Partial Least Squares Regression - PLS

Analisi Statistica dei Dati Multidimensionali[1]

[1]Corso di Laurea in Scienze Statistiche e Attuariali

Facoltà di Scienze Economiche e Aziendali
Università degli Studi del Sannio

Modelli di base

- Si consideri il modello lineare in termini matriciali $\mathbf{Y}_{(n,c)} = \mathbf{X}_{(n,p)}\mathbf{B}_{(p,c)} + \mathbf{E}_{(n,c)}$ dove $\mathbf{B}$ contiene i c coefficienti di regressione e $\mathbf{E}$ termine residuale del modello
- $\mathbf{T} = \mathbf{XW}$ matrice degli score tramite la matrice dei pesi $\mathbf{W}$
- Si consideri inoltre il modello $\mathbf{Y} = \mathbf{TQ}^T + \mathbf{E}$ dove $\mathbf{Q}$ è una matrice di coefficienti di regressione per $\mathbf{T}$ ed $\mathbf{E}$ termine residuale.
- Una volta calcolate le $\mathbf{Q}$ il precedente modello equivale a $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ (con $\mathbf{B} = \mathbf{WQ}^T$) che può essere considerato come un modello regressivo predittivo.
- Una matrice necessaria per una descrizione completa del PLS è la matrice $\mathbf{P}$ dei fattori del modello $\mathbf{X} = \mathbf{TP}^T + \mathbf{F}$ con $\mathbf{F}$ matrice residuale

Notazione e criterio

- Indichiamo con $\mathbf{E}_0 = \mathbf{X}$ e $\mathbf{F}_0 = \mathbf{Y}$ le matrici centrate e normalizzate rispetto alla metrica dei pesi $\mathbf{D}$
- Sia allora A il numero di componenti prescelte $k = 1, \ldots, A$
- Siano $\mathbf{t} = \mathbf{E}_{k-1}\mathbf{w}$ e $\mathbf{u} = \mathbf{F}_{k-1}\mathbf{q}$ le combinazioni lineari colonne delle matrici centrate $\mathbf{E}_{k-1}$ e $\mathbf{F}_{k-1}$ associate rispettivamente ai vettori dei pesi $\mathbf{w}$ e $\mathbf{p}$
- La covarianza fra $\mathbf{t}$ e $\mathbf{u}$ si scrive come il prodotto scalare rispetto alla metrica dei pesi $\mathbf{D}$

$$cov(\mathbf{t}, \mathbf{u}) = (\mathbf{t}, \mathbf{u})_{\mathbf{D}} = \mathbf{w}^T \mathbf{E}_{k-1}^T \mathbf{D} \mathbf{F}_{k-1} \mathbf{q}$$

- $var(\mathbf{t}) = \|\mathbf{t}\|_{\mathbf{D}}^2 = \mathbf{t}^T \mathbf{D} \mathbf{t}$ norma quadratica rispetto alla metrica dei pesi $\mathbf{D}$

Passo k

- Il passo k dell'algoritmo si può separare concettualmente in due parti. La prima parte fornisce le componenti $\mathbf{t}_k = \mathbf{E}_{k-1}\mathbf{w}_k$ e $\mathbf{u}_k = \mathbf{F}_{k-1}\mathbf{q}_k$ attraverso i pesi ottimali $\mathbf{w}_k$ e $\mathbf{q}_k$. La seconda parte aggiorna la matrice dei predittori e delle risposte $\mathbf{E}_k$ e $\mathbf{F}_k$ come residui della regressione su $\mathbf{t}_k$

1. Calcolo dei pesi	$\arg\max_{(\mathbf{w}_k,\mathbf{q}_k)} cov(\mathbf{t},\mathbf{u}) = \mathbf{w}_k^T\mathbf{E}_{k-1}^T\mathbf{D}\mathbf{F}_{k-1}\mathbf{q}_k$
2. Calcolo dei residui	$\mathbf{E}_k = \mathbf{E}_{k-1} - \mathbf{P}_{\mathbf{t}_k}\mathbf{E}_{k-1}$ $\mathbf{F}_k = \mathbf{F}_{k-1} - \mathbf{P}_{\mathbf{t}_k}\mathbf{F}_{k-1}$

dove $\mathbf{P}_{\mathbf{t}_k} = \frac{\mathbf{t}_k\mathbf{t}_k^T\mathbf{D}}{var(\mathbf{t}_k)} = \frac{\mathbf{t}_k\mathbf{t}_k^T\mathbf{D}}{\mathbf{t}_k^T\mathbf{D}\mathbf{t}_k} = \mathbf{t}_k(\mathbf{t}_k^T\mathbf{D}\mathbf{t}_k)^{-1}\mathbf{t}_k^T\mathbf{D}$ è l'operatore di proiezione $\mathbf{D}$-ortogonale sul sottospazio di $\mathbf{t}_k$

Problema di ottimizzazione

- Le soluzione sono ottenute mediante il metodo dei moltiplicatori di Lagrange.
- La funzione Lagrangiana è definita come

$$L = \mathbf{w}^T \mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{q} - \frac{\lambda}{2}(\mathbf{w}^T \mathbf{w} - 1) - \frac{\mu}{2}(\mathbf{q}^T \mathbf{q} - 1)$$

- Differenziando L rispetto a $\mathbf{w}$ e $\mathbf{q}$ e uguagliando a zero il risultato, otteniamo il seguente sistema di equazioni normali

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{q} - \lambda \mathbf{w} = 0$$
$$\frac{\partial L}{\partial \mathbf{q}} = \mathbf{Y}_{k-1}^T \mathbf{D} \mathbf{X}_{k-1} \mathbf{w} - \mu \mathbf{q} = 0$$
$$\frac{\partial L}{\partial \lambda} = -\mathbf{w}^T \mathbf{w} + 1 = 0$$
$$\frac{\partial L}{\partial \mu} = -\mathbf{q}^T \mathbf{q} + 1 = 0$$

- Premoltiplicando ambo i membri della prima equazione per $\mathbf{w}^T$ e quelli della seconda per $\mathbf{q}^T$ otteniamo rispettivamente

$$\mathbf{w}^T\mathbf{X}_{k-1}^T\mathbf{D}\mathbf{Y}_{k-1}\mathbf{q} - \lambda\mathbf{w}^T\mathbf{w} = 0$$
$$\mathbf{q}^T\mathbf{Y}_{k-1}^T\mathbf{D}\mathbf{X}_{k-1}\mathbf{w} - \mu\mathbf{q}^T\mathbf{q} = 0$$

e poichè $\mathbf{w}^T\mathbf{w} = \mathbf{q}^T\mathbf{q} = 1$, abbiamo

$$\mathbf{w}^T\mathbf{X}_{k-1}^T\mathbf{D}\mathbf{Y}_{k-1}\mathbf{q} = \lambda$$
$$\mathbf{q}^T\mathbf{Y}_{k-1}^T\mathbf{D}\mathbf{X}_{k-1}\mathbf{w} = \mu$$

da cui risulta che $\lambda = \mu$ che risulta essere l'ottimo ricercato.

- Premoltiplicando ambo i membri della prima equazione per μ ottenendo

$$\mu \mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{q} - \mu \lambda \mathbf{w} = 0$$
$$\mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mu \mathbf{q} - \lambda^2 \mathbf{w} = 0$$

dalla seconda equazione abbiamo che

$$\mu \mathbf{q} = \mathbf{Y}_{k-1}^T \mathbf{D} \mathbf{X}_{k-1} \mathbf{w}$$

da cui sostituendo nella precedente

$$\mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{Y}_{k-1}^T \mathbf{D} \mathbf{X}_{k-1} \mathbf{w} = \lambda^2 \mathbf{w}$$

- In modo analogo, premoltiplicando ambo i membri della seconda equazione per λ e considerando dalla prima equazione sappiamo che

$$\lambda \mathbf{w} = \mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{q}$$

otteniamo

$$\mathbf{Y}_{k-1}^T \mathbf{D} \mathbf{X}_{k-1} \mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{q} = \lambda^2 \mathbf{q}$$

quindi i vettori soluzione $\mathbf{w}$ e $\mathbf{q}$ saranno rispettivamente quelli associati all'autovalore dominante λ^2 soluzione delle quantità

$$\boxed{\begin{aligned} \mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{Y}_{k-1}^T \mathbf{D} \mathbf{X}_{k-1} \mathbf{w} &= \lambda^2 \mathbf{w} \\ \mathbf{Y}_{k-1}^T \mathbf{D} \mathbf{X}_{k-1} \mathbf{X}_{k-1}^T \mathbf{D} \mathbf{Y}_{k-1} \mathbf{q} &= \lambda^2 \mathbf{q} \end{aligned}}$$

Partial Least Squares Regression: Algoritmo - No missing value

Step 1

- Indichiamo con $\mathbf{E}_0 = \mathbf{X}$ e $\mathbf{F}_0 = \mathbf{Y}$ le matrici centrate e normalizzate rispetto alla metrica dei pesi $\mathbf{D}$ e $k = 1$;
- Sia A il numero di componenti prescelte $k = 1, \ldots, A$
- Calcoliamo i vettori soluzione $\mathbf{w}$ e $\mathbf{q}$ soluzioni dominanti delle quantità

$$\begin{aligned} \mathbf{X}^T\mathbf{D}\mathbf{Y}\mathbf{Y}^T\mathbf{D}\mathbf{X}\mathbf{w} &= \lambda^2\mathbf{w} \\ \mathbf{Y}^T\mathbf{D}\mathbf{X}\mathbf{X}^T\mathbf{D}\mathbf{Y}\mathbf{q} &= \lambda^2\mathbf{q} \end{aligned}$$

- Calcoliamo gli score $\mathbf{t}_1 = \mathbf{X}\mathbf{w}$;
- Sia $\mathbf{T}_1 = [\mathbf{t}_1]$ e $\mathbf{P}_{\mathbf{T}_1} = \mathbf{T}_1(\mathbf{T}_1^T\mathbf{T}_1)^{-1}\mathbf{T}_1^T$ l'operatore di proiezione ortogonale sul sottospazio $Im(\mathbf{T}_1)$
- Sia $k = 2$;

Partial Least Squares Regression: Algoritmo

Step 2

- Si procede a calcolare le matrici residuali

$$\mathbf{X}_k = \mathbf{X} - \mathbf{P}_{\mathbf{T}_k}\mathbf{X} = (\mathbf{I} - \mathbf{P}_{\mathbf{T}_k})\mathbf{X} = \mathbf{P}^{\perp}_{\mathbf{T}_k}\mathbf{X}$$
$$\mathbf{Y}_k = \mathbf{Y} - \mathbf{P}_{\mathbf{T}_k}\mathbf{Y} = (\mathbf{I} - \mathbf{P}_{\mathbf{T}_k})\mathbf{Y} = \mathbf{P}^{\perp}_{\mathbf{T}_k}\mathbf{Y}$$

- Calcoliamo i vettori soluzione **w** e **q** soluzioni dominanti rispettivamente delle quantità

$$\mathbf{X}_k^T\mathbf{D}\mathbf{Y}_k\mathbf{Y}_k^T\mathbf{D}\mathbf{X}_k\mathbf{w} = \lambda^2\mathbf{w}$$
$$\mathbf{Y}_k^T\mathbf{D}\mathbf{X}_k\mathbf{X}_k^T\mathbf{D}\mathbf{Y}_k\mathbf{q} = \lambda^2\mathbf{q}$$

N.B.: da un punto di vista di efficienza computazionale si diagonalizza la matrice più piccola e si usano le formule di transizione per ottenere l'altra soluzione

Partial Least Squares Regression: Algoritmo

Step 3

- Calcoliamo i nuovi score $\mathbf{t}_k = \mathbf{X}_k \mathbf{w}_k$;
- Aggiorniamo la matrice degli scores giustapponendo per colonna i nuovi score $\mathbf{T}_k = [\mathbf{t}_1, ..., \mathbf{t}_k]$;
- Calcoliamo il nuovo proiettore $\mathbf{P}_{\mathbf{T}_k}$;
- Incrementiamo l'indice k: $k = k + 1$;
- Riprendi dallo Step 2 fino a quando sia vera la condizione $k \leq A$.

Proprietà delle componenti

- Le formule d'attualizzazione delle variabili conducono alla relazione: $< \mathbf{t}_k, \mathbf{t}_l >_{\mathbf{D}} = < \mathbf{t}_k, \mathbf{w}_l >_{\mathbf{D}} = 0, \ \forall l > k$
- Si dimostra per ricorrenza che $\mathbf{t}_k = \mathbf{X}\mathbf{w}_k^*$ appartiene a $Im(\mathbf{X})$ spazio vettoriale generato dai predittori. Più precisamente i coefficienti $\mathbf{w}_k^*$ saranno dati dalle relazioni:

$$\mathbf{w}_1^* = \mathbf{w}_1$$

$$\mathbf{w}_k^* = [\mathbf{I}_p - \sum_{j=1}^{k-1} \frac{\mathbf{w}_j^* \mathbf{w}_j^{*T}}{\|\mathbf{t}_j\|_{\mathbf{D}}^2} \mathbf{X}^T \mathbf{D} \mathbf{X}] \mathbf{w}_k, \ \forall k > l$$

- In termini matriciali la relazione fra $\mathbf{w}_k^*$ e $\mathbf{w}_k$ è pari a

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1}$$

Proprietà delle componenti

- Gli autovettori **W** sono ortonormali;
- Le componenti **T** sono ortogonali;
- I loadings **P** non sono ortogonali;
- Le componenti **U** non sono ortogonali;
- Le componenti **T** e **U** sono ortogonali fra loro

$$\mathbf{u}_b^T \mathbf{t}_a = 0 \text{ se } b > a$$

- I coefficienti **W** e **P** sono ortogonali fra loro

$$\boxed{\mathbf{p}_b^T \mathbf{w}_a = 0 \text{ se } b > a} \text{ e } \boxed{\mathbf{p}_b^T \mathbf{w}_a = 1 \text{ se } b = a}$$

tale che $\mathbf{P}^T\mathbf{W}$ risulta essere una matrice triangolare superiore

Proprietà delle componenti

- La non correlazione delle componenti implica che l'operatore di proiezione ortogonale $\mathbf{P}_{\mathbf{T}_A}$ sul sottospazio generato da $\mathbf{T}_A = [\mathbf{t}_1, \ldots, \mathbf{t}_A]$ può scriversi

$$\mathbf{P}_{\mathbf{T}_A} = \sum_{k=1}^{A} \mathbf{P}_{\mathbf{t}_k}$$

  che può essere anche visto come l'operatore di proiezione ortogonale su $Im(\mathbf{X})$ generato dalle componenti $\mathbf{t}_1, \ldots, \mathbf{t}_A$
- Nel caso particolare che $A = rang(\mathbf{X})$ allora $\mathbf{P}_{\mathbf{T}_A} = \mathbf{P}_{\mathbf{X}}$.

I modelli PLS

- L'attualizzazione delle variabili e la non correlazione delle componenti ci consentono di scrivere più semplicemente i modelli parziali: $\hat{\mathbf{X}}_k = \mathbf{P}_{\mathbf{t}_k}\mathbf{X}$ e $\mathbf{Y}_k = \mathbf{P}_{\mathbf{t}_k}\mathbf{Y}$
- La non correlazione delle componenti consente inoltre di decomporre la varianza totale:

$$Var(\mathbf{Y}) = \sum_{j=1}^{c} Var(\mathbf{Y}^j) = \sum_{k=1}^{A} Var(\hat{\mathbf{Y}}_k) + Var(\mathbf{F}_A)$$

come anche scrivere in modo definitivo i modelli PLS in funzione delle componenti

$$\hat{\mathbf{Y}}_A = \mathbf{P}_{\mathbf{T}_A}\mathbf{Y} \text{ e } \hat{\mathbf{X}}_A = \mathbf{P}_{\mathbf{T}_A}\mathbf{X}$$

I modelli PLS

- L'operatore di proiezione ortogonale $\mathbf{P}_{\mathbf{T}_A}$ sul sottospazio generato da $\mathbf{T}_A = [\mathbf{t}_1, \ldots, \mathbf{t}_A]$ può essere anche scritto

$$\mathbf{P}_{\mathbf{T}_A} = \sum_{k=1}^{A} \mathbf{X} \frac{\mathbf{w}_k^* \mathbf{w}_k^{*T}}{\|\mathbf{t}_k\|_{\mathbf{D}}^2} \mathbf{X}^T \mathbf{D}$$

che implica che il modello PLS è lineare rispetto alle variabili predittrici iniziali

$$\hat{\mathbf{Y}}_A = \mathbf{X} \hat{\beta}_A$$

con

$$\hat{\beta}_A = \sum_{k=1}^{A} \frac{\mathbf{w}_k^* \mathbf{w}_k^{*T}}{\|\mathbf{t}_k\|_{\mathbf{D}}^2} \mathbf{X}^T \mathbf{D} \mathbf{Y}$$

I modelli PLS

- Se $A = rang(\mathbf{X})$ allora PLS($\mathbf{X}, \mathbf{Y}$)=OLS($\mathbf{X}, \mathbf{Y}$);
- PLS($\mathbf{X}, \mathbf{X}$)=ACP($\mathbf{X}$);

- Il modello predittivo si può scrivere

$$\begin{aligned}\mathbf{Y} &= \mathbf{T}\mathbf{Q}^T + \mathbf{E} \\ &= \mathbf{X}\mathbf{W}^*\mathbf{Q}^T + \mathbf{E} \\ &= \mathbf{X}\underbrace{\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T}_{=\mathbf{B}^{\text{PLSR}}} + \mathbf{E} \\ &= \mathbf{X}\mathbf{B}^{\text{PLSR}} + \mathbf{E}\end{aligned}$$

I modelli PLS

A sua volta $\mathbf{B}^{\text{PLSR}}$ può essere scritto

$$\boxed{\mathbf{B}^{\text{PLSR}}} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T$$

$$= \mathbf{W}^*\mathbf{Q}^T$$

$$= \mathbf{W}^*(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{Y}$$

$$= \mathbf{W}^*(\mathbf{W}^{*T}\mathbf{X}^T\mathbf{X}\mathbf{W}^*)^{-1}\mathbf{W}^{*T}\mathbf{X}^T\mathbf{Y}$$

$$= \mathbf{W}^* \underbrace{(\mathbf{W}^{*T}\mathbf{X}^T\mathbf{X}\mathbf{W}^*)^{-1}\mathbf{W}^{*T}\mathbf{X}^T\mathbf{X}}_{=\mathbf{P}^T} \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}_{=\mathbf{B}^{\text{OLS}}} = \boxed{\mathbf{W}^*\mathbf{P}^T\mathbf{B}^{\text{OLS}}}$$

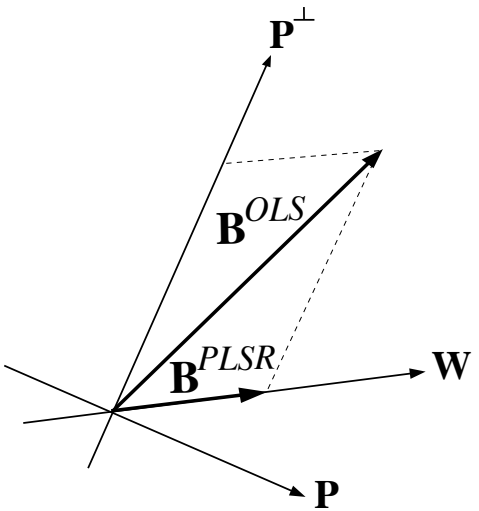
$$= \underbrace{\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{P}^T}_{=\mathbf{H}^A_{\mathbf{WP}^\perp}}\mathbf{B}^{\text{OLS}}$$

$$= \boxed{\mathbf{H}^A_{\mathbf{WP}^\perp}\mathbf{B}^{\text{OLS}}}$$

$$\boxed{\boxed{\mathbf{H}^A_{\mathbf{WP}^\perp} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{P}^T}}$$

I modelli PLS

- La matrice $\mathbf{H}^{A}_{\mathbf{WP}^{\perp}} = \mathbf{W}(\mathbf{P}^{T}\mathbf{W})^{-1}\mathbf{P}^{T}$ è idempotente ma non è simmetrica ed è quindi un operatore di proiezione obliquo. Una proiezione obliqua è una proiezione su un sottospazio ma non di tipo ortogonale rispetto a questo spazio, bensì lungo qualche altra direzione.
- La matrice $\mathbf{H}^{A}_{\mathbf{WP}^{\perp}}$ è quindi un operatore di proiezione obliquo sul sottospazio generato dalle colonne di **W** lungo la direzione ortogonale a **P**.



I modelli PLS

- Si può inoltre dimostrare (de Jong, 1995; Goutis, 1996) che

$$\|\mathbf{B}^{\text{PLSR}}\| = \|\mathbf{H}^{A}_{\mathbf{WP}^{\perp}}\mathbf{B}^{\text{OLS}}\| \leqslant \|\mathbf{B}^{\text{OLS}}\|$$