STATISTICA, anno LXII, n. 2, 2002

LA STIMA DI VARIABILI LATENTI DA VARIABILI OSSERVATE MISTE

Pietro Giorgio Lovaglio

1. INTRODUZIONE

In molti problemi di natura socioeconomica è di primario interesse per il ricercatore un indicatore che sintetizzi una performance o uno stato di benessere globale per ogni unità statistica (paziente, utente, azienda etc). Spesso la natura di tale indicatore è tipicamente latente, cioè viene concepito come un costrutto teorico inosservabile sottostante ad una serie di variabili manifeste (MV) e solo stimabile a partire da esse (Gertler, 1988; Fallowfield, 1996).

Lo scopo del presente lavoro è quello di fornire una metodologia per la stima dei punteggi di variabili latenti (LV), supportate da indicatori qualitativi e quantitativi.

La stima di costrutti latenti viene solitamente ottenuta (Sheehan, 1991; Romney *et al*. 1992) con LISREL (Joreskog, 1978), tuttavia la definizione di LV adottata da LISREL (Bentler, 1982) o da EQS (Bentler, 1982) nel modello fattoriale lascia indeterminati i punteggi latenti (Vittadini, 1999).

Il modello di analisi fattoriale,

$$\mathbf{Y} = \boldsymbol{\eta}\mathbf{B} + \mathbf{E} \tag{1}$$

dove $\mathbf{B}$ è la matrice $(r \times q)$ di coefficienti strutturali, oltre al problema di indeterminatezza dei punteggi latenti nella matrice $\boldsymbol{\eta}$ $(n \times r)$, ha il forte limite di escludere dall'analisi variabili (cause) che generano una LV, operando solamente con variabili di tipo-effetto (come colonne di $\mathbf{Y}$ $(n \times q)$): ad esempio per stimare i punteggi della LV *capitale umano* (Dagum e Vittadini, 1996) si dovrebbe rinunciare a quelle variabili (titolo di studio, anni di lavoro etc.) da cui il capitale umano è generato, non potendo tali indicatori essere assimilati a effetti del capitale umano.

2. VARIABILI LATENTI UNICHE E STIMA DEL MODELLO DI MISURA

La soluzione del problema della non unicità dei punteggi latenti ha giustificato in letteratura una nuova definizione di LV (RCD, Schonemann e Steiger, 1976;

RCDR, Haagen e Vittadini, 1991) come combinazione lineare ($\boldsymbol{\eta}$=**YA**) di una serie di q indicatori (**A** è di dimensione ($q \times r$)), e stimata, in un primo stadio, attraverso l'analisi in Componenti Principali (CP), la Correlazione Canonica (CC), il Partial Least Squares (PLS, Wold 1982) o con tecniche di trasformazione non lineare attraverso l'Analisi delle Corrispondenze (MCA), le Componenti principali non lineari (NPCA, Young *et al.* 1978) e la Rasch Analysis (Vittadini, 1997). Nel secondo stadio le stime di tali LV vengono inserite nel modello fattoriale originariamente ipotizzato dal ricercatore per stimare i parametri causali tra LV e MV.

Tale *modus operandi* ha le seguenti limitazioni:

(*i*) inefficienza: la stima dei punteggi delle LV avviene separatamente dalla stima dei parametri del modello di misura specificato nella (1), senza definire un criterio di ottimo globale nei due stadi;

(*ii*) inconsistenza logica: la stima delle LV si ottiene con tecniche che invertono il legame causa-effetto tra LV ed indicatori manifesti, rispetto alla specificazione classica del modello fattoriale in cui le MV nella (1) sono effetti di $\boldsymbol{\eta}$.

PLS stima i punteggi di due LV (vettori) $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, (unidimensionali) nel modello specificato, coerente con il path diagram del modello LISREL di figura1,

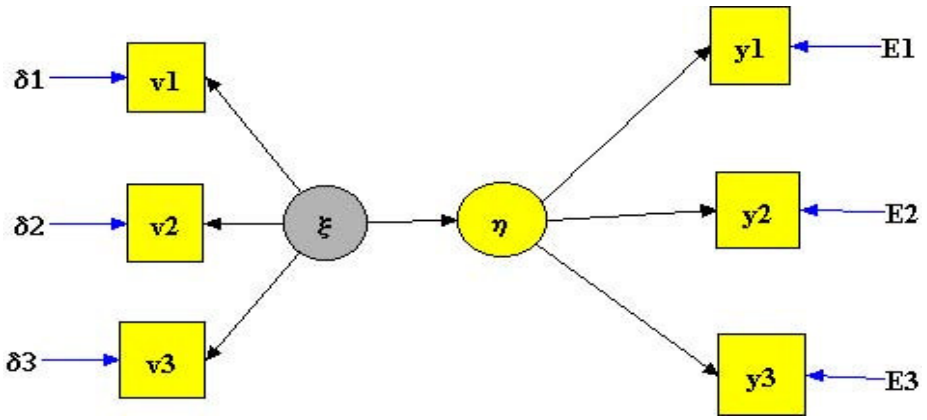

Figura 1 – Path diagram di LISREL e PLS, con $\boldsymbol{\eta}$ e $\boldsymbol{\xi}$ unidimensionali.

definendole come combinazioni $\boldsymbol{\eta}$= **Ya**, $\boldsymbol{\xi}$ = **Vc**, attraverso una procedura iterativa che alterna due fasi: 1) stima iniziale di $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ ($\boldsymbol{\xi}_0$, $\boldsymbol{\eta}_0$) come componenti principali dei propri indicatori; 2) si aggiornano i coefficienti (**a, c**) attraverso metodologie chiamate Mode (A,A), Mode (B,B), Mode (A,B) e Mode (B,A).

Il Mode (B,B) specifica due modelli di regressione: di $\boldsymbol{\eta}_0$ su $\mathbf{v}_1$ $\mathbf{v}_2$ $\mathbf{v}_3$ e di $\boldsymbol{\xi}_0$ su $\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_3$, secondo il path diagram di figura 2

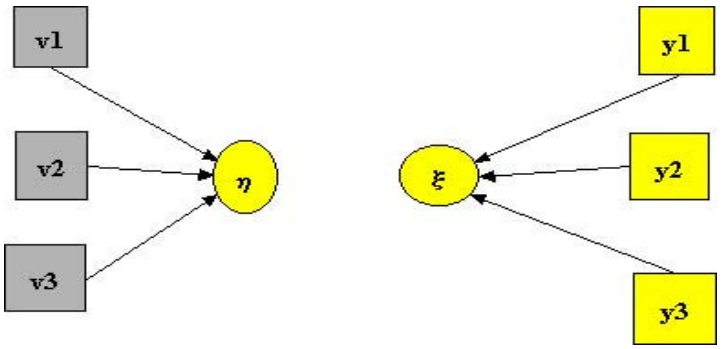

Figura 2 – Path diagram del Mode (B,B) di PLS.

Alternando i passi 1) e 2) si ottengono i punteggi di $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ attraverso le stime dei pesi (**a, c**).

Il Mode (A,A) invece aggiorna i pesi (**a, c**) attraverso regressioni univariate di $\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_3$ su $\boldsymbol{\xi}_0$ e di $\mathbf{v}_1$ $\mathbf{v}_2$ $\mathbf{v}_3$ su $\boldsymbol{\eta}_0$, il Mode (B,A) e il Mode (A,B) costituiscono soluzioni intermedie.

Tale metodologia illustrata in figura 2, non è coerente con i modelli fattoriali specificati per $\boldsymbol{\xi}$ ed $\boldsymbol{\eta}$ (figura 1), perchè:

(*i*) la stima iniziale di $\boldsymbol{\eta}$ ed $\boldsymbol{\xi}$ ($\boldsymbol{\eta}_0$ $\boldsymbol{\xi}_0$) si ottiene come componente principale degli indicatori del proprio blocco, rendendo ogni variabile manifesta *causa* della rispettiva LV, non rispettando invece il ruolo degli indicatori come effetti (figura 1);

(*ii*) Mode (B,B) aggiorna i punteggi delle LV trattando le variabili ($\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_3$) che costituiscono gli effetti di $\boldsymbol{\eta}$, (figura 1), come variabili che generano $\boldsymbol{\xi}$;

(*iii*) la stima dei punteggi latenti ottenuti attraverso il Mode (A,A) è ancora in contraddizione con la specificazione di figura 1;

(*iv*) Wold (1982) dimostra che il Mode (B,B) perviene alla stima dei punteggi delle due LV, massimizzando la correlazione canonica tra i due insiemi di variabili; tuttavia come è noto (Stewart e Love, 1968) una correlazione canonica elevata non necessariamente implica che le variabili canoniche catturino una significativa porzione della variabilità del proprio insieme.

Gli autori della Regression Component Decomposition (RCD) suppongono invece che le MV siano contemporaneamente sia effetti (secondo la logica del modello fattoriale) sia anche cause della LV, in modo da poter utilizzare la definizione di LV come combinazione dei suoi indicatori, introducendo così una ambiguità del ruolo degli indicatori manifesti.

La definizione di LV adottata, che utilizza la stessa informazione campionaria del modello MIMIC (Joreskog e Goldberger, 1975) risultandone però più flessibile (vedi paragrafo 5), consiste nella specificazione di una LV $\boldsymbol{\eta}$ (es. capitale umano) come una combinazione di variabili cause **X** che generano tale LV (es. anni di scolarità, anni di esperienza professionale, anni di disoccupazione) e che contemporaneamente "best predicts" un insieme di variabili manifeste **Y** (es. effetti monetari e professionali del capitale umano (Dagum, 1994)) che descrivo-

no gli effetti di tale LV: in tale modo si sfrutta appieno l'unica informazione disponibile per una LV cioè il ruolo che essa svolge nel veicolare le informazioni tra due insiemi di variabili nel modello.

Tale concezione di LV ci pone lontano dalla definizione classica di variabile latente come costrutto teorico inosservabile, tipico dell'analisi fattoriale, ma piuttosto identifica una LV come combinazione lineare (e non lineari per dati categoriali) delle variabili manifeste, come negli approcci della RCD, RCDR o del PLS.

La stima dei punteggi di LV in tale ambito deve tener conto anche della correlazione tra gli errori delle variabili dipendenti $(\mathbf{y}_1 \ldots . \mathbf{y}_q)=\mathbf{Y}$, per evitare distorsioni nei coefficienti di regressione, perdita di efficienza delle stime degli standard error e intervalli di confidenza non conservativi (Fitzmaurice e Laird, 1997; Liang *et al.*, 1992).

3. IL MODELLO PER LA STIMA DI LV E PARAMETRI STRUTTURALI

Con p variabili-cause $\mathbf{X}$ $(n\times p)$, e q indicatori-effetto $\mathbf{Y}$ $(n\times q)$ continui su n osservazioni il modello più ragionevole è:

$$\boldsymbol{Y} = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{XM} + \boldsymbol{E} \qquad r=\rho(\mathbf{M}) \leq \min(p,q) \qquad \mathrm{Vec}(\mathbf{E}) \sim (\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n) \tag{2}$$

con $\mathbf{1}_n$ un vettore unitario n-dimensionale, $\mathbf{M}$ una matrice $(p\times q)$ di coefficienti di rango r, $\rho(r)$, $\boldsymbol{\mu}$ il vettore q-dimensionale di medie di $\mathbf{Y}$, Vec($\mathbf{E}$) l'operatore che impila le colonne di $\mathbf{E}$.

Attraverso una Singular Value Decomposition (SVD) su $\mathbf{M}=\mathbf{AB}$, la (2) è equivalente a:

$$\boldsymbol{Y} = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{XAB} + \boldsymbol{E} \qquad \text{con} \qquad \boldsymbol{A'X'XA}=\boldsymbol{I}_r \tag{3}$$

$$\mathbf{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{\eta}\ \mathbf{B} + \mathbf{E} \qquad \rho(\mathbf{A})=r$$

La (3), il cui vincolo assicura l'unicità di $\mathbf{A}$ $(p\times r)$ nella SVD, contrariamente agli approcci che operano in due stadi (PLS, Componenti Principali, RCD), costituisce un modello sul quale è possibile stimare simultaneamente sia pesi $\mathbf{A}$ per definire la LV $\boldsymbol{\eta}=\mathbf{XA}$, ($r$ componenti ortonormali) come combinazione lineare di $\mathbf{X}$ (per evitare i problemi di indeterminatezza) sia la matrice $\mathbf{B}$ $(r\times q)$ dei pesi di regressione tra LV ed indicatori $\mathbf{Y}$ (parametri strutturali), tenendo conto della struttura di associazione tra indicatori dipendenti $\mathbf{y}_s$, evidenziata dalla matrice $\boldsymbol{\Sigma}$ degli errori di dimensione $(q\times q)$ nella (1).

Il significato dei parametri della (3), mostrata nel path diagram di figura 3, coincide con quello del modello fattoriale (1), risolvendo il problema dell'unicità dei punteggi latenti senza rinunciare alla stima dei legami causali tra LV e variabili manifeste.

Inoltre la scelta della dimensione di $\boldsymbol{\eta}$, attraverso il vincolo di rango (ρ) su $\mathbf{A}$, le possibili specificazioni per $\boldsymbol{\Sigma}$ (non strutturata, nota, diagonale omoschedastica, diagonale eteroschedastica) e l'approccio distribution free degli errori, costituisco-

no ulteriori aspetti di flessibilità del modello che numerosi approcci in letteratura non prevedono, tra cui LISREL, MIMIC e la Rasch Analysis

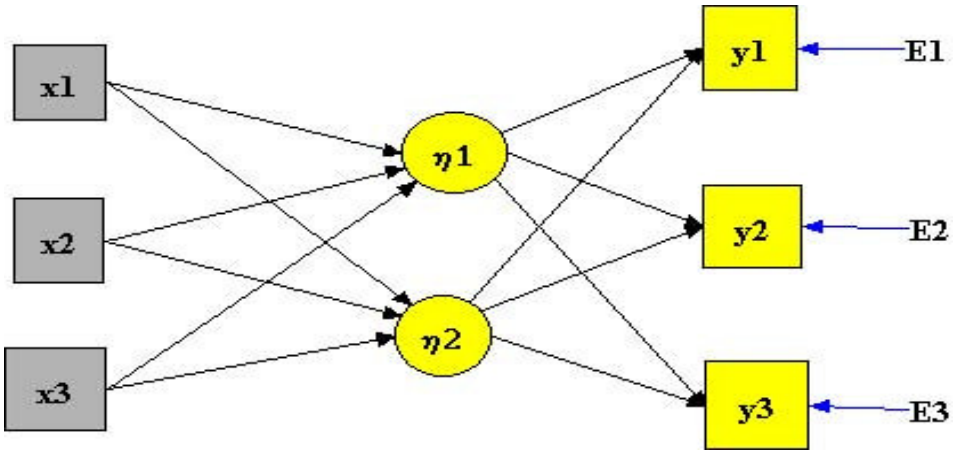

Figura 3 – Path diagram associato alla (2)

La stima della (3) rientra nell'ambito dei modelli di regressione multivariata con rango ridotto (RRR, Tso, 1981), ma non è stata estesa in letteratura a MV di tipo misto (qualitative e quantitative).

Il modello (3) con r noto, con $\mathbf{A'X'XA=I}_r$ e $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, si può ricondurre ad un modello equivalente (per la dimostrazione si veda Israels, 1984) specificato nella (4)

$$\mathbf{YC} = \mathbf{1}_n\boldsymbol{\mu}' + \mathbf{XB} + \mathbf{E} \quad \mathbf{C'C=I}_r \quad \rho(\mathbf{C})=r \quad \text{Vec}(\mathbf{E}) \sim (\mathbf{0}, \sigma^2\mathbf{I}_{nq}) \tag{4}$$

in cui la LV $\boldsymbol{\eta}_j$ (j-esima colonna di $\boldsymbol{\eta}$, con j=1,..r) coincide con la j-esima componente principale di $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$, dove $\hat{\mathbf{Y}}$ è la matrice che presenta nelle colonne i q vettori previsti attraverso le regressioni OLS per ogni $\mathbf{y}_s$.

Le quantità $\lambda_j\ c_{sj}^2$ ed $\Sigma_s^q\ \lambda_j\ c_{sj}^2$ misurano rispettivamente la varianza spiegata dell's-esimo (s=1,...q) indicatore ($\mathbf{y}_s$) attraverso la j-esima LV e la varianza totale di $\mathbf{Y}$ (tr$\mathbf{Y}$) spiegata dalla j-esima LV, con λ_j il j-esimo maggior autovalore di $\mathbf{Y'X(X'X)^{-1}X'Y}$.

4. GENERALIZZAZIONE E ALGORITMO PER DATI MISTI

Il caso più frequente nelle applicazioni economico-sociali riguarda il caso di variabili categoriali (nominali e/o ordinali) nelle matrici $\mathbf{X}$, $\mathbf{Y}$ o in entrambe.

In letteratura non esistono proposte[1] che estendono la metodologia descritta nella (3) per dati misti; in questo lavoro si estende la metodologia per LV suppor-

[1] Alcune proposte (Keller e Wansbeek, 1983; Israels,1984) non evidenziano differenza tra variabili nominali ed ordinali: le stime delle categorie per le variabili ordinali non rispettano il vincolo di rango.

tate da indicatori misti attraverso la metodologia dell'optimal scaling (O.S.) nel caso di $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$.

L'O.S., applicato all'interno di un modello statistico (Kruskal, 1965), stima i parametri strutturali del modello e simultaneamente una trasformazione ottima (non necessariamente lineare) delle MV categoriali presenti nell'analisi, massimizzando un criterio di ottimo (imposto dal modello), rispettando il livello di misurazione (quantitativa, nominale, ordinale) di ogni MV nell'analisi.

Siano $\mathbf{X}$ $(n\times p)$ ed $\mathbf{Y}$ $(n\times q)$ matrici di MV categoriali (o quantitative) di k_i e k_s categorie ognuna, rappresentabili senza perdita di generalità attraverso:

$$\mathbf{X} = (\mathbf{G}_1\boldsymbol{\omega}_1,\ldots \mathbf{G}_i\boldsymbol{\omega}_i.\ldots\mathbf{G}_p\boldsymbol{\omega}_p)' = \mathbf{G}^1\boldsymbol{\omega}^1 , \quad i= 1\ldots p \tag{5}$$

$$\mathbf{Y} = (\mathbf{G}_1\boldsymbol{\omega}_1,\ldots. \mathbf{G}_s\boldsymbol{\omega}_s \ldots\mathbf{G}_q\boldsymbol{\omega}_q)' = \mathbf{G}^2\boldsymbol{\omega}^2 , \quad s= 1\ldots.q$$

dove $\mathbf{G}_i$ e $\mathbf{G}_s$ sono le matrici indicatore delle variabili presenti in $\mathbf{X}$ ed $\mathbf{Y}$, $\mathbf{G}^1$ $\mathbf{G}^2$ le matrici indicatori complete $(n\times\Sigma_i^p\ k_i)$ e $(n\times\Sigma_s^q\ k_s)$, $\boldsymbol{\omega}^1$ ed $\boldsymbol{\omega}^2$ i vettori di dimensione $\Sigma_i^p\ k_i$ e $\Sigma_s^q\ k_s$ formati incolonnando i vettori di scaling $\boldsymbol{\omega}_i$ $(k_i\times 1)$ $\boldsymbol{\omega}_s$ $(k_s\times 1)$ di ogni variabile:

$$\boldsymbol{\omega}^1 = (\boldsymbol{\omega}_1', .\ \boldsymbol{\omega}_i'\ldots.\boldsymbol{\omega}_p')' \quad , \quad \boldsymbol{\omega}^2 = (\boldsymbol{\omega}_1', ..\ \boldsymbol{\omega}_s'\ \ldots\boldsymbol{\omega}_q')'$$

Se una variabile è continua i vettori di scaling $\boldsymbol{\omega}_i$ $\boldsymbol{\omega}_s$ coincidono con i punteggi osservati, mentre se è ordinale va inserito un vincolo di rango sulle k_s componenti (omettendo l'indice s) del vettore di scaling $\boldsymbol{\omega}_s$:

$$\omega_1 \leq\ldots\leq\ \omega_i.\leq\ldots..\leq \omega_k$$

Per dati categoriali e misti la metodologia descritta nel paragrafo 3 non è generalizzabile poiché la matrice indicatore completa $\mathbf{G}^{1\prime}\mathbf{G}^1$ associata ad $\mathbf{X}$ non è invertibile (Takeuki *et al.*, 1982) rendendo difficoltose le q regressioni OLS, che costituiscono generalmente il primo stadio per stimare i punteggi di $\boldsymbol{\eta}$, oltre al fatto che il livello di misurazione (continuo, ordinale, nominale) va specificato separatamente per ogni MV; tali complicazioni vengono facilmente risolte utilizzando un algoritmo iterativo che ammette nell'i-esima iterazione (mostrata in figura 5) quattro stadi.

Tra essi si utilizza un Algoritmo per la Regressione Multipla (Lovaglio, 1997) e un Algoritmo per la NPCA con dati misti (Young *et al.*, 1978) che permettono la specificazione della scala di misurazione di ogni variabile e non richiedono l'inversione di $\mathbf{G}^{1\prime}\mathbf{G}^1$.

L'algoritmo, appartenente alla famiglia Alternating Least Squares (ALS, Young *et al.*, 1978) stima i parametri della (4), per una LV unidimensionale, nella generica iterazione attraverso i seguenti stadi:

1) per ogni $\mathbf{y}_s$ si specificano q separati modelli di regressione con variabili miste (Lovaglio, 1997) ottenendo i punteggi delle p variabili esplicative $\mathbf{X}_s^{os}$ (ottima-

mente trasformati) in ognuna[2] delle q-equazioni (s=1,…,q) e le quantificazioni (ottime) delle q variabili dipendenti $\mathbf{y}_1^{os},\ldots,\mathbf{y}_q^{os}$ che coincidono con le previsioni ottenute dalle regressioni con dati $\hat{\mathbf{Y}} = (\hat{\mathbf{y}}_1.....\hat{\mathbf{y}}_q)$;

2) la prima C.P. $\hat{\mathbf{Y}}\mathbf{c}_1^*$ di $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ costituisce la stima iniziale di $\boldsymbol{\eta}$ ($\mathbf{c}_1^*$ è il primo autovettore di $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$);

3) attraverso una regressione con $\hat{\mathbf{Y}}\mathbf{c}_1^*$, variabile dipendente quantitativa, e i p regressori misti (quantitativi o categoriali) si ottiene il vettore dei coefficienti di regressione $\mathbf{b}$ e le quantificazioni dei p regressori (colonne della matrice $\mathbf{X}^{os}$);

4) attraverso la NPCA si ottengono le stime aggiornate di $\mathbf{y}_1^{os},\ldots,\mathbf{y}_q^{os}$ proiettando $\hat{\mathbf{Y}}\mathbf{c}_1^*$ sullo spazio delle colonne delle matrici indicatori associate ad ogni variabile dipendente[3],

$$\mathbf{y}_s^{os} = \mathbf{G}_s \mathbf{D}_s^{-1}\mathbf{G}_s' \hat{\mathbf{Y}}\mathbf{c}_1^* \tag{6}$$

Con i vettori $\mathbf{y}_s^{os}$ e le matrici $\mathbf{X}_s^{os}$, ottenute nel primo passo, si ritorna al punto 1) iniziando una nuova iterazione dell'algoritmo. La procedura si estende facilmente al caso di $\boldsymbol{\eta}$ r-dimensionale.

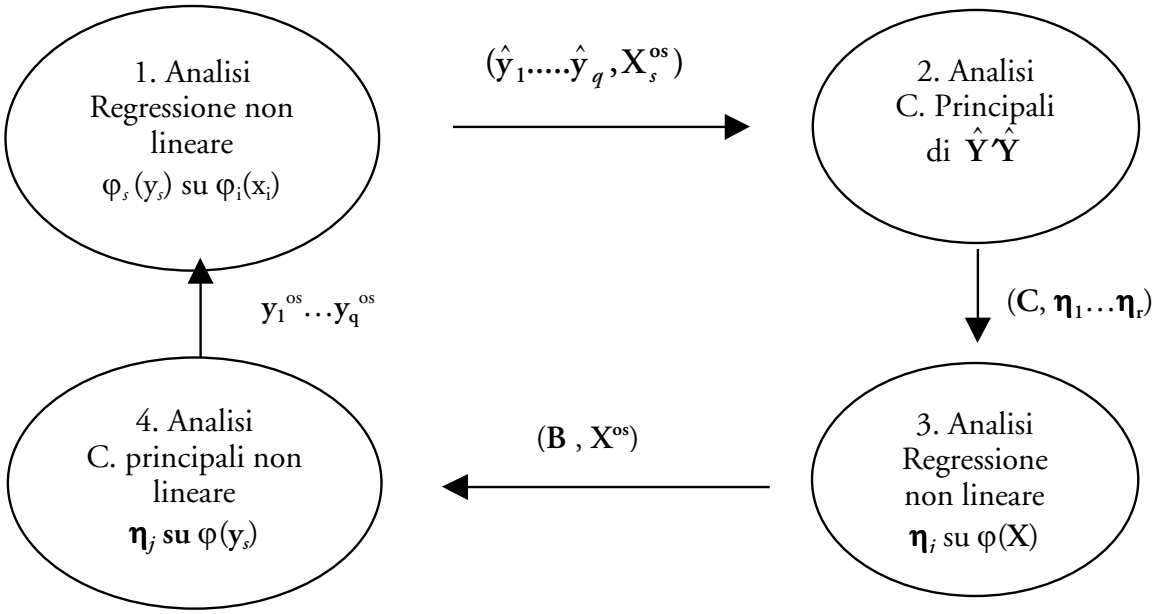

Figura 4 – Diagramma di flusso dell'algoritmo: l'i-esima iterazione è composta di 4 stadi (evidenziati nelle ellissi); le variabili presso le frecce costituiscono gli output di ogni stadio. $\varphi(.)$ è la trasformazione ottimale per la variabile in parentesi.

[2] Si noti che l'optimal scaling applicato su q equazioni di regressione che hanno stessi regressori X ma diversa variabile dipendente restituisce q differenti quantificazioni ottimali di ogni variabile indipendente.

[3] Se la variabile j è quantitativa si salta tale passo.

5. CONCLUSIONE

La metodologia proposta perviene alla stima simultanea dei punteggi (unici) di una LV e dei coefficienti strutturali tra LV e variabili manifeste, in modo più efficiente delle altre proposte che operano in due stadi separati; inoltre rispetta il ruolo che le variabili manifeste hanno nella specificazione del modello, contrariamente ad altri approcci basati sulla definizione di LV come combinazione lineare degli indicatori manifesti.

Infine l'algoritmo proposto nel paragrafo 4 generalizza la metodologia ad ogni combinazione dei livelli di misurazione delle variabili manifeste, rispettando la scala di misurazione di ognuna.

Le stime delle LV sono ottenute come combinazioni non lineari delle variabili categoriali e combinazioni lineari delle variabili ottimamente trasformate:

$$\boldsymbol{\eta}_j = \Sigma_s \, \mathbf{y}_s^{os} \, c_{js}^{*} \tag{7}$$

L'approccio mostrato è assai flessibile, non impone ipotesi forti sulle variabili (normalità) o sui parametri e risolve alcuni problemi lasciati irrisolti da approcci in letteratura (indicati in parentesi):

1. stima in un solo stadio le LV e i parametri di regressione su *r* dimensioni latenti (PLS);
2. restituisce LV non indeterminate nei punteggi (LISREL);
3. rispetta le relazione di causalità supposte nel modello tra LV-MV (PLS, CP, RCD);
4. evita pesanti ipotesi distributive (LISREL, MIMIC);
5. si estende a variabili categoriali (PLS, MIMIC, RCD, RCDR) e in un approccio di Optimal Scaling stima i parametri di scaling insieme ai parametri del modello statistico (NPCA).

6. UNA APPLICAZIONE: STIMA DELLO STATUS ANAGRAFICO-CULTURALE E STATUS PROFESSIONALE COME LV

La metodologia viene applicata alla stima di una LV, rappresentativa del capitale umano (**h**), bidimensionale: lo status anagrafico-culturale (*SAC*) che tiene conto anche della ricchezza familiare e lo status professionale (*SP*) valutato a livello familiare che incidono economicamente (reddito da lavoro e reddito finanziario) su un campione di 4103 famiglie americane sui dati dell'indagine Federal Reserve su reddito e ricchezza per il 1983 (Avery e Elliehausen, 1985).

Per la stima dei punteggi delle due componenti vanno specificate, secondo il path diagram in figura 3, per ogni dimensione latente $\boldsymbol{\eta}_j$ (j=1,2) le variabili dipendenti, nelle matrici $\mathbf{Y}_j$ (cfr. equazione (4))[4], e le variabili esplicative nelle due matrici $\mathbf{X}_j$ (cfr. equazione (4)).

Le variabili dipendenti che costituiscono gli effetti monetari di *SAC* e di *SP* (secondo la specificazione del modello ricorsivo di Dagum, 1994) sono il reddito da lavoro (**y1**) e il reddito familiare di tipo finanziario (**y2**), entrambe variabili quantitative.

In tabella 1 vengono mostrate le variabili indipendenti (per ogni componente latente), il relativo livello di misurazione e il vettore dei coefficienti di regressione

[4] L'indice *j* per le matrici **X** ed **Y** permette di distinguere le variabili cause e gli indicatori separatamente per ogni LV $\boldsymbol{\eta}_j$.

stimati ($\boldsymbol{\beta}_j$) nella (4) sia per *SAC* sia per *SP* con l'algoritmo del paragrafo 4.

La figura 6 mostra le distribuzioni dei punteggi latenti stimati per *SAC* e *SP*.

Alcune variabili fondamentali come titolo di studio, anni di lavoro (Dagum *et al.*, 2002) non potendo essere assimilate ad effetti delle due LV stimate non sono utilizzabili come informazione rilevante per la stima di una proxy di **h** nel modello fattoriale.

TABELLA 1

Variabili in analisi: descrizione, tipo di misurazione e stime dei parametri.

Variabili esplicative	Descrizione**	Scala	Coefficienti strutturali	
			SAC	SP
Interc			-162,427	-10,844
X1	Età H	Numerica	-,005	
X2*	Sesso H (1=Femmina)	Nominale	3,451	
X3*	Stato civile H	Nominale	-3,243	
X4	Anni di scolarità H	Numerica	1,040	0,616
X5	Anni scolarità S	Numerica	1,057	
X6	Anni di lavoro H	Numerica	0,053	0,210
X7	Anni di lavoro S	Numerica		-,0742
X8*	Posiz. Lavorativa H	Ordinale		-2,231
X9*	Tipo di occupazione H	Ordinale		-4,824
X10*	Settore lavorativo H	Nominale		-4,477
X11	Posiz. Lavorativa S	Numerica		0,877
X12	Ricchezza totale F	Numerica	3,021	
y1	Reddito da lavoro F	Numerica		
y2	Reddito finanziario F	Numerica		

* da trasformare con l'Optimal Scaling. Valori di categoria via via maggiori per tali variabili corrispondono a "situazioni, status e condizioni" più svantaggiate dal punto di vista socio-economico.

** nucleo familiare (F), capofamiglia (H), coniuge (S).

Fonte: dati Federal Reserve Board 1983 (Avery e Elliehausen, 1985).

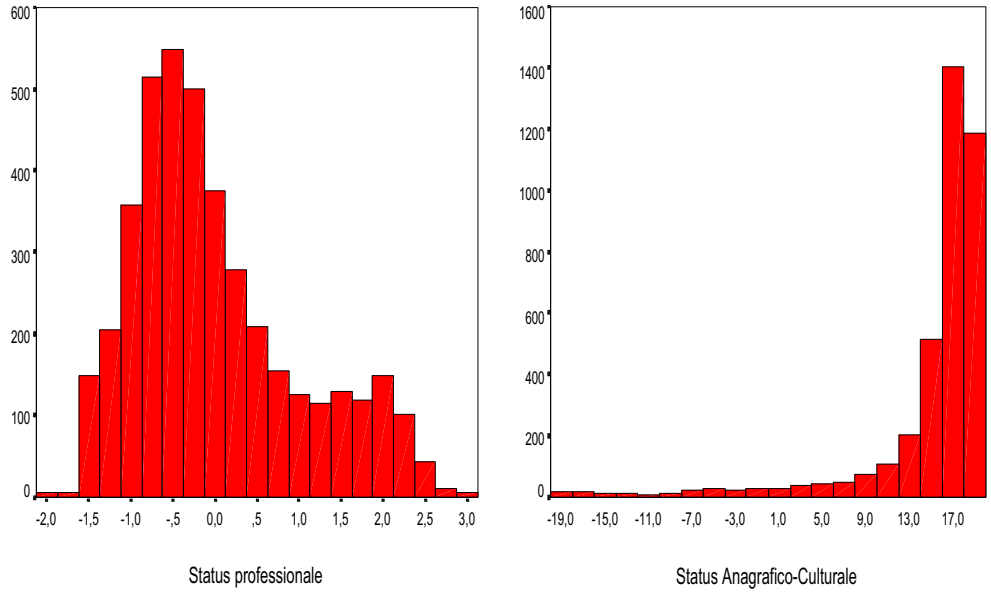

Figura 6 – Distribuzione dei punteggi stimati delle LV *SP* e *SAC* per le Famiglie Americane (1983).

Dipartimento di Statistica
Università di Milano-Bicocca
PIETRO GIORGIO LOVAGLIO

RIFERIMENTI BIBLIOGRAFICI

R.B.E. AVERY, G.E. ELLIEHAUSEN (1985), *1983 Survey Of Consumer Finances: Technical Manual And Codebook*, Federal Reserve Board, Washington.

P.M. BENTLER (1982), *Linear system with multiple levels and types of latent variables* in: "Sistem under indirect observation", in K. Joreskog, H. Wold (eds), North Holland, Amsterdam.

BENTLER (1989), *EQS Strumental equations program manual,* BMPD Statistical software Los Angeles.

V. CAREY, S.L. ZEGER, P., DIGGLE (1993), Mode*lling multivariate binary data with alternating logistic regressions*, "Biometrika", 80, 3, pp. 517-526.

C. DAGUM (1994), *Human Capital, Income and Wealth Distribution Models and Their Applications,* "Proceedings of the 154th Meeting of the American Statistical Association, Business and Economic Statistics Section", pp. 253-258.

C. DAGUM, G. VITTADINI (1996), *Human Capital Measurement and Distribution,* "Proceedings of the 156th Meeting of the American Statistical Association, Business and Economic Statistics Section", pp. 194-199.

C. DAGUM, G. VITTADINI, M. COSTA, P. LOVAGLIO (2002), *An estimation methodology for variable "human capital" in the U.S.A.*, preprint.

L. FALLOWFIELD (1996), *Quality of quality of life data*, "Lancet", 348, pp. 421-422.

G.M. FITZMAURICE, N.M. LAIRD (1997), Regression models for mixed discrete and continuous responses with potentially missing values, "Biometrics", 53, pp. 110-122.

P.J. GERTLER (1988), A *latent variable model of quality determination*, "Journal of Businness and Economics Statistics", 6, pp. 96-104.

K. HAAGEN, G. VITTADINI (1991), *Regression Component Decomposition in structural analysis*, "Communications in Statistics", 20 pp 1153-1161.

A.G. ISRAELS (1984), *Redundancy analysis for qualitative variables*, "Psychometrika", 49, pp. 331-346.

K. JORESKOG (1978), *Structural Analysis of covariance and correlation matrices*, "Psychometrika", 43, pp. 443-477.

K. JORESKOG, S. GOLDBERGER (1975), *Estimation of a model with multiple indicators and multiple causes of a single latent variable*, "Journal of American Statistical Society", 70, pp. 631-639.

W.J. KELLER, T.J. WAANSBEEK (1983), *Multivariate methods for quantitative and qualitative data*, "Journal of Econometrics", 22, pp. 91-111.

J.B. KRUSKAL (1965), *Analysis of factorial experiments by estimating monotone trasformations of the data*, "Journal of Royal Statistical Society B", 27, pp. 251-263.

K. Y. LIANG, S.L. ZEGER, B. QAQISH (1992), *Multivariate regression analyses for categorical data*, "Journal of Royal Statistical Society B" 54, 1, pp. 3-40.

P.G. LOVAGLIO (1997), *Un algoritmo per la regressione multipla con dati categoriali*, "Quaderni di Matematica e Statistica Applicata alle Scienze Socio-Economiche" 19, No 3, Trento.

P.G. LOVAGLIO (2001), *The estimate of latent outcomes*, "Atti del Convegno Intermedio della Società Italiana di Statistica", Processi e Metodi Statistici di Valutazione, CISU,Roma, pp. 393-396.

P.G. LOVAGLIO, G. VITTADINI (2002), *Latent variables in structural model: an alternative to PLS*, "Classification and Data Analysis", LA, Usa, in corso di stampa.

D.M. ROMNEY, C.D., JENKINS, J.M. BYNNER (1992), *A structural analysis of Health-related Quality of life dimensions*, "Human Relations", 45, pp. 165-176.

P., J. SCHONEMANN, K. STEIGER (1976), *Regression Component Analysis*, "British Journal of Mathematical and Statistical Psychology", 29, pp. 175-189

T. SHEEHAN (1991), *A structural equation model of Factors affecting neonatal outcomes*, "International Workshop on Statistical Modelling and Latent Variables", Trento.

D. STEWART, W. LOVE (1968), *A general canonical correlation index,* "Psychological Bulletin", 70, pp. 160-163.

K. TAKEUCHI, H. YANAI, B.N. MUKHERJEE (1982), *The Foundations of Multivariate Analysis, A Unified approach by means of projection onto linear subspaces*, Wiley Eastern Limited.

M.K.S. TSO (1981), *Reduced rank regression and Canonical analysis*, "Journal of Royal Statistical Society", B, 43, pp.183-189.

G. VITTADINI (1997), *Una metodologia statistica per la performance evaluation dei servizi alla persona di pubblica utilità*, "Atti del convegno SIS La statistica per le imprese", Tirrenia, Torino.

G. VITTADINI (1999), *Analysis of qualitative variables in Structural Models with unique solutions*, in: "Classifications and data analysis-Theory and Applications", Vichi M., Opitz O. (eds.), Springer and Verlag, pp. 203-210.

G. VITTADINI, P.G. LOVAGLIO (2001), T*he estimate of latent variables in a structural model: an alternative approach to PLS*, "2nd International Symposium on PLS and Related Methods", October 1-3, 2001 Capri (Italy), Cisia Ceresta, Montreuil, France pp.423-434.

H. WOLD (1982), *Soft modeling: the basic design and some extension* in: "System under indirect observation", Joreskog K. Wold H. (Eds.), North Holland, Amsterdam.

F. YOUNG, Y. TAKANE, J. DE LEEUW (1978), *The Principal Component of Mixed Measurement Level Multivariate Data: an Alternating Leasts Squares Method with Optimal Scaling Features*, "Psychometrika", 43, pp. 279-281.

RIASSUNTO

La stima di variabili latenti da variabili osservate miste

Il presente lavoro cerca di formulare, un modello equivalente, ma alternativo al modello di analisi fattoriale (AF) che ha l'obiettivo di stimare una o più variabile latente (LV) e di evidenziare i legami causalità tra LV e variabili manifeste (MV) all'interno di un modello statistico. L'attuale proposta risolve i problemi di consistenza logica manifestati da altre metodologie in letteratura nate come alternativa ad AF per stimare i punteggi di una o più LV (PLS, RCD, MIMIC). In particolare il modello introdotto ha il pregio di stimare in un solo stadio sia punteggi di una LV (concepita come combinazione di MV) sia i parametri di regressione del modello di misura tra LV e MV, rispettando i legami logici supposti nel modello specificato. La metodologia di stima viene infine estesa al caso di variabili manifeste miste (qualitative e categoriali), distinguendo accuratamente la scala di misurazione (ordinale, nominale) di ogni MV, per stimare una proxy bidimensionale del capitale umano.

SUMMARY

The estimate of latent variables with mixed manifest variables.

The aim of this paper is to propose a general nonparametric model to estimate latent variables with scores non indeterminate; in this paper, following other approaches (PLS, RCD, RCDR), a latent variable (LV) is conceived as a linear combination of predictors (causes) which best predicts a set of dependent variables (indicators), maximising, in this manner, all available information about a LV in the specified model.

The model is also extented to categorical variables (nominal, ordinal) by means of optimal scaling methodology and applied to the estimate of a bidimensional LV as a proxy of human capital for US families in 1983.