Rilevanza delle analisi di misture di distribuzioni nelle valutazioni di efficacia con metodi di inferenza causale

Andrea Mercatanti[1]
Dipartimento di Statistica e Matematica Applicata all'Economia – Università di Pisa

Riassunto. Alcune problematiche metodologiche relative all'indebolimento delle usuali condizioni di applicabilità dei metodi di inferenza causale riguardano l'analisi di misture di distribuzioni. In particolare il presente contributo intende prendere in considerazione la questione dell'eliminazione dei vincoli di esclusione nell'utilizzo di variabili strumentali ai fini della valutazione dell'efficacia di una variabile di tipo binario sotto l'ipotesi che la variabile di risposta sia distribuita secondo una normale. Rispetto alle usuali analisi su misture di distribuzioni si evidenzia un maggiore contenuto informativo riguardo alle probabilità di appartenenza ai gruppi componenti le misture. Di converso emergono però maggiori difficoltà inferenziali connesse alla plurimodalità della funzione di verosimiglianza prodotta dalla presenza di più misture con componenti comuni. Il contributo prende inoltre in considerazione una procedura di massimizzazione vincolata della verosimiglianza che sfrutta le maggiori informazioni relative alle probabilità di appartenenza ai gruppi, al fine di risolvere i problemi legati alla plurimodalità della funzione di verosimiglianza.
Parole chiave: misture di distribuzioni normali, variabili strumentali, vincolo di esclusione.

1. Introduzione

L'importanza delle applicazioni di inferenza causale alle problematiche della valutazione di efficacia è ormai consolidata, e spazia dall'utilizzazione delle variabili strumentali e dei *propensity score* alle stratificazioni principali solo per citare alcune me-

[1] Il presente lavoro è stato realizzato nell'ambito del progetto "Transizioni Università-Lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionale delle determinanti", cofinanziato dal MIUR; coordinatore nazionale è Luigi Fabbris, coordinatore del gruppo di Firenze è Bruno Chiandotto.

todologie di largo uso. Nelle analisi riguardanti la valutazione delle transizione Università-lavoro, i suddetti metodi possono trovare applicazione ad esempio nella valutazione dell'effetto del conseguimento di una laurea su di un outcome post-laurea quali ad esempio il tempo di disoccupazione, il reddito, la soddisfazione professionale, o la congruenza tra le materie di studio e le competenze necessarie alla professione svolta.

Da un punto di vista più strettamente metodologico si può mettere in evidenza che alcune problematiche relative all'indebolimento delle condizioni di applicabilità dei modelli causali riguardano l'analisi delle misture di distribuzioni. In particolare questo concerne la rimozione del cosiddetto vincolo di esclusione nell'uso delle variabili strumentali a fini causali. Nella loro applicazione più semplice le variabili strumentali possono essere introdotte per la valutazione dell'efficacia di una certa variabile binaria su di un outcome di qualsiasi tipo (Imbens e Angrist, 1994). Tra le condizioni necessarie all'identificazione di effetti causali con l'ausilio di variabili strumentali una delle più problematiche e difficili da soddisfare è il vincolo di esclusione in base al quale la variabile strumentale non può avere effetti diretti sull'outcome di interesse.

La problematica sorge, ad esempio, nella valutazione dell'effetto scolarizzazione sul reddito mediante l'uso di variabili strumentali legate alla coorte di nascita. In questi casi (Card e Lemieux, 2001) la teoria microeconomica suggerisce, in base a modelli con imperfetta sostituibilità tra individui con scolarizzazione simile, che l'effetto della scuola sul reddito riflette anche variazioni nell'offerta relativa di individui con scolarizzazione simile tra le varie coorti di nascita. Ecco presentarsi quindi una critica microeconomica all'uso di variabili strumentali legate alla coorte di nascita nella valutazione del *return to schooling*. La suddetta motivazione si basa sul concetto di equilibrio economico generale e inficia la soddisfazione del vincolo di esclusione poiché la coorte di nascita ragionevolmente agisce sul reddito oltre che in base al trend storico della scolarizzazione anche in base a questioni di mercato legate alla numerosità delle coorti. In altre parole si può ragionevolmente ipotizzare che in questo caso esista un effetto diretto della variabile strumentale sull'outcome.

Il presente contributo, di tipo metodologico, si basa su di una impostazione parametrica dell'analisi causale con variabili strumentali, ossia su di una formulazione della funzione di verosimiglianza per un esperimento randomizzato con non-compliance che mette in particolare evidenza la presenza di misture di distribuzioni. Sulla base di una proposta di massimizzazione vincolata della verosimiglianza, viene svolta un'analisi di tipo simulativo finalizzata ad un primo giudizio sulla bontà e sui limiti della proposta stessa.

2. Proposta di analisi vincolata della funzione di verosimiglianza

Già a partire dal contributo di Imbens e Rubin (1997) si è data una formalizzazione di tipo parametrico al modello di regressione lineare semplice con variabili strumentali per l'identificazione e la stima di effetti causali, nel caso di variabile trattamento binaria. Il punto di vista filosofico causale preso in considerazione dagli autori nell'esplicitazione della funzione di verosimiglianza è quello basato sull'idea di controfattualità ed a questo vogliamo continuare ad attenerci nel presente lavoro. In termini formali, si fa riferimento alla struttura teorica di un'esperimento randomizzato per il quale indichiamo con y_i la variabile di risposta, con D_i il trattamento di tipo binario (0,1), e con Z_i la variabile strumentale da intendersi come assegnazione al trattamento di tipo binario.

Di conseguenza occorre ricordare che la popolazione complessiva si può dividere in quattro gruppi, denominati *compliance status*, ognuno dei quali si caratterizza per come gli individui reagiscono dal punto di vista controfattuale all'assegnazione al trattamento. Si parla infatti di *always-takers* per indicare il gruppo di individui che assumono sempre il trattamento (ossia presentano $D_i = 1$ indipendentemente dal valore assunto dall'assegnazione al trattamento Z_i); di *never-takers* per indicare gli individui che non assumono mai il trattamento (ossia presentano $D_i = 0$ indipendentemente dal valore assunto dall'assegnazione al trattamento Z_i); di *compliers* per gli individui che assumono o meno il trattamento in base a quanto assegnatoli (ossia presentano $D_i = 1$ se $Z_i = 1$, e $D_i = 0$ se $Z_i = 0$); e di *defiers* per gli individui che assumono il trattamento in maniera opposta all'assegnazione. Imbens e Angrist (1994) definiscono le condizioni in base alle quali un'analisi di regressione della variabile y_i sul trattamento D_i, supportata dalla variabile strumentale Z_i, identifica l'effetto causale del trattamento per il gruppo dei compliers. Tra queste condizioni spicca per difficoltà di soddisfacimento il vincolo di esclusione, in base al quale la variabile Z_i non può avere effetti diretti su y_i. Al fine della rimozione completa del vincolo di esclusione e partendo dalla funzione di verosimiglianza proposta dai suddetti autori, si può arrivare mediante opportune riparametrizzazioni (Mercatanti, 2004) alla scrittura della stessa in una forma che ne permetta la massimizzazione vincolata ad un opportuno sottospazio parametrico. Questo risulta individuabile senza far ricorso ad informazioni aggiuntive rispetto alle ipotesi necessarie all'identificazione di effetti causali mediante variabili strumentali, a parte l'ipotizzata forma funzionale per la distribuzione dell'outcome, essendo in ambito parametrico.

In estrema sintesi, si intende far riferimento alla seguente funzione di verosimiglianza:

$$L(\boldsymbol{\theta}) = \prod_{i\in(D_i=1,Z_i=0)} \omega_{a0}\cdot N(y_i \mid \mu_{a0},\sigma_{a0}) \times \prod_{i\in(D_i=0,Z_i=1)} \omega_{n1}\cdot N(y_i \mid \mu_{n1},\sigma_{n1})$$
$$\times \prod_{i\in(D_i=1,Z_i=1)} \left[\omega_{a1}\cdot N(y_i \mid \mu_{a1},\sigma_{a1}) + \omega_{c1}\cdot N(y_i \mid \mu_{c1},\sigma_{c1})\right]$$
$$\times \prod_{i\in(D_i=0,Z_i=0)} \left[\omega_{n0}\cdot N(y_i \mid \mu_{n0},\sigma_{n0}) + \omega_{c0}\cdot N(y_i \mid \mu_{c0},\sigma_{c0})\right], \qquad (1)$$

dove[2] si è indicato: con ω_{tz} la probabilità di appartenenza al gruppo di individui nel compliance status *t=a (always-takers), n(never-takers), c(compliers)* e con assegnazione al trattamento $Z_i = z$; con μ_{tz} e σ_{tz} rispettivamente la media e lo standard error per il gruppo di individui nel compliance status *t* e con assegnazione al trattamento $Z_i = z$.

La presenza nella (1) di due misture di distribuzioni normali comporta problematiche di tipo analitico e computazionale nell'esecuzione di un'analisi MLE. Le misture di distribuzioni normali assumono infatti caratteristiche analitiche che le rendono di non facile analisi. I principali elementi perturbartivi in un'analisi MLE della (1) possono essere sintetizzati nei seguenti tre punti:

- la (1) non è limitata sopra (Day, 1969) quindi in generale l'analisi MLE è mal posta poiché non esiste un massimizzatore assoluto; è stato però dimostrato che esiste un massimizzatore locale consistente, efficiente e asintoticamente normale (Kiefer, 1978) sul quale può quindi essere dirottata la ricerca;
- la (1) è multimodale;
- la massimizzazione locale della (1) produce massimi *spuri*, ossia punti di massimo locale tipicamente in corrispondenza di raggruppamenti di poche unità anomale; questi punti possono tuttavia essere facilmente individuati poiché presentano una componente di varianza prossima allo zero.

Numerose proposte sono state avanzate in letteratura per l'analisi MLE di misture. Tra quelle che appaiono particolarmente convincenti si può citare un'approccio di tipo generale (Priebe, 1994), ossia la conduzione di una serie di massimizzazioni non vincolate seguite da un'analisi dei punti di massimo locali al fine di individuare e scartare quelli spuri. Successivamente la stima ML del vettore parametrico può essere considerata quella corrispondente al massimo tra i rimanenti punti. La proposta appare semplice e non introduce informazioni extra nell'analisi, anche se una ricerca sufficientemente esauriente dei punti di massimo locale si può rilevare particolarmente dispendiosa in termini di tempo di calcolo.

[2] Le ipotesi in base alle quali vale la descritta funzione di verosimiglianza sono le seguenti: distribuzione normale per l'outcome; *Stable Unit Treatment Value Assumption* in base alla quale per ogni individuo i comportamenti controfattuali non dipendono dal trattamento degli altri individui; identica probabilità di assegnazione al trattamento per ogni individuo; inesistenza di defiers.

Oltre alle problematiche caratteristiche delle analisi di misture, l'analisi della (1) comporta delle complicazioni aggiuntive dovute al cosiddetto *label switching,* inconveniente dovuto ad eventuali permutazioni per alcune variabili indicanti l'appartenenza ai gruppi componenti le misture (etichette). La problematica del label switching concerne l'identificabilità delle misture di distribuzioni. E' risaputo infatti (Hjort, 1986) che in una mistura di distribuzioni appartenenti alla stessa famiglia parametrica, $f(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{g} \omega_j \; f_j(\mathbf{x};\boldsymbol{\theta}_j)$, il vettore parametrico $\boldsymbol{\theta}$ non è identificato; viene invece identificata una classe di distribuzioni in quanto $f(\mathbf{x};\boldsymbol{\theta})$ è invariante alle *g!* permutazioni nelle etichette delle componenti in $\boldsymbol{\theta}$. Il label switching sebbene non sia un problema rilevante nella stima MLE di una mistura di distribuzioni appartenenti alla stessa famiglia parametrica a fini di cluster analysis, lo diventa però in un'analisi MLE della (1). Occorre infatti considerare che gli effetti causali in un'ottica controfattuale sono definiti dalle tre differenze $\Delta_t = (\mu_{t1} - \mu_{t0})$ con *t=a,n,c*, di conseguenza l'identificazione degli effetti causali necessita dell'esatta etichettatura di tutte le componenti.

Una diversa strategia di analisi della funzione di verosimiglianza (1) viene suggerita dalla considerazione che, senza l'aggiunta di ulteriori ipotesi, esiste la possibilità di stimare facilmente le probabilità di appartenenza ai gruppi componenti le misture anche al di fuori di un contesto di massima verosimiglianza. Questi elementi informativi possono essere sfruttati nella stima di massima verosimiglianza del vettore parametrico, vincolando la ricerca ad opportuni sottospazi parametrici.

Sotto le ipotesi che hanno portato alla scrittura della (1) è infatti possibile stimare (Mercatanti, 2004) le probabilità ω_{tz}, caratterizzanti le due misture, con le quantità $\hat{\phi}_{tz}$:

$$\hat{\phi}_{a1} = [\#(D_i = 1, Z_i = 0) / \#(Z_i = 0)\,] - [\#(D_i = 1, Z_i = 0) \cdot N^{-1}],$$

$$\hat{\phi}_{n0} = [\#(D_i = 0, Z_i = 1) / \#(Z_i = 1)\,] - [\#(D_i = 0, Z_i = 1) \cdot N^{-1}],$$

$$\hat{\phi}_{c0} = [\#(D_i = 0, Z_i = 0) \cdot N^{-1}] - \hat{\phi}_{n0},$$

$$\hat{\phi}_{c1} = [\#(D_i = 1, Z_i = 1) \cdot N^{-1}] - \hat{\phi}_{a1}.$$

Al fine di sfruttare al massimo le informazioni disponibili risulta allora proponibile la massimizzazione della (1) vincolata ad un intorno del punto $(\hat{\phi}_{a1}, \hat{\phi}_{n0}, \hat{\phi}_{c0}, \hat{\phi}_{c1})$, ossia la ricerca del punto di massimo $\hat{\boldsymbol{\theta}}$ soddisfacente, per un certo valore di *k*, le condizioni:

$$\left|\hat{\phi}_{a1} - \hat{\omega}_{a1}\right| < k, \; \left|\hat{\phi}_{n0} - \hat{\omega}_{n0}\right| < k, \left|\hat{\phi}_{c1} - \hat{\omega}_{c1}\right| < k, \; \left|\hat{\phi}_{c0} - \hat{\omega}_{c0}\right| < k.$$

3. Analisi esemplificativa basata su dataset artificiali

La sezione presenta un'analisi di tipo simulativo condotta su dataset artificiali relativi ad esperimenti randomizzati con non-compliance e senza vincoli di esclusione; i dataset verranno estratti da popolazioni ipotetiche soddisfacenti le ipotesi espresse nella nota 3 della precedente sezione. L'obiettivo è la verifica empirica dell'uso della procedura di massimizzazione vincolata ad un intorno sferico del punto $(\hat{\phi}_{a1}, \hat{\phi}_{n0}, \hat{\phi}_{c0}, \hat{\phi}_{c1})$. Si consideri allora un primo campione artificiale composto da 10000 unità estratte da una popolazione ipotetica i cui parametri vengono riportati in Tabella 1. Al fine di identificare i punti di massima verosimiglianza locale sono state condotte 100 procedure di massimizzazione libera utilizzando l'algoritmo EM e partendo ogni volta con valori casuali del vettore parametrico. È stato inoltre identificato il massimo consistente, $\hat{\boldsymbol{\theta}}_1$, come il punto al quale converge l'algoritmo EM partendo con il vero vettore parametrico. Come previsto la funzione è risultata multimodale, dalla Tabella 2[3], si può notare infatti come nelle 100 prove si sia ottenuto:

- per 22 volte convergenza al massimo consistente, $\hat{\boldsymbol{\theta}}_1$,
- per 4 volte convergenza a massimi spuri, cioè punti con una componente di varianza prossima allo zero $(\hat{\boldsymbol{\theta}}_5, \hat{\boldsymbol{\theta}}_6, \hat{\boldsymbol{\theta}}_7, \hat{\boldsymbol{\theta}}_8)$,
- per 74 volte convergenza ad altri punti di massimo locale $(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_3, \hat{\boldsymbol{\theta}}_4)$, che vedremo rappresentano una tipologia di massimi spuri dovuti al label switching e anomala rispetto alle usuale analisi di misture. Da notare che in ogni soluzione le stime dei due parametri μ_{a0}, μ_{n1}, sono identiche in quanto calcolate sempre come medie delle unità appartenenti ai gruppi $(D_i = 1, Z_i = 0)$ e $(D_i = 0, Z_i = 1)$ rispettivamente.

Per analizzare le caratteristiche degli otto punti di massimo torna utile utilizzare le probabilità di imputazione calcolate durante l'ultimo E-step dell'algoritmo EM. Per probabilità di imputazione si intende la probabilità di appartenenza ad ognuno dei tre compliance-status (always-takers, never-takers, compliers) e che per ogni unità viene calcolata ad ogni iterazione durante il passo "E" dell'algoritmo EM. Dalle probabilità di imputazione è inoltre possibile calcolare l'*imputation rate* (Holgersson e Jorner, 1998) il quale rappresenta un'utile indice per la bontà della scissione di una mistura. L'imputation rate è dato dalla media della più alta probabilità di imputazione osservata per ogni unità. Nel nostro caso, l'imputation rate complessivo assume un valore molto alto in ogni soluzione e non consente quindi una discriminazione tra le stesse.

---

[3] Per analogia con la parametrizzazione classica si sono riportate direttamente le stime delle probabilità di appartenenza ai compliance status $(\omega_a, \omega_n, \omega_c)$ ottenute come medie ponderate delle stime di massima verosmiglianza vincolate $(\hat{\omega}_{a0}, \hat{\omega}_{a1}, \hat{\omega}_{n0}, \hat{\omega}_{n1}, \hat{\omega}_{c0}, \hat{\omega}_{c1})$. Per ragioni di spazio non vengono riportate le stime delle componenti di varianza σ_{tz}.

Essendo però in ambito simulativo il compliance status di ogni singola unità è conosciuto. Il confronto tra i veri compliance status delle unità statistiche e le probabilità di imputazione agli stessi rende possibile verificare il grado e la bontà delle scissioni delle misture per ogni punto di massimo locale. Per rendere chiara l'idea consideriamo la Tabella 3 che riporta, per i gruppi (*t,z*), la media e lo scarto quadratico medio delle probabilità di imputazione ad ognuno dei tre compliance status calcolate all'ultima iterazione dell'algoritmo EM, per alcuni punti di massimo locale[4].

Tabella 1. *Valori parametrici della popolazione ipotetica utilizzata per l'analisi simulativa.*

t	ω_t	(μ_{t0}, σ_{t0})	(μ_{t1}, σ_{t1})
a	0.4	(0, 1)	(1, 1.2)
n	0.25	(1, 1.15)	(2,1)
c	0.35	(6, 0.85)	(7, 0.7)
$P(Z_i = 1) = 0.25$			

Tabella 2. *Punti di massimo locale identificati da 100 procedure di massimizzazione non vincolata.*

	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$	$\hat{\theta}_7$	$\hat{\theta}_8$
ω_a	0.400	0.387	0.400	0.387	0.387	0.400	0.486	0.387
ω_n	0.250	0.250	0.323	0.323	0.062	0.512	0.512	0.062
ω_c	0.349	0.361	0.276	0.288	0.549	0.087	0.001	0.549
μ_{a0}	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
μ_{a1}	1.074	6.999	1.076	6.998	7.002	1.093	3.854	7.002
μ_{n0}	1.022	1.020	5.993	5.994	-2.377	3.913	3.913	-2.431
μ_{n1}	2.076	2.076	2.076	2.076	2.076	2.076	2.076	2.076
μ_{c0}	5.988	5.987	1.032	1.035	3.913	2.377	-2.379	3.913
μ_{c1}	7.000	1.072	7.002	1.070	1.076	7.012	0.855	1.076
Log Lik.	-30164	-30177	-30225	-30267	-32684	-33208	-33232	-32692
Imp. rate	0.9938	0.9970	0.9968	0.9968	0.9995	0.9995	0.9995	0.9997

[4] Non vengono riportate le probabilità di imputazione ai gruppi (*a*,0) e (*n*,1) poichè per le unità appartenenti a questi due gruppi le informazioni a disposizione consentono un'esatta imputazione ai rispettivi compliance status fin dalla prima iterazione dell'algoritmo EM.

Si osserva per il punto di massimo consistente $\hat{\boldsymbol{\theta}}_1$ una soddisfacente attribuzione delle unità ai compliance status. Ad esempio per le unità appartenenti al gruppo (a,1) la probabilità di imputazione al gruppo always-takers ha media 0.997 con s.e. di 0.036; questo significa che nel successivo M step le unità nel gruppo (a,1) vengono in sostanza correttamente considerate come always-takers. Analogamente per le unità nel gruppo (c,1) la probabilità di imputazione al gruppo compliers ha media 0.990 e s.e. 0.066, e quindi nel successivo M step queste unità vengono in sostanza correttamente considerate come compliers. Considerando che le unità nei gruppi (a,1) e (c,1) formano una delle due misture caratterizzanti la (1) ciò è indice di un'ottima scissione della mistura. Discorso analogo vale per le unità nei gruppi (n,0) e (c,0) e per la rispettiva mistura.

Si consideri adesso il punto di massimo $\hat{\boldsymbol{\theta}}_2$ dove a differenza della precedente soluzione, $\hat{\boldsymbol{\theta}}_1$, la scissione della mistura formata dai gruppi (a,1) e (c,1) non è più soddisfacente. Dalla Tabella 3 si vede infatti come le unità nel gruppo (a,1) vengano in sostanza erroneamente attribuite al gruppo dei compliers, e come le unità nel gruppo (c,1) vengano erroneamente attribuite al gruppo degli always-takers. Situazioni simili si riscontrano per le soluzioni $\hat{\boldsymbol{\theta}}_3$ e $\hat{\boldsymbol{\theta}}_4$. Precisamente per la soluzione $\hat{\boldsymbol{\theta}}_3$ si osserva un'errata scissione della mistura formata dai due gruppi (n,0) e (c,0), e per la soluzione $\hat{\boldsymbol{\theta}}_4$ l'errata scissione di entrambe le misture. Per questi punti $(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_3, \hat{\boldsymbol{\theta}}_4)$ il valore dell'imputation rate resta comunque alto.

Tabella 3. *Probabilità di imputazione per alcuni punti di massimo locale.*

soluzione	(t,z)	t					
		a		n		c	
		media	s.e.	media	s.e.	media	s.e.
$\hat{\boldsymbol{\theta}}_1$	(a,1)	0.997	0.036	0	0	0.002	0.036
	(n,0)	0	0	0.990	0.069	0.009	0.069
	(c,0)	0	0	0.009	0.066	0.990	0.066
	(c,1)	0.009	0.066	0	0	0.990	0.066
$\hat{\boldsymbol{\theta}}_2$	(a,1)	0.002	0.039	0	0	0.997	0.039
	(n,0)	0	0	0.990	0.070	0.009	0.070
	(c,0)	0	0	0.009	0.068	0.990	0.068
	(c,1)	0.997	0.037	0	0	0.002	0.037
$\hat{\boldsymbol{\theta}}_5$	(a,1)	0.001	0.034	0	0	0.998	0.034
	(n,0)	0	0	0.001	0.032	0.998	0.032
	(c,0)	0	0	0.000	0.000	1	0
	(c,1)	0.996	0.041	0	0	0.003	0.041

Finora l'errata scissione di una mistura si è concretizzata nell'attribuzione di tutte le unità al compliance status errato. I restanti punti di massimo locale assumono anche le caratteristiche dei punti di massimo spuri usualmente identificabili nelle analisi di misture. Infatti per questi l'errata scissione di una mistura si manifesta anche con l'attribuzione di quasi tutte le unità ad uno solo dei due compliance status. Per chiarire consideriamo il punto di massimo locale $\hat{\boldsymbol{\theta}}_5$ sempre in Tabella 3; si può osservare come le unità nella mistura formata dai due gruppi (*n*,0) e (*c*,0) vengono in sostanza attribuite quasi tutte al gruppo dei compliers. Lo stesso modo di scindere le misture si manifesta anche per i restanti punti di massimo $\hat{\boldsymbol{\theta}}_6$ e $\hat{\boldsymbol{\theta}}_7$.

L'errata attribuzione delle unità nelle misture produce conseguenze negative nella stima dei componenti del vettore parametrico. Tornando infatti a considerare la soluzione $\hat{\boldsymbol{\theta}}_2$, si osservino i diversi valori delle stime delle probabilità $(\omega_a, \omega_n, \omega_c)$ rispetto a $\hat{\boldsymbol{\theta}}_1$. Questo risultato deriva dal fatto che ad ogni iterazione dell'algoritmo EM le stime delle probabilità $(\omega_a, \omega_n, \omega_c)$ vengono calcolate durante il passo "M" come media delle probabilità di imputazione ai compliance status. Per esser chiari si faccia riferimento alla Tabella 4, la prima riga della quale riporta le quote relative di popolazione, $\psi_{t,z}$, appartenenti ai sei gruppi (*t,z*) per un grande campione estratto dalla popolazione ipotetica considerata. Si osservi come le quote relative di popolazione appartenenti ai tre compliance status si possano facilmente ottenere come:

$$\psi_a = (\psi_{a,0} + \psi_{a,1}) = (0.30 + 0.10) = 0.40,$$

$$\psi_n = (\psi_{n,0} + \psi_{n,1}) = (0.1875 + 0.0625) = 0.25,$$

$$\psi_c = (\psi_{c,0} + \psi_{c,1}) = (0.2625 + 0.0875) = 0.35.$$

Questi valori corrispondono alle stime $\hat{\omega}_a, \hat{\omega}_n, \hat{\omega}_c$ in $\hat{\boldsymbol{\theta}}_1$, a parte piccole differenze dovute sia alla variabilità campionaria che al fatto che le probabilità di imputazione osservate all'ultima iterazione dell'algoritmo EM non sono sempre esattamente binarie (vedi Tabella 2). I valori poc'anzi calcolati di ψ_a, ψ_n, ψ_c costituiscono infatti dei valori limite delle medie aritmetiche delle probabilità di imputazione ai compliance status conseguenti ad una corretta scissione delle misture che caratterizzano la (1). Riconsiderando adesso la soluzione $\hat{\boldsymbol{\theta}}_2$, dalla Tabella 3 si osserva come le unità nel gruppo (*a*,1) vengono erroneamente attribuite al gruppo (*c*,1) e viceversa. Dopo l'errata scissione della mistura composta dai due suddetti gruppi, le quote relative

Tabella 4. *Quote relative di popolazione per compliance status, t, e assegnazione, z.*

	ψ_{a0}	ψ_{a1}	ψ_{n0}	ψ_{n1}	ψ_{c0}	ψ_{c1}
$\hat{\boldsymbol{\theta}}_1$	0.30	0.10	0.1875	0.0625	0.2625	0.0875
$\hat{\boldsymbol{\theta}}_2$	0.30	0.0875	0.1875	0.0625	0.2625	0.10

di popolazione nei gruppi (*t,z*) per un grande campione sono quelle riportate nella seconda riga della Tabella 4. Ora le quote relative di popolazione appartenente ai tre compliance status sono:

$$\psi_a = (\psi_{a,0} + \psi_{a,1}) = (0.30 + 0.0875) = 0.3875,$$
$$\psi_n = (\psi_{n,0} + \psi_{n,1}) = (0.1875 + 0.0625) = 0.25,$$
$$\psi_c = (\psi_{c,0} + \psi_{c,1}) = (0.2625 + 0.10) = 0.3625,$$

che, a parte piccole differenze, corrispondono alle stime $\hat{\omega}_a, \hat{\omega}_n, \hat{\omega}_c$ in $\hat{\boldsymbol{\theta}}_2$. Considerazioni analoghe valgono per tutti gli altri punti di massimo locale.

Oltre che sulle stime delle probabilità $(\omega_a, \omega_n, \omega_c)$, l'errata scissione delle misture comporta conseguenze prevedibili anche sul resto degli elementi del vettore parametrico. Infatti le stime dei parametri μ_{tz} e σ_{tz} di ogni gruppo (*t,z*) vengono calcolate durante il passo "M" dell'algoritmo EM come stime di massima verosimiglianza ponderata dove ogni unità ha peso uguale alla probabilità di imputazione al gruppo (*t,z*) calcolata al precedente passo "E". Avendo preso in considerazione outcome distribuiti secondo distribuzioni normali, e date le caratteristiche delle probabilità di imputazione già illustrate, allora è comprensibile come ad esempio per la soluzione $\hat{\boldsymbol{\theta}}_2$ si ottengano valori di stima dei parametri $\mu_{a1}, \mu_{c1}, \sigma_{a1}, \sigma_{c1}$ sostanzialmente scambiati rispetto alla soluzione $\hat{\boldsymbol{\theta}}_1$. Questo ragionamento vale per tutte le altre soluzioni. L'analisi delle probabilità di imputazione ha quindi permesso lo studio delle caratteristiche dei punti di massima verosimiglianza locale. Si è visto che oltre ai massimi spuri, facilmente identificabili poiché presentano sempre una componente di varianza prossima a zero, la plurimodalità della funzione di verosimiglianza sia dovuta al label switching.

Si può anche mettere in evidenza come i punti di massimo spuri corrispondono a piccoli gruppi di outliers. Ad esempio, per il punto $\hat{\boldsymbol{\theta}}_5$, al gruppo (*n*,0) vengono assegnate soltanto due unità la cui media è -2.377 e la cui varianza è 0.053; per il punto $\hat{\boldsymbol{\theta}}_8$, al gruppo (*n*,0) viene assegnata soltanto una unità il cui valore è -2.431.

La Tabella 5 mostra le performance della procedura di massimizzazione vincolata ad un intorno del punto $(\hat{\phi}_{a1}, \hat{\phi}_{n0}, \hat{\phi}_{c0}, \hat{\phi}_{c1})$ proposta nella precedente sezione. Per alcuni valori del vincolo *k* (0.03, 0.01, e 0.005) sono state effettuate 100 procedure di massimizzazione vincolata ognuna su di un dataset di numerosità 10000 estratto sempre dalla medesima popolazione ipotetica. Ogni procedura di massimizzazione è stata iniziata con valori casuali del vettore parametrico ad eccezione delle componenti $(\omega_{g0}, \omega_{a1}, \omega_{n0}, \omega_{n1}, \omega_{c0}, \omega_{c1})$ che in partenza vengono sempre poste uguali a $(\hat{\phi}_{a0}, \hat{\phi}_{a1}, \hat{\phi}_{n0}, \hat{\phi}_{n1}, \hat{\phi}_{c0}, \hat{\phi}_{c1})$. Si può osservare come la procedura di massimizzazione vincolata non sempre converge al punto di massimo consistente, ma ciò non costitui-

Tabella 5. *Frequenze assolute dei vari tipi di massimo locale identificati dalla procedure di massimizzazione vincolata per alcuni valori di k (100 replicazioni per ogni valore di k).*

k	Convergenza al massimo consistente	Convergenza a punti sulla frontiera di $\Omega_k^{\hat{\phi}}$	Convergenza a massimi spuri	
			con almeno una comp. var. prossima a zero	dovuti al label switching
0.03	25	73	2	0
0.01	30	68	2	0
0.005	35	63	2	0

sce un problema data la facile individuabilità degli altri punti di massimo locale. La Tabella 5 mostra infatti che l'algoritmo, oltre al punto di massimo consistente, converge anche a punti di massimo spuri con una componente di varianza prossima allo zero, e a punti sulla frontiera dello spazio parametrico vincolato $\Omega_k^{\hat{\phi}}$. Si osservi inoltre come, al diminuire di *k*, aumenta il numero di volte in cui la procedura converge al massimo consistente nelle 100 prove.

Per valutare la bontà della procedura di analisi vincolata (1) presentata nella sezione precedente, sono poi stati estratti 100 dataset artificiali di numerosità 10000 sempre dalla stessa popolazione ipotetica. Per ognuno di questi dataset è stato identificato il punto di massimo interno ad un intorno sferico del punto $(\hat{\phi}_{a1}, \hat{\phi}_{n0}, \hat{\phi}_{c0}, \hat{\phi}_{c1})$ ponendo k=0.01. Sui vettori di stima così ottenuti si è poi provveduto al calcolo per ogni parametro della distorsione media, della radice quadrata dell'errore quadratico medio, dell'ampiezza media dell'intervallo di confidenza al 95% e della frazione di volte che tale intervallo contiene il vero valore del parametro. A fini comparativi sugli stessi dataset artificiali sono state applicate altre procedure standard che non necessitano dell'introduzione di informazioni ausiliarie. Precisamente, sono state calcolate anche:

- le stime di massima verosimiglianza, ipotizzando l'esistenza del vincolo di esclusione in forma debole, ossia imponendo nella (1):
  $\mu_{a1} = \mu_{a0}$, $\mu_{n1} = \mu_{n0}$, $\sigma_{a1} = \sigma_{a0}, \sigma_{n1} = \sigma_{n0}$;
- la stima del C.A.C.E. (Compliers Average Causal Effect), $\mu_{c1} - \mu_{c0}$, ottenuta con il metodo delle variabili strumentali.

I risultati per alcuni parametri sono illustrati nella Tabella 6. Da evidenziare il fatto che sui campioni artificiali estratti dalla popolazione ipotetica l'analisi di massima verosimiglianza condotta sotto il vincolo di esclusione in forma debole non produce un'unica soluzione; per questa ragione anche in questo caso l'analisi è vincolata ad un intorno sferico di $(\hat{\phi}_{a0}, \hat{\phi}_{a1}, \hat{\phi}_{n0}, \hat{\phi}_{n1}, \hat{\phi}_{c0}, \hat{\phi}_{c1})$. Com'era prevedibile l'analisi

Tabella 6. *Performance comparativa della procedura vincolata su 100 dataset ognuno di 10000 unità estratti dalla popolazione ipotetica di cui alla Tabella 1.*

Parametro	Stimatore	Distorsione Media	$\sqrt{MSE}$	Intervallo al 95%	
				Grado di copertura	Ampiezza media
μ_{c0}	ML vincolata	0.002	0.079	0.947	0.312
	ML*	0.204	0.220	0.240	0.306
μ_{c1}	ML vincolata	0.002	0.024	0.991	0.072
	ML*	0.256	0.272	0.237	0.377
σ_{c0}	ML vincolata	0.004	0.041	0.947	0.163
	ML*	0.042	0.088	0.846	0.156
σ_{c1}	ML vincolata	-0.00049	0.054	0.940	0.224
	ML*	-0.006	0.061	0.920	0.216
C.A.C.E.	ML vincolata	0.00011	0.096	0.940	0.368
	ML*	0.051	0.111	0.912	0.368
	IVE**	-1.844	1.857	1.000	15.99

***** stime di massima verosimiglianza ipotizzando l'esistenza del vincolo di esclusione in forma debole;

****** stima del C.A.C.E. (Compliers Average Causal Effect) ottenuta con il metodo delle variabili strumentali.

condotta assumendo il vincolo di esclusione in forma debole soffre di una distorsione media e di un errore quadratico medio sistematicamente maggiore rispetto all'analisi condotta senza vincoli di esclusione, in particolare per quanto riguarda le stime dei parametri relativi alle distribuzioni per i compliers. Ancora peggiore risulta la stima del C.A.C.E. calcolata con il metodo delle variabili strumentali, per la quale si ottiene un alto valore del grado di copertura degli intervalli di confidenza ma al costo di un'ampiezza media esagerata.

4. Considerazioni conclusive

Il lavoro ha inteso mettere in evidenza come alcune problematiche relative all'indebolimento delle condizioni di applicabilità di una metodologia largamente utilizzata nelle valutazioni di efficacia (ossia l'impostazione parametrica all'analisi causale con variabili strumentali) possano essere affrontate facendo riferimento alla teoria delle misture di distribuzioni. In tal senso si è proposto l'uso di una procedura di analisi di massima verosimiglianza vincolata; una successiva analisi di tipo simu-

lativo ha poi consentito un primo giudizio sulla bontà della proposta. L'estrazione ripetuta di campioni causali semplici da una popolazione ipotetica ha evidenziato una buona performance anche comparativamente ad altri metodi usuali. Resta però il fatto che l'analisi simulativa è stata condotta basandosi su di una sola popolazione ipotetica di riferimento. Appare quindi interessante un'eventuale approfondimento basato su altre popolazioni ipotetiche che possa evidenziare aspetti più difficoltosi nelle analisi di misture, come ad esempio un peggior grado di scissione conseguente ad una maggiore vicinanza dei valori delle medie e delle varianze delle componenti le misture.

Riferimenti bibliografici

CARD D., T. LEMIEUX (2001) Can falling supply explain the rising return to college for younger men? A cohort-based analysis, *Quaterly Journal of Economics*, **116**: 705-746.

DAY N.E. (1969) Estimating the components of a mixture of normal distributions, *Biometrika*, **56**: 463-474.

HJORT N.L. (1986) Contribution to the discussion of paper by P.Dianconis and D.Freedman, *The Annals of Statistics*, **14**: 49-55.

HOLGERSSON M., U. JORNER (1998) Decomposition of a mixture into normal components: a review, *International Journal of Biomedical Computing*, **29**: 367-392.

IMBENS G.W., J.ANGRIST (1994) Identification and estimation of local average treatment effects; *Econometrica*, **62**: 467-476.

IMBENS G.W., D.R. RUBIN (1997) Bayesian inference for causal effects in randomized experiments with non-compliance, *The Annals of Statistics*, **25**: 305-327.

KIEFER M. (1978) Discrete parameter variation: efficient estimation of a switching regression model, *Econometrica*, **46**: 427-439.

MERCATANTI A. (2004) Causal inference methods without exclusion restrictions: an economic application, *Report n.250 del Dip. di Statistica e Matematica Applicata all'Economia, Università di Pisa.*.

PRIEBE C.E. (1994) Adaptive mixtures, *J.A.S.A.*, **89**: 796-806.

The importance of Mixture models in efficacy evaluation with causal methods

Summary: *Some methodological issues regarding the weakening of the assumptions usually adopted for causal inference methods concern the analysis of mixture models. In particular, this paper considers the complete relaxation of the exclusion restriction when using the instrumental variables method for identifying and estimating causal effects. We are supposing a binary treatment and a normally distributed outcome. With respect to a standard analysis of mixture models, we can exploit a larger set of a priori information in particular as concerns the mixing proportions; conversely, the presence of common distribution mixtures produces a likelihood function having more than one maximum point. This paper also takes into account a constrained maximisation procedure that uses the greater information regarding the probability of group belonging, in order to resolve the problems tied to the multiple mode of the likelihood function.*
Keywords: *Normal mixtures, instrumental variables, exclusion restriction.*