Análisis de variables canónicas (CVA): planteamiento y justificación del problema

Antonio Sala Piqueras

Notas de clase sobre identificación multivariable

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

Video-presentación disponible en:

http://personales.upv.es/asala/YT/V/**cva1**.html

Presentación

Motivación:

En estadística monovariable, la correlación $-1 \leq \rho \leq 1$ (covz./(prod. desv. típicas), adimensional) es la predicción entre variables normalizadas a varianza 1. Con ese escalado, coinciden regresión directa e inversa $\hat{y}_{norm} = \rho x_{norm}$, $\hat{x}_{norm} = \rho y_{norm}$, y el error de predicción puede entenderse en términos porcentuales, relativos. ¿existe un análisis similar para el caso multivariable?

Objetivos:

Comprender el significado de la correlación entre vectores de variables y la descripción independiente de escalado que consigue de un sistema.

Contenidos:

Motivación y distinción en el planteamiento respecto PLS. Blanqueado de entradas y salidas. Matriz de correlación. Conclusiones.

Canonical Variate Analysis (CVA) versus PCR, PLS

Consideremos una serie de datos, x, y, con el objetivo de predecir y a partir de x.
Su matriz VC conjunta es:

$$\begin{bmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma_{yx}^T & \Sigma_x \end{bmatrix}$$

Tenemos varias opciones anteriores para intentar obtener una "estructura" de modelo, que, quizás, permita reducir la complejidad sin perder capacidad de explicación. (PCR basada en SVD de X, PLS y O-PLS basados en SVD de covarianzas).

Ahora, una más: **CVA**, basada en SVD de **correlación**.

Nota: Los desarrollos están realizados para x, y vectores *columna*. Todas las matrices deben transponerse para obtener resultados con datos en forma de vector fila.

CVA: Blanqueado (ortonormalización) de entradas y salidas

En el CVA se propone blanqueado (media cero, varianza identidad) tanto de **entradas** x como de **salidas a predecir** y.

- El blanqueado de entradas no cambia la varianza de error de predicción (el predictor incorporaría la inversa del cambio de variable).
- El blanqueado de salidas **sí** cambia el resultado de la regresión: al escalar de forma diferente unas variables de otras, el error cuadrático (variación total , traza de VC) es ponderado de forma diferente.

*Sólo un cambio ortogonal $T^{-1} = T^T$ deja inalterada la variación total.

El escalado de los componentes principales de y a varianza 1 tiene una interpretación "porcentual": un error de predicción de varianza 0.5 en un componente de varianza 5 es tratado igual que un error de varianza 0.02 en un componente de varianza 0.2. Es razonable en predicción de variables heterogéneas (unidades, fuentes de datos, ... diferentes)... salvo que componentes de poca varianza se consideren ruido y se proponga una

Matriz de correlación entre variables ortonormalizadas

Con variables $(\tilde{x}, \tilde{y})$ blanqueadas respecto a unas originales (x,y), esto es $\tilde{x} = \Sigma_x^{-\frac{1}{2}} x$, $\tilde{y} = \Sigma_y^{-\frac{1}{2}} y$, la matriz de varianzas covarianzas es:

$$\begin{bmatrix} I & \Sigma_{\tilde{y}\tilde{x}} \\ \Sigma_{\tilde{y}\tilde{x}}^T & I \end{bmatrix}$$

siendo, en función de las variables originales (no blanqueadas):

$$\Sigma_{\tilde{y}\tilde{x}} = \Sigma_y^{-\frac{1}{2}} \Sigma_{yx} \Sigma_x^{-\frac{1}{2}}$$

En caso escalar, sería $\frac{\Sigma_{xy}}{\sigma_x \sigma_y}$ siendo σ la desviación típica (raíz cuad. de varianza), esto es, sería el **coeficiente de correlación** ρ de una regresión lineal.

En multivariable, $\Sigma_{\tilde{y}\tilde{x}}$ es una **matriz de correlación**.

Conclusiones

- El preblanqueado de entradas y salida implica regresión entre componentes principales de ambas, en términos "relativos" (todo escalado a base 1).
- La covarianza entre variables ortonormalizadas es una matriz de correlación.
- PLS con esas nuevas variables se convierte en CVA, desarrollado en otros materiales.