Los datos de metagenómica son composicionales. El problema y la solución.

A. Blasco[1], M. Martínez Alvaro[2], M. Greenacre[3]

[1] Instituto de Ciencia y Tecnología Animal. Universitat Politècnica de València, Apartado 22012, Valencia 46022, España.
ablasco@dca.upv.es
[2] Department of Agriculture, Scotland's Rural College, The Roslin Institute, Easter Bush Campus, Midlothian, EH25 9RG, Scotland.
[3] Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27, Barcelona 08005, Spain.

Resumen

Los datos composicionales (e.g. datos metagenómicos) provienen de variables no negativas expresadas como proporciones que suman 1. En datos composicionales aparecen correlaciones espurias; por ejemplo, imaginemos una composición formada por las proporciones de dos bacterias ambas perjudiciales para la salud; aumentando una disminuye la otra. Si los pacientes empeoran aumentando la proporción de una, parecerá que empeoran disminuyendo la segunda. Además, al extraer un subconjunto de variables y volver a expresarlas como proporciones que suman 1, las relaciones entre las variables cambian, las correlaciones entre ellas ya no son las mismas que en la base de datos original, no hay *coherencia subcomposicional*.

Para evitar los problemas citados, los datos composicionales se analizan como ratios, que no presentan estos problemas. Se usa el logaritmo del ratio para que los datos estén en la recta real evitando que sean porcentajes y además las distribuciones son más simétricas.

Con p variables composicionales, el "espacio completo" está definido por todos los $p(p-1)/2$ log-ratios posibles, pero como en metagenómica hay una enorme cantidad de variables, se usan subconjuntos de menos dimensiones. El más simple y menos recomendado, por el que nos decantaremos, es el log-ratio aditivo (ALR), en donde las variables x_i se representan con respecto a una variable de referencia: $\log(x_i/x_{ref})$. Otra alternativa (CLR) es usar como referencia la media geométrica G de las variables dentro de cada muestra $\log(x_i/G)$ centrando así los $\log(x_i)$, y otra alternativa (ILR) es usar medias geométricas de subconjuntos de variables $\log(G_i/G_k)$, llamados "balances".

El CLR es *isométrico* respecto al "espacio completo", conserva las distancias entre variables, pero no tiene coherencia subcomposicional porque cada subconjunto de variables tiene una G diferente. El ILR es isométrico y tiene coherencia subcomposicional, pero los balances son muy difíciles de interpretar.

El ALR tiene ventajas. Si escogemos una x_{ref} que tenga muy poca variabilidad entre individuos de la muestra, comparar dos $\log(x_i/x_{ref})$ es casi como comparar dos $\log(x_i)$, algo fácil de interpretar, y además casi desaparecen las correlaciones espurias. La media geométrica G del CLR es mucho más variable entre individuos y por eso el CLR no tiene esta facilidad de interpretación. Sin embargo, el ALR no está recomendado porque no es isométrico respecto al "espacio completo" y el resultado depende de la x_{ref}. ¿Hasta qué punto los dos espacios son aproximadamente iguales? Podemos proyectar el espacio reducido del ALR en el "espacio completo", usando *análisis Procrustes*; si la correlación de las coordenadas en ambos espacios es próxima a 1, los dos espacios tienen una geometría similar.

En metagenómica hay tal cantidad de variables, que probablemente encontraremos una poco variable entre individuos y que dé lugar a un ALR con elevada correlación Procrustes con el “espacio completo”. Esto es exactamente lo que hemos hecho; hemos tomado varias bases de datos metagenómicas de diferentes orígenes y para todos ellos hemos encontrado varias posibles variables candidatas a variable de referencia, que permiten trabajar en un espacio aproximadamente isométrico con el completo, pero con variables mucho más fáciles de interpretar.

Palabras clave: Datos composicionales. Metagenómica. Transformación ALR