Statistica & Applicazioni
Vol. III, n. speciale 1, 2005
pp. 93-100

Alcune note sull'impiego delle isomappe nell'analisi della soddisfazione degli utenti di servizi sanitari

Marilena Pillati [§]

Angela Montanari [§]

Erika Massimiliani[§]

Summary: *The paper deals with non linear dimension reduction methods in presence of ordinal data. More precisely, our attention focuses on isomaps, a multidimensional scaling method based on the so called geodesic distance, which connects a pair of units along a path which goes trough its k nearest neighbours. We propose a criterion for the choice of k, discuss some issues posed by the use of isomaps for ordinal data and present an application to the study of health service user satisfaction.*

Keywords: *Non linear dimension reduction, multidimensional scaling, geodesic distance, isomap.*

1. Introduzione

Il problema della riduzione delle dimensioni rappresenta uno dei temi più ampiamente discussi nella letteratura statistica multivariata. In generale le motivazioni sono legate ai vantaggi interpretativi che la rappresentazione di un fenomeno attraverso un numero ristretto di caratteristiche offre e talvolta sono dettate anche dalla impossibilità di trattare analiticamente lo spazio di dimensioni elevate. Le soluzioni devono coniugare il rigore metodologico con l'effettiva capacità di riassumere in poche variabili tutto il contenuto informativo presente nei dati originali.

I metodi esplorativi classici quali l'analisi delle componenti principali e il *multidimensional scaling* metrico consentono di trarre indicazioni sulla struttura lineare sottostante i dati osservati, ma falliscono in presenza di non linearità. Un esempio di questa situazione è rappresentato in figura 1.

Recentemente sono stati proposti alcuni metodi di proiezione di tipo non lineare che derivano in modo più o meno diretto dal *multidimensional*

[§] Dipartimento di Scienze Statistiche - Università di Bologna - via Belle Arti, 41, 40126 BOLOGNA (e-mail: marilena.pillati@unibo.it, angela.montanari@unibo.it, massimiliani@stat.unibo.it).

scaling classico, ma sostituiscono alla distanza euclidea una diversa metrica che meglio si presta alla misura della prossimità tra osservazioni che giacciono su sottospazi non lineari. Il riferimento è alle cosiddette Isomappe di Tenenbaum *et al.* (2000) e alla *Curvilinear Distance Analysis* di Lee *et al.* (2002).

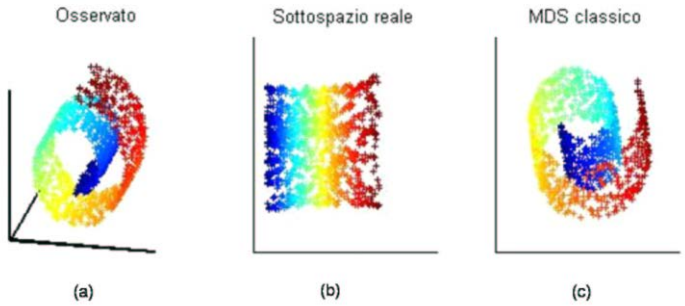

Figura 1. *Swiss roll: i dati osservati giacciono su una spirale nello spazio tridimensionale (a), ma la reale dimensionalità del data set è bidimensionale (b). Il multidimensional scaling non riesce a riconoscere la vera struttura bidimensionale (c).*

Entrambi i metodi nascono in un'ottica di analisi di dati quantitativi e si sono rivelati in grado di cogliere particolari strutture sottostanti i dati osservati che si manifestano negli stessi in forme non lineari.

La presenza di non linearità può caratterizzare anche variabili osservate ordinali, quali sono quelle tipiche dei questionari per valutare la soddisfazione degli utenti di servizi, ad esempio quelli offerti da una struttura sanitaria. L'impiego dei metodi sopra menzionati in questo contesto è ancora totalmente inesplorato e pone questioni di carattere metodologico e interpretativo che saranno oggetto di riflessione nel presente lavoro. Per la natura del metodo, che lo rende più facilmente adattabile all'analisi di dati ordinali, ci si concentrerà sulle Isomappe piuttosto che sulla *Curvilinear Distance Analysis.*

2. Le isomappe

L'idea principale alla base delle Isomappe (Tenenbaum *et al.*, 2000) è quella di superare i limiti del *multidimensional scaling* classico, che è lineare, sostituendo alla distanza euclidea la cosiddetta distanza di geodesia.

A partire dalla matrice delle distanze euclidee, la determinazione della distanza di geodesia richiede innanzitutto di individuare, per ciascuna unità, le k unità a essa più vicine, o in alternativa le unità che giacciono in un suo intorno di diametro d (con k e d definiti a priori). Successivamente si modifica la matrice delle distanze: solo quelle relative a unità comprese

nell'insieme dei *k* vicini restano inalterate e coincidenti con le distanze euclidee, mentre quelle relative a punti lontani sono sostituite dalla lunghezza del percorso più breve che collega i due punti, passando attraverso i vicini. La ricerca del percorso di minima lunghezza che congiunge due distinte unità può essere affrontata mediante gli algoritmi di Floyd o di Dijkstra, che consentono quindi di ottenere la matrice delle distanze di geodesia.

Lo spazio di dimensioni ridotte viene poi ricostruito applicando il *multidimensional scaling* metrico alla matrice delle distanze di geodesia così ottenuta.

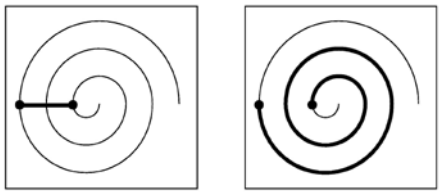

Figura 2. *Spirale: in grassetto a) distanza euclidea (a sinistra) e b) distanza di geodesia (a destra)*

La figura 2 illustra in modo chiaro la differenza tra le due metriche, quella euclidea e quella di geodesia, e una situazione tipica in cui il *multidimensional scaling* metrico non riesce a evidenziare il reale spazio ridotto su cui giacciono i dati: la spirale è bidimensionale, ma la dimensione sottostante è unidimensionale.

3. Problemi aperti nell'uso delle isomappe per dati ordinali

3.1 *Alcuni criteri per la scelta di k*

L'impiego in concreto delle isomappe richiede di definire il numero *k* di unità che compongono ciascun intorno e, una volta ricostruita la matrice delle distanze di geodesia, di determinare la dimensione del sottospazio "reale" in cui si collocano i dati. I due aspetti sono fortemente connessi: scelte diverse per *k* possono condurre alla individuazione di sottospazi di diversa dimensionalità, con strutture diverse da quella reale. Osservando il comportamento delle isomappe nell'analisi dello *swiss roll* si nota che, per valori di *k* compresi tra 2 e 3 l'algoritmo non riesce a riprodurre le distanze di geodesia per tutte le unità. Per valori di *k* compresi tra 4 e 10 si ricostruisce il reale sottospazio bidimensionale. Per valori di *k* superiori a 10 l'algoritmo identifica una struttura sottostante diversa da quella che ha

effettivamente generato i dati.

Poiché i due problemi, scelta di *k* e scelta della dimensione dello spazio di proiezione, non possono essere affrontati in modo simultaneo, in ragione della loro circolarità, opereremo in modo sequenziale scegliendo prima un valore di *k* "ottimo" e successivamente determinando la corrispondente dimensione del sottospazio.

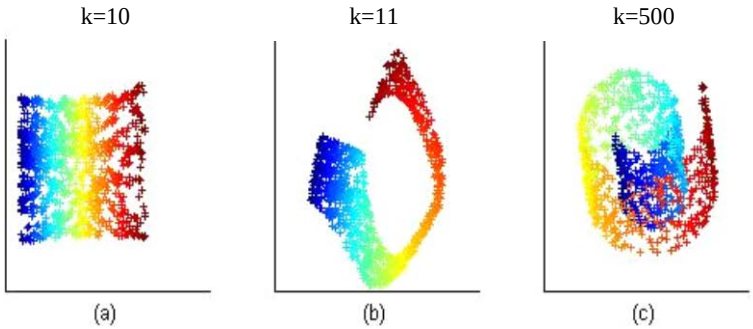

Figura 3. *Proiezioni bidimensionali dello swiss roll ottenute con le isomappe per diversi valori di k. In (c) la struttura bidimensionale ottenuta con k=500 coincide con quella ottenuta utilizzando il multidimensional scaling metrico (figura 1)*

E' opportuno osservare che un valore di *k* piccolo non consente di costruire le distanze di geodesia per tutte le coppie di unità e preserva la struttura locale dei dati, mentre un valore di *k* elevato non fa altro che riprodurre la matrice delle distanze di partenza e annulla il vantaggio della distanza di geodesia. Ancora, per quanto mostrato in precedenza, ci si aspetta che, in presenza di non linearità le distanze di geodesia siano, rispetto alle distanze euclidee, mediamente più elevate, più variabili, con distribuzione asimmetrica. Queste considerazioni conducono alla proposta di criteri che si configurano anche come strumenti empirici di diagnosi della non linearità.

Per ciascun valore di *k*, si può determinare, ad esempio, il rapporto tra la media aritmetica delle distanze di geodesia e la media delle distanze euclidee a partire dalle quali la distanza di geodesia è stata costruita e rappresentare le quantità in un grafico cartesiano, così da ottenere una sorta di *scree diagram*. Ci si aspetta che, in presenza di nonlinearità, il grafico mostri uno o più salti in corrispondenza dei valori di *k* che generano un cambiamento nella struttura della proiezione dei dati nello spazio ridotto.

Anche il rapporto tra la varianza delle distanze di geodesia e la varianza delle distanze euclidee, o tra le asimmetrie o i *range* delle stesse possono fornire utili indicazioni allo stesso fine.

In figura 4 è riportato l'andamento dei quattro indicatori per lo *swiss roll* e per un insieme di dati generato da una normale multivariata e, quindi,

caratterizzato unicamente da una struttura lineare. I grafici riflettono perfettamente il comportamento dell'algoritmo delle isomappe nello *swiss roll* al variare di *k*, come precedentemente descritto, e mostrano una sensibilità dei criteri alla presenza o meno di non linearità.

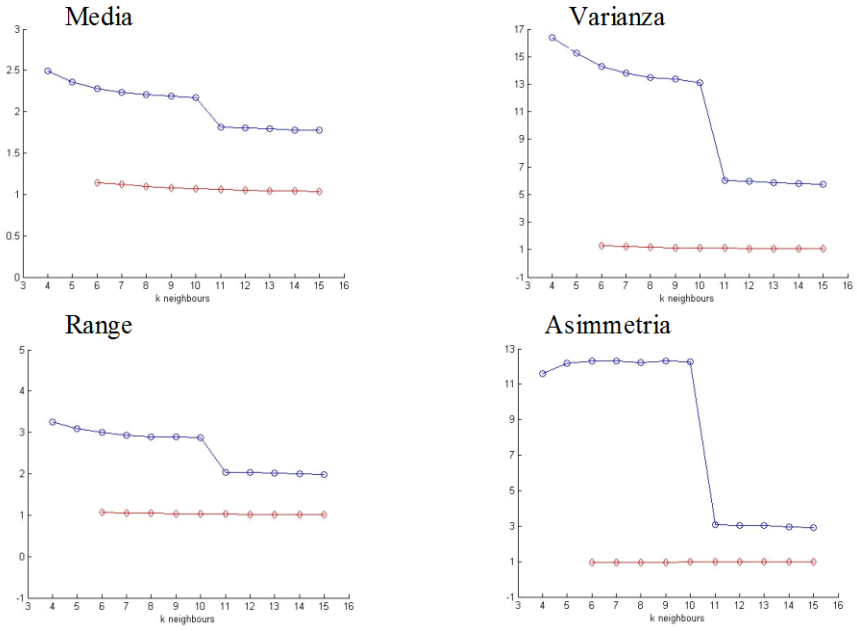

Figura 4. *Andamento dei 4 indicatori al variare di k per lo swiss roll (linea superiore) e per dati gaussiani (linea inferiore).*

Qualora le isomappe vengano impiegate per l'analisi di dati ordinali, oltre al problema della scelta di *k*, che può essere affrontato secondo i metodi appena descritti, si pone l'ulteriore problema che il minimo valore di *k* che consente di ricostruire tutte le distanze di geodesia è sempre molto grande a causa dell'elevato numero di unità che possono presentare lo stesso pattern di risposte. Una soluzione possibile consiste nel perturbare i dati in modo da eliminare i valori ripetuti aggiungendo alle risposte valori da una normale N(0;0.001).

3.2 *Scelta della metrica*

Il contesto in cui le isomappe sono state sviluppate è quello tipico del riconoscimento di immagini. Dovendo trattare dati metrici, la scelta naturale

per misurare le prossimità nello spazio originale è la distanza di geodesia costruita a partire da quella euclidea tra coppie di unità, e la stessa distanza euclidea rappresenta anche la metrica naturale per la proiezione dei dati sul sottospazio di dimensioni ridotte, giustificando in questo modo il ricorso al *multidimensional scaling* metrico come strumento per la riduzione dimensionale.

Lo studio di variabili ordinali pone innanzitutto il problema della scelta della misura di prossimità tra coppie di unità statistiche più coerente con la natura stessa delle informazioni. Il coefficiente semplice di dissomiglianza, che rapporta il numero di caratteri su cui due unità presentano livelli differenti al numero totale di caratteri osservati rappresenta una delle scelte più diffuse nella letteratura, ma presenta il ben noto limite di trascurare totalmente la natura ordinale del carattere.

La quantificazione del carattere ordinale secondo *l'underlying variable approach*, fondato sull'ipotesi che i livelli osservati altro non siano che l'espressione discreta di una sottostante variabile gaussiana, mal si presta a descrivere dati di soddisfazione che presentano, nella maggior parte delle circostanze, una distribuzione fortemente asimmetrica.

Una soluzione parzialmente coerente con la natura dei dati, e totalmente svincolata da qualsivoglia assunzione sulla legge distributiva, può essere rappresentata dall'impiego della distanza *city block* tra coppie di unità statistiche calcolata dopo avere codificato le modalità di ciascun carattere con i numeri naturali da 1 al numero massimo di modalità che esso presenta (si veda, ad esempio, Zani, 2000). La distanza euclidea sui dati codificati allo stesso modo può comunque essere impiegata come termine di paragone.
Tali soluzioni, di diffuso impiego, sottintendono tuttavia l'ipotesi di equidistanza tra le modalità delle singole variabili, che solo raramente trova effettivo riscontro nell'analisi empirica.

4. Analisi di dati di soddisfazione degli utenti di un servizio sanitario

Un'analisi delle performance delle isomappe per la riduzione delle dimensioni in presenza di caratteri ordinali è stata condotta su dati raccolti nel 2003 all'interno di un progetto nato dalla collaborazione tra il Policlinico S. Orsola-Malpighi e il Dipartimento di Scienze Statistiche dell'Università di Bologna, volto a ottenere un giudizio complessivo di soddisfazione sull'esperienza vissuta dai pazienti a contatto con la struttura sanitaria.

I dati derivano da un questionario autocompilato da parte dei pazienti ricoverati, composto da una serie di *item* con risposte espresse su scala Likert a cinque modalità, da pessimo a ottimo, e da alcune domande relative alle caratteristiche socio-demografiche. Ciascun paziente è stato chiamato a esprimere un giudizio complessivo di soddisfazione e giudizi specifici su aspetti fondamentali della degenza ospedaliera quali: gentilezza e

disponibilità del personale medico (2 *item*), accoglienza al momento del ricovero, gentilezza e disponibilità del personale infermieristico (3 *item*), informazioni ricevute (4 *item*), comfort alberghiero (5 *item*).

Si sono analizzate le performance del *multidimensional scaling* metrico, variando la misura utilizzata per valutare la prossimità tra pazienti rispetto alle variabili di soddisfazione sopra descritte. In particolare si è scelto di impiegare la distanza euclidea (perché elemento fondamentale del *multidimensional scaling* metrico nella sua versione classica), la distanza *city block* e il coefficiente semplice di dissomiglianza, nonché le distanze di geodesia da esse derivate.

I criteri descritti nel paragrafo 3.1 hanno mostrato, per le distanze di geodesia ottenute a partire da tutte le metriche, un andamento monotono decrescente senza salti, a indicare una sostanziale linearità nei dati (Figura 5), confermata anche dai risultati del *multidimensional scaling*. Per diversi valori di *k* tutte le analisi hanno identificato in 2 la dimensione dello spazio ridotto su cui proiettare i dati, e hanno prodotto vettori dei coefficienti per tali dimensioni praticamente coincidenti quando le metriche utilizzate erano la distanza euclidea, la distanza *city block* e le distanze di geodesia da esse derivate, ma profondamente differenti per il coefficiente semplice di dissomiglianza.

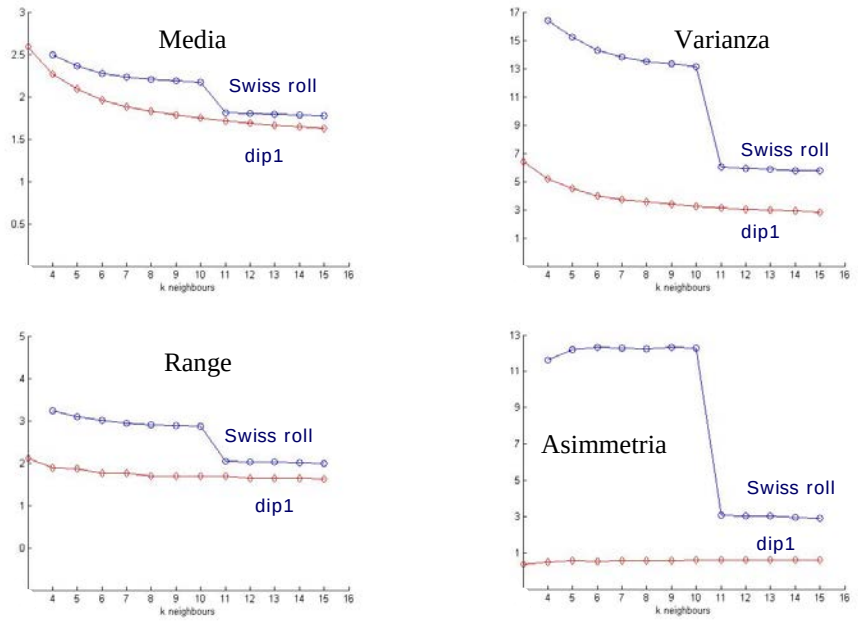

Figura 5. *Andamento dei 4 indicatori al variare di k per lo swiss roll e per i dati di uno dei dipartimenti della struttura ospedaliera analizzata.*

Per interpretare le dimensioni ottenute si sono calcolati i coefficienti di correlazione per ranghi tra ciascuna di esse e le singole variabili che compongono il questionario.

I risultati peggiori sono relativi all'impiego del coefficiente semplice di dissomiglianza, sia nella sua formulazione originale sia in quella che lo vede come base per la distanza di geodesia. La sua incapacità di tener conto dell'ordinamento implicito nelle modalità dei caratteri osservati fa sì che le due dimensioni ottenute presentino correlazione per ranghi pressoché nulla con tutte le domande del questionario.

Molto simili tra loro risultano invece i risultati ottenuti con i restanti quattro metodi: la prima dimensione presenta correlazioni per ranghi positive e elevate con tutte le variabili osservate e può quindi essere interpretata come espressione di un livello generale di soddisfazione; la seconda è un contrasto tra le variabili che descrivono il cosiddetto "comfort alberghiero" e i restanti aspetti indagati dal questionario.

Questa somiglianza di risultati sembra indicare, da un lato, l'ininfluenza del tipo di metrica utilizzata, a patto che questa tenga conto della natura ordinale delle variabili rilevate e, dall'altro, la tendenziale linearità della struttura presente nei dati.

Ringraziamenti

Gli autori desiderano ringraziare il Settore Comunicazione e Informazione del Policlinico S.Orsola – Malpighi per aver gentilmente concesso l'uso dei dati.

Lavoro svolto nell'ambito del progetto di ricerca nazionale prin2004 dal titolo "Nuovi metodi statistici multivariati di classificazione e riduzione dimensionale per la valutazione e la *customer satisfaction* nei servizi" (Coordinatore nazionale: Prof. Maurizio Vichi).

Riferimenti Bibliografici

Borg I., Groenen P. (1997), *Modern Multidimensional Scaling – Theory and Application*, Springer-Verlag, New York.

Lee J. A., Landasse A., Verleysen M. (2002), Curvilinear distance analysis versus ISOMAP, in M. Verleysen (Eds.), *ESANN'2002 Proceedings European Symposium on Artificial Neural Networks*, Bruges, Belgium: D-Side Publications, pp.185-192.

Tenenbaum J., V. de Silva, J. Langford (2000), A global geometric framework for non-linear dimensionality reduction, *Science*, **290**, n. 5500, pp. 2319–2323.

Zani S. (2000), *Analisi dei dati statistici*, Vol II, Giuffrè editore, Milano.