métrico

La estrategia evolutiva de adaptación de la matriz de covarianzas o Covariance Matrix Adaptation Evolution Estrategy (CMA-ES, por sus siglas en inglés) es uno de los algoritmos de optimización continua más exitoso de los últimos años. Este algoritmo controla mediante la adaptación de la matriz de covarianzas los pasos individuales en cada dirección y las relaciones entre las coordenadas.[2]

Mediante una distribución normal se generan las mutaciones, creando la nueva población. CMA-ES adapta la matriz de covarianzas de la distribución normal multivariante de mutaciones, captando las relaciones de dependencia entre las variables, ya que la matriz de covarianzas define la dependencia por pares entre las variables de la distribución.

CMA-ES está especialmente orientada para el escenario provisto de problemas "black-box" y puede sobrellevar problemas de alta dimensionalidad de forma rápida. Está diseñada para tomar ventaja de espacios escarpados, problemas no separables, no linealidad, no suavidad y multimodalidad del STRESS. El STRESS es considerado una función de "black-box", pues su dominio no se conoce explícitamente, pero el valor de cada representación puede calcularse, siendo esta la única información disponible.[1]

3. Resultados

Se generaron 20000 matrices de disimilitud aleatoriamente (distribuidas como se describe anteriormente), y se procesaron por MDS Métrico con CMA-ES Absoluto (disimilitudes no escaladas) y MDS No Métrico con Evolución Diferencial.

El procesamiento de las muestras permitió la elaboración de tablas para cada algoritmo [Figuras 1 y 3] que representan los valores Mínimos y Máximos, Medias y percentiles de nivel 1 para cada cantidad de objetos estudiada (5-25).

No. de Objetos	Stress Mínimo	Media	Stress Máximo	Percentil Nivel 1
5	0,0748	0,2777	0,5628	0,092
6	0,1057	0,3113	0,5620	0,117
7	0,1615	0,3363	0,5218	0,182
8	0,2247	0,3586	0,5306	0,234
9	0,2246	0,3733	0,5340	0,259
10	0,2540	0,3934	0,5378	0,278
11	0,2998	0,4104	0,5461	0,308
12	0,3217	0,4274	0,5524	0,329
13	0,3298	0,4425	0,5637	0,345
14	0,3535	0,4554	0,5674	0,357
15	0,3623	0,4687	0,5669	0,387
16	0,3838	0,4815	0,5908	0,399
17	0,3917	0,4902	0,5651	0,406
18	0,4155	0,5008	0,5816	0,426
19	0,4308	0,5128	0,5898	0,440
20	0,4243	0,5224	0,6038	0,461
21	0,4487	0,5303	0,6238	0,460
22	0,4470	0,5365	0,6129	0,460
23	0,4514	0,5440	0,6183	0,476
24	0,4849	0,5506	0,6221	0,490
25	0,4879	0,5579	0,6269	0,496

Figura 1. Tabla Stress. MDS con CMAES

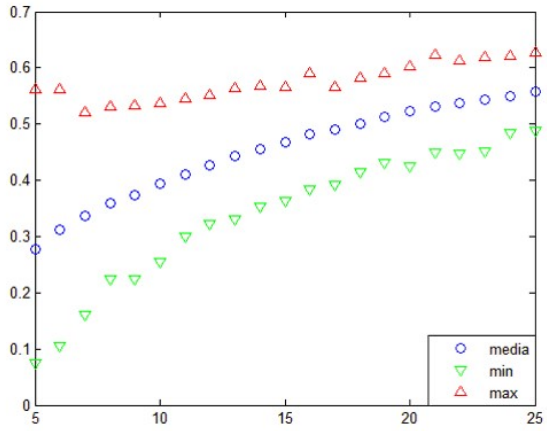

Figura 2. Curva de crecimiento CMAES

El percentil de nivel 1 aporta una cota superior de los valores esperados de STRESS para un número específico de objetos cuando se trabaje con datos reales.

Como era esperado, los valores de STRESS crecen al crecer el número de objetos. En las Figuras 2 y 4 puede apreciarse el comportamiento de los valores de STRESS al crecer la dimensión.

Los resultados obtenidos para los dos métodos empleados coinciden con los reportados por Sturrock & Rocha (2000).

No. de Objetos	Stress Mínimo	Media	Stress Máximo	Percentil Nivel 1
5	0.0351	0.2428	0.8371	0.066
6	0.0661	0.2746	0.6281	0.110
7	0.1252	0.3117	0.5659	0.155
8	0.1571	0.3317	0.5657	0.199
9	0.1874	0.3503	0.5493	0.224
10	0.2352	0.3692	0.5848	0.259
11	0.2568	0.3844	0.5627	0.281
12	0.2855	0.4012	0.5533	0.313
13	0.2972	0.4085	0.5276	0.325
14	0.3216	0.4220	0.5536	0.339
15	0.3218	0.4318	0.5494	0.354
16	0.3562	0.4421	0.5563	0.368
17	0.3695	0.4533	0.5510	0.385
18	0.3680	0.4606	0.5482	0.395
19	0.3842	0.4679	0.5688	0.400
20	0.3788	0.4742	0.5742	0.414
21	0.3817	0.4813	0.5833	0.425
22	0.4048	0.4873	0.5670	0.434
23	0.4269	0.4933	0.5701	0.441
24	0.4190	0.4993	0.5701	0.449
25	0.4338	0.5058	0.5824	0.455

Figura 3. Tabla Stress. MDS con Evolución Diferencial

Conclusiones

Se generaron 20000 matrices de disimilitud aleatoriamente, y se procesaron por MDS No Métrico con Evolución Diferencial y MDS Métrico con CMA-ES absoluto. El procesamiento de las muestras permitió la elaboración de una tabla que representa los percentiles de nivel 1. La tabla elaborada ofrece cotas superiores de calidad para la evaluación del STRESS en estos métodos. Las cifras obtenidas sirven como herramienta al investigador para juzgar a partir del valor de STRESS de su representación si el método empleado reconoce la estructura original de los datos. Se trabaja en el incremento del número de objetos y la validación de los resultados mediante su comparación con ejemplos reales.

Referencias

[1] A Auger and N. Hansen. Tutorial cma-es: evolution strategies and covariance matrix adaptation. *In GECCO (Companion).*, 2012.

[2] A. Bolufé Röhler. Búsqueda de población mínima–un algoritmo de optimización escalable para problemas multimodales. *Tesis de Doctorado.*, 2015.

[3] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling. Theory and Applications.* Springer., 2005.

[4] Ingwer Borg, Patrick J. F. Groenen, and Patrick Mair. *Applied multidimensional scaling.* Springer., 2013.

[5] Trevor F. Cox and Michael A. A. Cox. *Multidimensional Scaling.* Chapman & Hall., 1994.

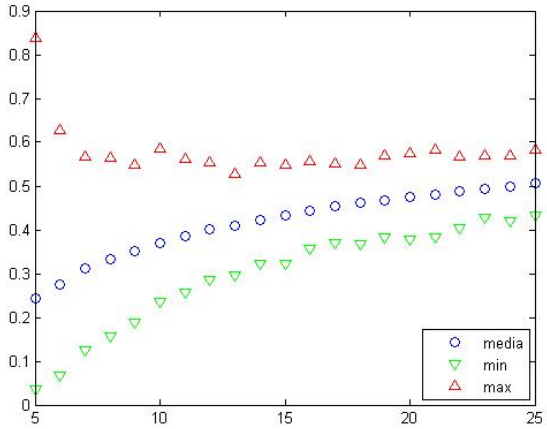

Figura 4. Curva de crecimiento Evolución Diferencial

[6] Carles M. Cuadras. *Nuevos Métodos de Análisis Multivariante.* CMC Editions., 2008.

[7] Stefan Etschberger and Andreas Hilbert. Multidimensional scaling and genetic algorithms : A solution approach to avoid local minima. *Arbeitspapiere zur mathematischen Wirtschaftsforschung.*, No. 181, 2003.

[8] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.*, volume Tenth printing.1995. Academic Press Inc., 1979.

[9] Elina Miret. Un enfoque unificado para técnicas de representación euclidiana. *Tesis de Doctorado.*, 2005.

[10] Javier A. Quintero. Algoritmo de escalamiento multidimensional métrico con el empleo de la estrategia evolutiva de adaptación de la matriz de covarianzas. *Tesis de Licenciatura.*, 2016.

[11] María E. Reyes. Escalamiento multidimensional empleando metaheurísticas. *Tesis de Licenciatura.*, 2015.

[12] Kenneth Sturrock and Jorge Rocha. A multidimensional scaling stress evaluation table. *Field Methods*, 2000.

[13] El-Ghazali Talbi. *Metaheuristics From Design to Implementation.* John Wiley & Sons., 2009.