Mínimos Cuadrados Parciales (PLS)

Antonio Sala Piqueras

Notas de clase sobre identificación multivariable

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

Presentación en vídeo en:
http://personales.upv.es/asala/YT/V/pls1.html, plso.html, plss.html

Presentación

Motivación:

Al precedir y a partir de x, los componentes más "activos" de X pueden no ser los más importantes para la predicción. Descomposición basada en la covarianza Σ_{yx} en vez de en Σ_x (PCR) o $\Sigma_{[x\,y]}$ (TLS).

Objetivos:

Comprender la motivación, planteamiento y alguna de las soluciones al problema PLS.

Contenidos:

Preliminares y planteamiento del problema. Preparación de datos. Regresión PLS con información ortonormalizada. Regresión SIMPLS (idea básica). Discusión y conclusiones.

Componentes principales en regresión

Ideas preliminares:

- Tenemos un conjunto de datos x, para predecir y a partir de ellos. Tenemos N muestras (y, x) que constan de $y_{m_y \times 1}$ datos a precedir junto a $x_{m_x \times 1}$ datos conocidos de entrada al modelo.
- El PCA básico (TLS) no distingue entradas de salidas, el PCR usa unos componentes en x que no dependen de y.
- Es posible que **los datos x tengan correlación entre sí**, así como los datos y (información "repetida").
  - No queremos identificar relaciones $x_5 - 3x_2 \approx 0$, $y_1 - 2y_2 + 3y_6 \approx 0$.
- Es posible que haya mucha información en x que **no es útil** para predecir y, e información en y que **no puede ser predicha** por x.

Planteamiento del problema

- El **objetivo** de la identificación o del análisis de datos suele ser determinar:
  1. "qué factores/componentes de x realmente **influyen** en y",
  2. "qué componentes de y son influenciados (**explicados**) por x".

- Predicción lineal óptima: $\hat{y} = H \cdot x = \Sigma_{yx}\Sigma_x^{-1} \cdot x$ (norm. media cero)

- H es una matriz "enorme" en caso multivariable... ¿podría informarnos el SVD de H sobre los objetivos buscados?
  - La respuesta es **afirmativa**, y discutir sobre el significado de ese SVD es el objetivo de este material...

Preparación de datos

Para que el ajuste por mínimos cuadrados y las conclusiones sobre el SVD de H tengan sentido, en un entorno multivariable, hace falta un escalado/preprocesado previo.

- Escalado de las variables y a predecir: Escalar los elementos de y para que error de ajuste en todos los elementos tenga significado "comparable" en la aplicación: mezclar errores de estimación en "metros" con otros errores en "milivoltios" requiere ponderación.
- Escalado de las variables x de entrada al modelo: Los valores singulares y direcciones del SVD dependeb de escalado, hacen falta refinamientos. Además se requieren componentes "no correlados entre sí" en x: si Σ_x no es diagonal, un coeficiente de H puede ser 0.5 porque el de otro componente "muy correlado" también es 0.5... si no estuviera el segundo componente, el predictor con sólo el primero sería 0.98.

Ortonormalización (preblanqueado) de la información

- Normalización de X: En una predicción por mínimos cuadrados de y dado x cualquier cambio de variable lineal, T, en x producirá el mismo error de predicción (distintos parámetros $\Theta x = \underbrace{\Theta T^{-1}}_{\widetilde{\Theta}} \cdot \underbrace{Tx}_{\widetilde{x}}$).

  Suponemos, entonces, tras **escalado** y **preblanqueado**, $T = \Sigma_x^{-\frac{1}{2}}$, que disponemos de muestras de datos x de varianza IDENTIDAD tras el preprocesado adecuado.

- Formar matrices de datos $Y_{m_y \times N}$, $X_{m_x \times N}$, y varianzas-covarianzas $\frac{1}{N-1} XX^T = I$, $\Sigma_y = \frac{1}{N-1} YY^T$, $\Sigma_{yx} = \frac{1}{N-1} YX^T$.

  Por ejemplo, si X antes de escalar fuera $X = U_x S_x V_x^T$, $X_{new} := \sqrt{N-1} V_x^T = \underbrace{\sqrt{N-1} S_x U_x^T}_{T} X$ sería el resultado del escalado y preblanqueado del x original, porque $\frac{1}{N-1} X_{new} X_{new}^T = I$.

Nota: Si los datos x, y fueran vectores **fila** podrían transponerse, o bien, cambiar todas las fórmulas de estas transparencias a su transpuesta.

Regresión PLS con información preblanqueada (ortonormalizada)

1. Identificar H, en un modelo $Y = H \cdot X$ (pred. óptima lineal)
   - Si $\Sigma_x = I$, la fórmula de mejor predicción es, directamente, $H \equiv \Sigma_{yx}$.
2. Hacer **SVD** (econ.) de $H = USV^T$, $S = diag(\sigma_1, \ldots, \sigma_{m_1})$, $m_1 = \min(m_y, m_x)$.
3. Si hacemos el cambio $\hat{y} = U^T y$, $\hat{x} = V^T x$, entonces:
   - las matrices $\Sigma_{\hat{y}} = E[\hat{y}\hat{y}^T] = U^T E[yy^T]U = U^T \Sigma_y U$,
   - $\Sigma_{\hat{x}} = V^T \Sigma_x V = I$.
   - Como $\Sigma_{\hat{y}\hat{x}} = U^T E[yx^T]V = U^T(USV^T)V = S$, la matriz de Varianzas-Covarianzas conjunta queda:

$$\Sigma_{[\hat{y},\hat{x}]} = E\left(\begin{bmatrix}\hat{y}\\ \hat{x}\end{bmatrix}\begin{bmatrix}\hat{y}^T & \hat{x}^T\end{bmatrix}\right) = \begin{pmatrix}\Sigma_{\hat{y}} & S\\ S^T & I\end{pmatrix}$$

Regresión PLS con información preblanqueada (ortonormalizada)

1. Identificar H, en un modelo $Y = H \cdot X$ (pred. óptima lineal)
   - Si $\Sigma_x = I$, la fórmula de mejor predicción es, directamente, $H \equiv \Sigma_{yx}$.
2. Hacer **SVD** (econ.) de $H = USV^T$, $S = diag(\sigma_1, \ldots, \sigma_{m_1})$, $m_1 = \min(m_y, m_x)$.
3. Si hacemos el cambio $\hat{y} = U^T y$, $\hat{x} = V^T x$, entonces:
   - las matrices $\Sigma_{\hat{y}} = E[\hat{y}\hat{y}^T] = U^T E[yy^T]U = U^T \Sigma_y U$,
   - $\Sigma_{\hat{x}} = V^T \Sigma_x V = I$.
   - Como $\Sigma_{\hat{y}\hat{x}} = U^T E[yx^T]V = U^T(USV^T)V = S$, la matriz de Varianzas-Covarianzas conjunta queda:

$$\Sigma_{[\hat{y},\hat{x}]} = E\left(\begin{bmatrix}\hat{y}\\ \hat{x}\end{bmatrix}\begin{bmatrix}\hat{y}^T & \hat{x}^T\end{bmatrix}\right) = \begin{pmatrix}\Sigma_{\hat{y}} & S\\ S^T & I\end{pmatrix}$$

   - la mejor predicción lineal de $\hat{y}$ es $\nu := S\hat{x}$.

Regresión PLS con información preblanqueada (ortonormalizada)

1. Identificar H, en un modelo $Y = H \cdot X$ (pred. óptima lineal)
   - Si $\Sigma_x = I$, la fórmula de mejor predicción es, directamente, $H \equiv \Sigma_{yx}$.
2. Hacer **SVD** (econ.) de $H = USV^T$, $S = diag(\sigma_1, \ldots, \sigma_{m_1})$, $m_1 = \min(m_y, m_x)$.
3. Si hacemos el cambio $\hat{y} = U^T y$, $\hat{x} = V^T x$, entonces:
   - las matrices $\Sigma_{\hat{y}} = E[\hat{y}\hat{y}^T] = U^T E[yy^T]U = U^T \Sigma_y U$,
   - $\Sigma_{\hat{x}} = V^T \Sigma_x V = I$.
   - Como $\Sigma_{\hat{y}\hat{x}} = U^T E[yx^T]V = U^T(USV^T)V = S$, la matriz de Varianzas-Covarianzas conjunta queda:

$$\Sigma_{[\hat{y},\hat{x}]} = E\left(\begin{bmatrix}\hat{y}\\ \hat{x}\end{bmatrix}\begin{bmatrix}\hat{y}^T & \hat{x}^T\end{bmatrix}\right) = \begin{pmatrix}\Sigma_{\hat{y}} & S\\ S^T & I\end{pmatrix}$$

   - la mejor predicción lineal de $\hat{y}$ es $\nu := S\hat{x}$.

PLS ortonormalizado/preblanqueado (2)

Tras el segundo cambio de variable (el primero era el preblanqueado y escalado), hemos obtenido:

- **Predictores monovariables sin interacción:** La mejor predicción lineal de $\hat{y}_i$ dado $\hat{x}_i$ es $\nu_i = \sigma_i \hat{x}_i$; La mejor predicción de $\hat{y}_i$ dado cualquier otro componente de x es CERO.

  Si Σ_x no fuera diagonal (en este caso, identidad), los predictores tendrían interación (y el coef. de la mejor predicción dado "x_i" sería diferente que en el caso de conocer "x_i y x_j".

- El error de predicción (en coords $\hat{y}$) tiene matriz de varianza-covarianza $\Sigma_{\hat{e}} = \Sigma_{\hat{y}} - SS^T$. La variación total del error es $\mathbb{V}_e = \mathbb{V}_y - \sum_{j=1}^{m_1} \sigma_j^2$.

  Nota: $\mathbb{V}_y = \mathbb{V}_{\hat{y}}$, porque son un cambio ortogonal (no cambia valores propios, no cambia traza).

PLS ortonormalizado/preblanqueado: resumen, discusión

Hemos ordenado componentes **no correlados** de x según influyen en y (según los valores en $diag(S)$), y, simultáneamente, obtenido las características de y mejor predichas a partir de x, $\nu = S\hat{x}$.

*Los elementos de $\hat{y}$ no son independientes (tienen correlación, $\Sigma_{\hat{y}} \neq I$): aunque el predictor (diagonal) ν lo forman componentes no correlados entre sí, $\hat{y}$ (y el original y) puede estar afectado por otras terceras variables (no correladas con x).

*Los componentes ($\hat{y} = U^T y$, $\hat{x} = V^T x$) no coinciden con los componentes principales (PCA) de y o x (antes de normalizar), respectivamente: U y V vienen del SVD de la covarianza, y no de las matrices de datos Y, X por separado.

PLS ortonormalizado/preblanqueado: resumen, discusión

Hemos ordenado componentes **no correlados** de x según influyen en y (según los valores en $diag(S)$), y, simultáneamente, obtenido las características de y mejor predichas a partir de x, $\nu = S\hat{x}$.

*Los elementos de $\hat{y}$ no son independientes (tienen correlación, $\Sigma_{\hat{y}} \neq I$): aunque el predictor (diagonal) ν lo forman componentes no correlados entre sí, $\hat{y}$ (y el original y) puede estar afectado por otras terceras variables (no correladas con x).

*Los componentes ($\hat{y} = U^T y$, $\hat{x} = V^T x$) no coinciden con los componentes principales (PCA) de y o x (antes de normalizar), respectivamente: U y V vienen del SVD de la covarianza, y no de las matrices de datos Y, X por separado.

Modelos simplificados: Descartar valores pequeños en $diag(S)$.

PLS ortonormalizado/preblanqueado: resumen, discusión

Hemos ordenado componentes **no correlados** de x según influyen en y (según los valores en $diag(S)$), y, simultáneamente, obtenido las características de y mejor predichas a partir de x, $\nu = S\hat{x}$.

*Los elementos de $\hat{y}$ no son independientes (tienen correlación, $\Sigma_{\hat{y}} \neq I$): aunque el predictor (diagonal) ν lo forman componentes no correlados entre sí, $\hat{y}$ (y el original y) puede estar afectado por otras terceras variables (no correladas con x).

*Los componentes ($\hat{y} = U^T y$, $\hat{x} = V^T x$) no coinciden con los componentes principales (PCA) de y o x (antes de normalizar), respectivamente: U y V vienen del SVD de la covarianza, y no de las matrices de datos Y, X por separado.

Modelos simplificados: Descartar valores pequeños en $diag(S)$.

PLS ante entrada X no "blanqueada" a varianza I

El escalado a varianza unidad de x realmente es denominado **Orthonormalised PLS**.

Hay otras variantes en literatura. Por ejemplo, **SIMPLS** (el que implementa Matlab en su comando `plsregress`).

SIMPLS busca maximizar $q_i^T \Sigma_{yx} r_i$, $i = 1, \ldots, m_x$ sujeto a:

1. $q_i^T q_i = 1$, $r_i^T r_i = 1$
2. $r_i^T \Sigma_x r_j = 0$ si $i \neq j$, esto es, ortogonalidad ($\approx$ no correlación) entre $t_i := X r_i$ y $t_j := X r_j$.

La mejor predicción del componente escalar $\nu := q_i^T y$ dado el componente escalar $\tau := r_i^T x$ es $\sigma_i \cdot \tau$, donde σ_i es el mayor valor singular de una secuencia de matrices $S_0 = \Sigma_{yx}^T$, $S_1 = P_0 S_0$, $S_j = P_{j-1} S_{j-1}$, siendo P_j matrices de proyección ortogonal sobre cierto subespacio.

Detalles en (https://doi.org/10.1016/0169-7439(93)85002-X).

Comparación

Nota: sin la condición 2, los máximos-silla-mínimos del problema vendrían del SVD de Σ_{yx}, y tendríamos $r_i^T r_j = 0$ para $i \neq j$.

En cambio, SIMPLS sustituye $r_i^T r_j = 0$ por $r_i^T \Sigma_x r_j = 0$.

Si los datos de entrada están preblanqueados $\Sigma_x = I$, por lo que SIMPLS coincide con lo presentado en transparencias anteriores (O-PLS).

En otros casos, no: SIMPLS es sensible al escalado en X.

Hay más variaciones de PLS, en concreto NIPALS es también popular.

Comparación (2)

A partir de: covarianza=desv.típ(x)*correlación*desv.típ(y)

1. PCR: seleccionar para regresión componentes no correlados de x con gran "varianza de x", tengan o no correlación con y.

2. SIMPLS (no ortogonalizado): componentes no correlados de x con mucha covarianza con y... grosso modo, mucha "correlación" + mucha "varianza de y explicada" + mucha "varianza de x".

   Aproximadamente, un **intermedio** entre "PLS ortogonal $>\Sigma_x$ no importa$<$" y "PCR $>\Sigma_x$ lo es todo$<$".

3. PLS (ortogonalizado, con cambio $\Sigma_{\hat{x}} = I$): componentes no correlados de $\hat{x}$ cuyo estimado reduzca mucho la varianza del error... grosso modo, mucha "correlación" + mucha "varianza de y explicada".

Conclusiones

- EL SVD del modelo de predicción (regresión PLS) descompone las entradas en componentes no correlados según su grado de "*utilidad para predecir*" las salidas.
  - Permite determinar que, por ejemplo, el 88% de la covarianza entre un vector de 20 salidas y un conjunto de 150 entradas es explicado por 4 variables "*latentes*". Si se cogen "todos" los componentes, el resultado es el estimado estandard de mínimos cuadrados, pero, claro, eso no es para lo que PLS está concebido.
- Existen variaciones del concepto PLS según la sensibilidad o no al escalado de Σ_x.
- Al usar la **co**varianza entre x e y, suele **explicar más** varianza de y que la regresión PCR (donde la covarianza no se considera), para mismo número de componentes.