

Análisis PCA,PLS,CVA: discusión, conclusiones

Antonio Sala Piqueras

Notas de clase sobre identificación y control multivariable

Dept. Ing. Sistemas y Automatica (DISA)
Universitat Politècnica de València (UPV)

Presentación en vídeo en:

<http://personales.upv.es/asala/YT/V/plsdisc.html>

Conclusiones

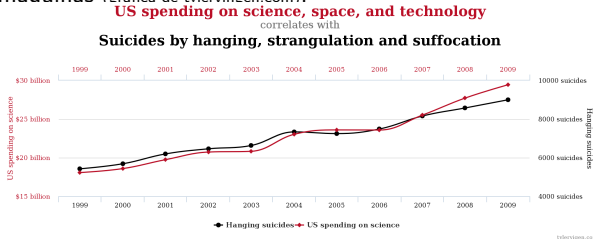
- El SVD de una matriz de datos tiene interpretación estadística de **componentes principales** descomponiendo un conjunto de señales en componentes *no correlacionados*.
 - Separa entre causas de la variabilidad (σ_i alta) –información– y modelos (σ_i baja).
 - No distingue entradas y salidas.
- EL SVD del modelo de predicción (regresión) descompone las **entradas** según su grado de “*utilidad para predecir*” las **salidas**.
 - Permite determinar que, por ejemplo, el 98% de la covarianza entre un vector de 20 salidas y un conjunto de 150 entradas es explicado por **4** variables “*latentes*”.

*Es un ingrediente fundamental en identificación “subespacio” de sistemas multivariados, esas variables latentes serán lo que conocemos por variables de **estado**. Si las 150 entradas contienen elementos que son funciones no-lineales de un número menor de entradas “reales”, permite escoger automáticamente aquellas características más relevantes para integrar en un modelo no lineal de predicción (en inteligencia computacional, automatic **feature extraction**).

Discusión, perspectivas

EL SVD es eficiente y variaciones están presentes en muchos algoritmos de bancos, aseguradoras, firmas de inversión, análisis de datos empresariales, servicios de inteligencia...

- Correlación no implica causalidad: la asociación puede ser debida a terceros factores, o no existir... hay tantos PetaBytes de información disponible que realizaciones de baja muy probabilidad son encontradas por las máquinas (gráfica de tvlervingen.com):



La probabilidad de 20 caras seguidas en monedas “sin trugar” es 2^{-20} (una entre un millón). Con 1 terabyte ($8 \cdot 10^{12}$ bits) en tiradas, es casi seguro que encuentre 20 caras seguidas ($p > 0.99999$); en youtube hay 400000 TB de datos; ver también es.wikipedia.org/wiki/Teorema_del_mono_infinito

Discusión, perspectivas (II)

- Engañar con estadística es más fácil... la gráfica anterior puede dar lugar a una ley.
- La selección de los datos a los algoritmos es subjetiva, el algoritmo usado también da resultados ligeramente diferentes (PCA, FA, CVA, ANOVA,...): seleccionar entradas, salidas a predecir y algoritmo hasta que salga *lo que yo quiero que salga*.
- Aún sin ser conscientemente deshonesto, los datos pasados tienen prejuicios, estereotipos, racismo, que los algoritmos estadísticos contribuyen a perpetuar en el futuro.

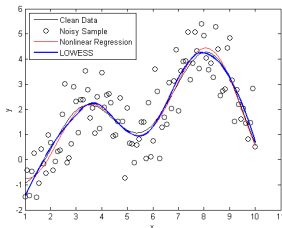
Pero estos algoritmos son "secretos", y a veces hasta para sus creadores son difíciles de analizar (¿cómo interpretar tres componentes principales de 420 elementos?, "caja negra")...

En definitiva, la profusión a todos los niveles de este tipo de análisis tiene consecuencias políticas/sociológicas importantes... Todo es complicado...

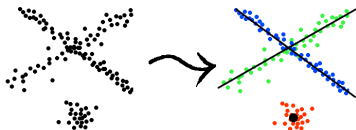
Limitaciones

Formalmente, PCA y CVA buscan ajustar un modelo lineal a los datos y otro modelo lineal (perpendicular) a los residuos, etc... Básicamente, se adapta a nubes de datos **elipsoidales**.

En algunos casos, los datos son una mezcla de varios modelos lineales (**mixture** models), o no-lineales:



<http://blogs.mathworks.com/loren/2011/01/13/data-driven-fitting/>



<http://perception.csl.illinois.edu/gpca/introduction/index.html>

El tratamiento de esos datos con SVD obtiene resultados **muy** malos. En datos de dimensión 2/3 se puede “ver” dichos problemas, pero en datos de mayor dimensión es un problema difícil, que requiere **clustering**, **generalised SVD**, **redes neuronales**, etc...