

La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 7

Escalado óptimo

Primera edición: julio 2008

ISBN: 978-84-96515-71-0

Traducción: Jordi Comas Angelet
Revisión: Carles M. Cuadras Avellana

© **Michael Greenacre, 2008**
© **de la edición en español, Fundación BBVA, 2008**

www.fbbva.es

Escalado óptimo

Hasta ahora hemos presentado el AC como un método geométrico de análisis de datos. Hemos destacado tres conceptos fundamentales: perfil, masa y distancia χ^2 , y cuatro conceptos derivados: centroide (media ponderada), inercia, subespacio y proyección. Los perfiles son puntos multidimensionales, ponderados por masas. Medimos las distancias entre perfiles mediante distancias χ^2 . Visualizamos los perfiles proyectándolos sobre el subespacio de pocas dimensiones que mejor se ajusta a los perfiles. A continuación, para su interpretación, proyectamos los vértices como puntos de referencia en dicho subespacio. De todas formas, existen muchas maneras distintas de definir y de interpretar el AC. Por ello, la misma metodología de base se ha redescubierto muchas veces en diferentes contextos. Una de estas metodologías alternativas es el *escalado óptimo*. Una discusión sobre esta aproximación nos permitirá profundizar en el conocimiento del AC.

Contenido

Cuantificación de un conjunto de categorías	76
Cálculo de la media global utilizando una escala entera	76
Cálculo de la media de los grupos de edad mediante una escala entera	77
Cálculo de la varianza utilizando la escala entera	77
Cálculo de las puntuaciones en una escala desconocida	77
La maximización de la varianza proporciona la escala óptima	78
Los valores de la escala óptima de la dimensión del AC que mejor se ajusta	78
Interpretación de la escala óptima	79
Condiciones de identificación de una escala óptima	80
Cualquier transformación lineal de la escala sigue dando una escala óptima	80
La escala óptima no es única	81
Un criterio basado en las distancias entre las filas y las columnas	82
RESUMEN: Escalado óptimo	83

Cuantificación de un conjunto de categorías

Consideremos una vez más el ejemplo de la tabla de la imagen 6.1, la tabla de contingencia que cruza los grupos de edad con las categorías sobre la autopercepción de la salud. Tanto las variables fila como las variables columna son variables categóricas, que hemos guardado en un archivo de datos utilizando códigos de 1 a 7 para la edad, y de 1 a 5 para la salud. Si queremos calcular estadísticos como la media y la varianza o hacer un análisis de regresión en el que intervenga, por ejemplo, la variable autopercepción de la salud, necesitamos valores numéricos para cada categoría de salud. Si utilizamos los valores de 1 a 5, estamos asumiendo de forma implícita que la separación entre estas categorías es exactamente la unidad, lo que no tiene porqué ser cierto. El hecho de que hayamos ordenado las categorías de salud (la autopercepción de la salud es una variable categórica ordinal), justifica, en parte, que hayamos utilizado los valores de 1 a 5, pero, ¿que ocurriría si la variable fuera nominal, como, por ejemplo, la variable país que vimos en el capítulo 1 (imagen 1.3)? ¿Y si fuera el estado civil? La variable grupo de edad es también una variable ordinal establecida mediante intervalos en la escala original de la edad, de manera que podemos razonablemente utilizar los puntos medios de cada intervalo de la edad como valores de escala. Sin embargo, no es claro el valor que hemos asignado al grupo de edad 7, que hemos dejado abierto (de 75 o más años). Cuando no haya una alternativa mejor, y las categorías presenten una ordenación natural, como en nuestro caso, en los cálculos utilizaremos por defecto valores enteros (en nuestro caso de 1 a 7 y de 1 a 5), que denominan una *escala entera*. Vamos a ver cómo el escalado óptimo ofrece un camino, resultante de un determinado criterio de optimización, que nos permite asignar valores numéricos a una variable categórica.

Cálculo de la media global utilizando una escala entera

Vamos a utilizar la escala entera para efectuar algunos cálculos simples. Sin embargo, primero invertiremos la codificación de las categorías de salud, de manera que el mayor valor corresponda a la mejor salud; así, 5 indicará *muy buena* salud, descendiendo hasta 1, que indicará *muy mala* salud. En la encuesta constituida por 6371 individuos, hay 817 individuos con *muy buena* salud (código 5), 3542 con *buena* salud (código 4), y así sucesivamente. Utilizando estos códigos enteros como escala para las categorías de salud, podemos calcular la *salud media* de la siguiente manera:

$$[(817 \times 5) + (3542 \times 4) + \dots + (103 \times 1)] / 6371 = 3,72$$

Es decir,

$$(0,128 \times 5) + (0,556 \times 4) + \dots + (0,016 \times 1) = 3,72 \quad (7.1)$$

donde $817/6371 = 0,128$, $3542/6371 = 0,556$, etc. son los elementos del perfil fila medio (última fila de la tabla de la imagen 6.2). Por tanto, esta media de todos

* En mi experiencia como consultor en estadística, una vez me mostraron una encuesta con la variable «afiliación religiosa» tomando los valores: 0 = ninguna, 1 = católica, 2 = protestante, etc. ¡El investigador calculó la religión media de la muestra!

los encuestados no es más que el centroide de los valores de la escala obtenido ponderando con los elementos del perfil fila medio.

Consideremos ahora un determinado grupo de edad, por ejemplo, el de 16 a 24 años. En la primera fila de datos de la tabla de la imagen 6.1, vemos que de los 1223 encuestados de este grupo, hay 243 individuos con *muy buena* salud, 789 con *buena*, y así sucesivamente. Utilizando otra vez los valores enteros de la escala de 5 a 1 para las categorías de la salud, la *salud media* de este grupo de 16-24 es:

$$[(243 \times 5) + (789 \times 4) + \dots + (6 \times 1)]/1223 = 4,02$$

es decir

$$(0,199 \times 5) + (0,645 \times 4) + \dots + (0,005 \times 1) = 4,02 \quad (7.2)$$

donde, en la segunda línea, aparecen nuevamente los valores del perfil (del grupo de edad de 16 a 24 años), $243/1223 = 0,199$, $789/1223 = 0,645$, etc., que hemos utilizado como pesos. Vemos que el grupo de edad más joven tiene una autopercepción media de la salud mayor que la media general, 4,02 con relación a 3,72. Podríamos repetir el cálculo anterior para los restantes seis grupos de edad, obteniendo las medias siguientes:

16-24	25-34	35-44	45-54	55-64	65-74	75+	Media global
4,02	3,97	3,86	3,66	3,39	3,30	3,19	3,72

Ahora que ya hemos calculado las medias de las categorías de salud de cada grupo de edad, podemos calcular su varianza. Este cálculo es similar al cálculo de inercia del capítulo 4 ya que ponderaremos cada grupo de edad proporcionalmente al tamaño de su muestra. Otra posibilidad sería asignar a cada uno de los 6371 encuestados el valor correspondiente a su respectivo grupo de edad, y hacer el cálculo habitual de la varianza. Hemos calculado la varianza como (véase la fila de totales de la tabla de la imagen 6.1):

$$\frac{1223}{6371}(4,02 - 3,72)^2 + \frac{1234}{6371}(3,97 - 3,72)^2 + \dots + \frac{396}{6371}(3,19 - 3,72)^2 = 0,0857$$

con desviación típica $\sqrt{0,0857} = 0,293$.

Todos los cálculos anteriores dependen de la escala entera asignada a las categorías de salud, una elección arbitraria verdaderamente difícil de justificar, especialmente después de ver los resultados del capítulo 6. La pregunta es: ¿existe una escala más justificable o, al menos, más interesante? La respuesta depende de lo que entendamos por «más interesante». Vamos a considerar un posible criterio que

Cálculo de la media de los grupos de edad mediante una escala entera

Cálculo de la varianza utilizando la escala entera

Cálculo de las puntuaciones en una escala desconocida

nos lleve a una escala con valores directamente relacionados con el AC. Supongamos que indicamos la escala asignada a las categorías de la salud por los valores desconocidos v_1, v_2, v_3, v_4 y v_5 . La media de todos los encuestados sería, en función de estos valores desconocidos, igual que en (7.1):

$$\text{media global de la salud} = (0,128 \times v_1) + (0,556 \times v_2) + \dots + (0,016 \times v_5) \quad (7.3)$$

y la media del grupo de edad de 16 a 24 años sería, igual que en (7.2):

$$\text{media de salud 16-24 años} = (0,199 \times v_1) + (0,645 \times v_2) + \dots + (0,005 \times v_5) \quad (7.4)$$

Llamamos *puntuaciones* a las medias calculadas de esta manera, así (7.3) es la puntuación media y (7.4) es la puntuación del primer grupo de edad, que indicamos como s_1 . Calcularíamos las puntuaciones, s_1, s_2, \dots, s_7 , de todos los grupos de edad de la misma manera, siempre en función de los valores desconocidos de la escala. Dado que cada uno de los 6371 encuestados pertenece a un grupo de edad, les podemos asignar la puntuación correspondiente de la escala de salud. Por ejemplo, a los 1223 encuestados del grupo de edad de 16 a 24 años les asignaremos la puntuación calculada en (7.4). Podríamos imaginar a los 6371 encuestados distribuidos en las siete puntuaciones de la escala de salud, independientemente de los valores que éstas tomen.

La maximización de la
varianza proporciona
la escala óptima

Para determinar los valores de la escala, propondremos que las 6371 puntuaciones cumplan determinadas propiedades. Una propiedad deseable sería que las puntuaciones estuvieran bien separadas entre sí, de manera que pudiéramos distinguir al máximo los grupos de edad. Dicho de otra forma, sería muy indeseable que las puntuaciones se encontraran muy cerca entre sí, de manera que nos fuera difícil distinguir entre grupos de edad en términos de sus categorías de salud. Una manera de expresar este requerimiento de forma más precisa es exigir que la varianza de las puntuaciones de los 6371 encuestados sea máxima. En términos numéricos, tenemos 1223 encuestados del primer grupo de edad (primera fila de la imagen 6.1) a los que hemos asignado la puntuación s_1 , 1234 del segundo grupo de edad a los que hemos asignado la puntuación s_2 , y así sucesivamente. Calcularemos la varianza de las 6371 puntuaciones, como hicimos en la página anterior. La *escala óptima* vendrá definida por los valores v_1, v_2, \dots, v_5 que hagan que la varianza de las s_1, s_2, \dots, s_7 puntuaciones sea máxima.

Los valores de la escala
óptima de la dimensión
del AC que mejor se
ajusta

Afortunadamente, las posiciones de las categorías de salud, en la dimensión del AC que mejor se ajusta, resuelve de forma exacta este problema del escalado óptimo. La varianza máxima es igual a la inercia de esta dimensión óptima del AC. Es decir, los valores de las coordenadas de los vértices de la imagen 6.5 son los valores de la escala óptima de v_1 a v_5 . A partir de los elementos de los perfiles de las filas podemos calcular sus correspondientes puntuaciones, de s_1 a s_7 .

CATEGORÍA DE LA SALUD	Coordenadas de vértices
<i>Muy buena</i>	1,144
<i>Buena</i>	0,537
<i>Regular</i>	-1,188
<i>Mala</i>	-2,043
<i>Muy mala</i>	-2,076

GRUPO DE EDAD	Coordenadas de perfiles
16-24	0,371
25-34	0,330
35-44	0,199
45-54	-0,071
55-64	-0,396
65-74	-0,541
75+	-0,658

Imagen 7.1:
Valores de las coordenadas de los puntos de la imagen 6.5, es decir, las coordenadas de los vértices de las columnas y de los perfiles de las filas en la dimensión que mejor se ajusta a los perfiles de las filas

En la tabla de la imagen 7.1 se muestran los valores de las coordenadas de los vértices y los de las coordenadas de los perfiles. En el capítulo 3, vimos que la posición de un determinado grupo de edad corresponde al centroide de los vértices de los cinco grupos de la salud. Esta propiedad también se cumple para las proyecciones de los perfiles en cualquier subespacio. Así, por ejemplo, obtenemos la puntuación del grupo de edad de 16 a 24 años (imagen 6.2), ponderando las posiciones de los vértices de las cinco categorías de salud con los perfiles correspondientes de este grupo de edad (imagen 7.1), de la siguiente manera:

$$(0,199 \times 1,144) + (0,645 \times 0,537) + \dots + (0,005 \times -2,076) = 0,371$$

lo que concuerda con la coordenada del perfil 16-24 de la imagen 7.1

También podríamos plantear el escalado óptimo al revés; es decir buscaríamos los valores de la escala de los grupos de edad que maximizaran la varianza de las categorías de salud. La solución viene dada por las coordenadas de los vértices de los cinco grupos de edad, siendo las coordenadas de los perfiles las puntuaciones de las categorías de la salud. En el siguiente capítulo veremos más a fondo la simetría existente entre el análisis de filas y el análisis de columnas. Esta simetría, o *dualidad*, del escalado óptimo ha llevado a algunos autores a denominar este método como *optimización dual de la escala*.

A diferencia de la escala entera original, la escala óptima no sitúa las cinco categorías de salud a distancias iguales. En el mapa de la imagen 6.5 vimos que existía una gran diferencia entre *buena* y *regular*, y una diferencia muy pequeña entre *mala* y *muy mala*. Estos valores de la escala óptima de las categorías de la salud hacen que las puntuaciones de los grupos de salud estén lo más separadas posible según el criterio de la varianza. Es decir, utilizando la escala óptima de las categorías de salud, obtenemos la máxima discriminación entre los grupos de edad. En el mapa de la imagen 6.3, en la que sólo representamos las puntuaciones de los

Interpretación de la escala óptima

grupos de edad, vemos que hasta el grupo de edad de 34 a 45 años, existen pequeños cambios en la autopercepción de la salud, luego vemos grandes cambios en los grupos de edad media, especialmente entre los grupos de 45 a 54 y de 55 a 64 años, y finalmente cambios más pequeños en los grupos de mayor edad. Revisando otra vez los perfiles de la tabla de la imagen 6.2, podemos comprobar que entre los grupos de edad de 45 a 54 años y de 55 a 64 años se produce una caída de aproximadamente el 50% en la categoría *muy buena* y un incremento de más del doble en la categoría *mala*. Estos cambios en los valores de los perfiles explican los cambios observados en las puntuaciones.

Condiciones de
identificación de una
escala óptima

Los valores de la escala óptima obtenidos para las categorías de la salud son 1,144, 0,537, -1,188, -2,043 y -2,076, respectivamente (imagen 7.1). Hemos calculado estos valores en determinadas restricciones, necesarias para poder hallar una sola solución. Estas restricciones son que, para los 6371 encuestados, la media de los valores de la escala de salud sea 0 y que su varianza sea 1:

$$(0,128 \times 1,144) + (0,556 \times 0,537) + \dots + (0,016 \times (-2,076)) = 0 \quad (\text{media } 0)$$

$$(0,128 \times 1,144^2) + (0,556 \times 0,537^2) + \dots + (0,016 \times (-2,076)^2) = 1 \quad (\text{varianza } 1)$$

Estos prerrequisitos para los valores de la escala son las *condiciones de identificación* o *restricciones* en el lenguaje utilizado en la teoría matemática de optimización. La primera condición es necesaria, ya que podrían existir dos escalas distintas que tuvieran la misma varianza pero medias diferentes. Es decir, sería imposible identificar una solución sin especificar la media. La segunda condición es necesaria ya que, si multiplicamos de forma arbitraria los valores de la escala por un valor grande, la varianza de las eventuales puntuaciones se incrementaría mucho, lo que no tendría sentido alguno, pues estamos intentando maximizar la varianza. En consecuencia, es necesario que busquemos una escala que tengan una determinada media y un determinado rango de variación. Aunque las condiciones de «media 0 y varianza 1» son una elección arbitraria, conducen a unas coordenadas adecuadas para los vértices del AC, que también cumplen estas condiciones.

Cualquier transformación
lineal de la escala sigue
dando una escala óptima

Para determinar la escala óptima, las dos condiciones de identificación que hemos descrito anteriormente son simples instrumentos técnicos que aseguran una sola solución matemática de nuestro problema. Sin embargo, una vez obtenidos los valores de la escala, tenemos la posibilidad de transformarlos en una escala más conveniente, siempre y cuando recordemos que la media y la varianza de la escala transformada no tienen ninguna trascendencia sustantiva o relevancia estadística. En general, llevamos a cabo la redefinición de esta escala, fijando los puntos extremos, de manera que tengan valores con algún significado. Por ejemplo, en este caso, podríamos dar el valor 0 a la categoría *muy mala* salud, y el va-

CATEGORÍA DE LA SALUD	Valor en la escala óptima	Valor en la escala transformada
<i>Muy buena</i>	1,144	100,0
<i>Buena</i>	0,537	81,1
<i>Regular</i>	-1,188	27,6
<i>Mala</i>	-2,043	1,0
<i>Muy mala</i>	-2,076	0,0

Imagen 7.2:
Valores de la escala óptima del AC y valores transformados para que la escala esté entre 0 y 100

lor 100 a la de *muy buena*. En tal caso, necesitamos hacer que una transformación asigne el valor 0 a -2,076 y el valor 100 a 1,144. Para ello, en primer lugar, podemos sumar 2,076 a todos los valores de la escala, de manera que el valor más pequeño sea 0. Ahora la escala va de 0 a 1,144 + 2,076 = 3,220. Para asignar 100 al mayor valor de la escala, podemos multiplicar todos los valores por 100/3,220. En este caso en concreto la fórmula de cálculo para pasar de la antigua a la nueva escala es simplemente:

$$nueva = (antigua + 2,076) \times \frac{100}{3,220}$$

o, para el caso general:

$$nueva = \left[(antigua - \text{antiguo límite inferior}) \times \frac{\text{rango nuevo}}{\text{rango antiguo}} \right] + \text{nuevo límite inferior} \quad (7.5)$$

(en nuestro ejemplo el nuevo límite inferior es 0). Aplicando esta fórmula a los cinco valores de la escala óptima, obtenemos los valores transformados de la imagen 7.2.

La escala anterior de 5 a 1, con cuatro intervalos iguales entre los puntos de la escala, tendría los valores 100, 75, 50, 25, 0 en la escala transformada de rango 100 (recordemos que hemos invertido la escala de manera que *muy buena* sea 100). Sin embargo, en la escala óptima transformada, *regular* no se halla en el punto medio (50) de la escala, se halla mucho más cerca del extremo «mala» salud de la escala.

Debemos insistir en que la escala óptima depende del criterio establecido para su determinación, así como de las condiciones de identificación escogidas. Aparte de los criterios puramente técnicos, también depende, de forma clara, de la tabla de contingencia original. Si tuviéramos una tabla de contingencia que cruzara autopercepción de la salud con otra variable demográfica, por ejemplo, nivel de educación, obtendríamos, una escala óptima distinta para las categorías de la salud, ya que ahora el procedimiento discriminaría de manera óptima las diferencias entre niveles de educación.

La escala óptima no es única

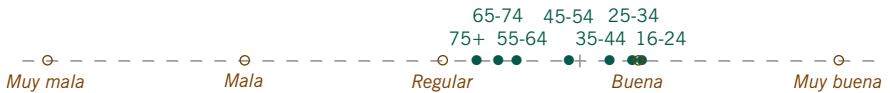
Un criterio basado en las distancias entre las filas y las columnas

En contraste con el criterio de maximización que hemos descrito anteriormente, a continuación presentamos un criterio de minimización para hallar los valores de la escala óptima que también conduce a la solución del AC. Este criterio se basa en las distancias entre las filas y las columnas —en el ejemplo que nos ocupa estas distancias serán las distancias entre las categorías de salud y los grupos de edad—. Imaginemos, en primer lugar, las categorías de salud en una determinada escala, por ejemplo, la escala entera de 1 a 5, de *muy mala* a *muy buena* salud, que representamos gráficamente en la imagen 7.3. Ahora el objetivo es hallar, en la misma escala, las posiciones de los grupos de edad que se hallen tan «próximos» como sea posible a las categorías de salud, en el sentido de que un grupo de edad, que tenga una frecuencia elevada para una determinada categoría de salud, tienda a aproximarse a esta categoría. Supongamos ahora que los valores de las categorías de salud (en este ejemplo inicial eran los valores de 1 a 5) son los valores h_1, h_2, \dots, h_5 y que los valores de los grupos de edad son a_1, a_2, \dots, a_7 . La distancia entre un grupo de edad y una categoría de salud es igual al valor absoluto de la diferencia $|a_i - h_j|$; sin embargo, preferimos, como medida de proximidad, la distancia al cuadrado $(a_i - h_j)^2$.^{*} Para que las distancias dependan de las frecuencias de ocurrencia de las tablas de contingencia, ponderaremos cada distancia al cuadrado con p_{ij} , la frecuencia relativa como la definimos en la página 51 del capítulo 4. Es decir, los valores de la imagen 6.1 divididos por la suma total 6371 (por tanto, la suma de todos los p_{ij} es 1). Nuestro objetivo sería entonces minimizar la función siguiente:

$$\sum_i \sum_j p_{ij} d_{ij}^2 = \sum_i p_{ij} (a_i - h_j)^2 \tag{7.6}$$

que tenderá a acortar las distancias cuando p_{ij} sea mayor. Dados unos determinados valores h_j de las categorías de salud, es fácil demostrar que obtenemos un mínimo de (7.6) con las medias ponderadas de los grupos de edad. Para los valores de las categorías de salud de 1 a 5, estas medias ponderadas son las puntuaciones que calculamos antes —en la fórmula (7.2) y las puntuaciones que le siguen—, que también hemos representado en el mapa de la imagen 7.3. Las dos escalas que mostramos en el mapa de la imagen 7.3 minimizan (7.6), pero si los valores de la escala de la salud fueran otros, ¿cuál sería el valor del mínimo? Para poder

Imagen 7.3:
La escala 1-5 de las categorías de salud y las medias ponderadas de los grupos de edad



^{*} De nuevo, como anteriormente, es siempre más fácil trabajar con las distancias al cuadrado —la raíz cuadrada de las expresiones que conducen a distancias euclídeas causan muchos problemas de optimización—; estas dificultades desaparecen cuando consideramos la optimización mínimo-cuadrática.

responder con sentido a esta pregunta, necesitamos, otra vez, definir unas condiciones de identificación; en caso contrario podríamos llegar a una solución que situara a todas las categorías de salud en el mismo punto. Si consideramos las mismas condiciones de identificación que vimos anteriormente para los valores de las categorías de la salud, es decir, media 0 y varianza 1, obtendremos, de nuevo, el mínimo con la dimensión óptima del AC. Comparando las posiciones de los grupos de edad en el mapa de la imagen 7.3 con las posiciones óptimas en el mapa de la imagen 6.5, vemos que la dispersión de los grupos de edad es mayor en el mapa 6.5, lo que significa que en el mapa 7.3, los grupos de edad quedan más cerca de las categorías de salud en términos del criterio (7.6). El valor del mínimo alcanzado en el mapa de la imagen 6.5 es igual a 1 menos la varianza (maximizada) de la dimensión óptima del AC, lo que llamamos *pérdida de homogeneidad*. (Volveremos a este concepto en el capítulo 20, cuando tratemos sobre el análisis de homogeneidad.) El criterio (7.6) lo podemos generalizar fácilmente a dos o más dimensiones, digamos K dimensiones, simplemente sustituyendo a_i y h_j por vectores con K elementos y sustituyendo los cuadrados de las diferencias $(a_i - h_j)^2$ por el cuadrado de las distancias euclídeas en un espacio de dimensión K .

1. El *escalado óptimo* asigna valores a las categorías (o atributos) de una variable categórica mediante algún criterio de optimización que separe, o discrimine, los grupos de casos que hemos formado al cruzar los casos con dicha variable.
2. Las posiciones de las categorías son los vértices en la dimensión óptima del AC que proporcionan unos valores de escala óptimos, en el sentido de que maximizan la varianza entre los grupos. Las puntuaciones de los grupos son las proyecciones de sus perfiles sobre esta dimensión. La varianza máxima de las puntuaciones es igual a la inercia de las proyecciones de los perfiles.
3. Es característico de la geometría del AC que las posiciones de las coordenadas de las proyecciones de las categorías sobre la dimensión óptima estén estandarizadas. De todas formas, en la práctica podemos recentrar y redimensionar los valores de la escala, por ejemplo, para que sus valores vayan de 0 a 1 o de 0 a 100. En tal caso variarán los valores de la media y la varianza.
4. También podemos hallar la escala óptima a partir de un criterio basado en las distancias entre filas y columnas. Concretamente, se trata de situar en el mapa las coordenadas de filas y columnas de manera que se minimicen las distancias ponderadas entre filas y columnas —ponderadas con las frecuencias relativas obtenidas de la tabla de contingencia—. Este valor mínimo es igual a 1 menos la varianza (máxima) de las puntuaciones en la escala óptima.