UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI INFORMATICA

Dottorato di ricerca in Informatica e Matematica
XXXI ciclo

Progetto di ricerca

Dottorando: Dott.ssa *Flavia Esposito*
Tutor: Prof.ssa *Nicoletta Del Buono*

Firma del dottorando _____

Firma del tutor _____

1 Titolo della ricerca:

Tecniche di decomposizione low-rank per analisi di dati biomedici

2 Area nella quale si inquadra la ricerca:

Metodi numerici per l'algebra lineare

3 Obiettivi della ricerca:

La ricerca che si intende svolgere sarà focalizzata sullo studio dei meccanismi di approssimazione low rank di dati strutturati al fine di sviluppare nuovi algoritmi di approssimazione più efficienti e adatti a condurre delle analisi esplorative di dati di grandi dimensioni (quali ad esempio microarray di espressioni geniche o immagini di risonanze magnetiche) derivanti da contesti sensibili quali quello biomedico. Particolare attenzione sarà rivolta alle tecniche di fattorizzazione non negativa (matriciali e/o tensoriali) di cui si analizzeranno i fattori di maggiore criticità, quali le problematiche legate alla scelta del rango della fattorizzazione (che determina il numero di fattori latenti), le performance degli algoritmi in base alle diverse inizializzazioni, le possibili connessioni con le reti e i grafi.
Obiettivo cardine della ricerca sarà determinare una metodologia ottimale per estrarre –dall'insieme dei dati oggetto di analisi– fattori latenti semanticamente rilevanti e in grado di rappresentare al meglio l'insieme dei dati analizzato.

4 Motivazioni della ricerca

Durante l'ultimo ventennio sono state sviluppate sempre più tecniche di storage di dati con un consequenziale aumento di informazioni immagazzinate. Per analizzare e estrarre informazioni rilevanti all'interno di questi numerosissimi dati è stato quindi necessario sviluppare parallelamente tecniche di analisi ad hoc.
Lo sviluppo di tecnologie innovative per l'immagazzinamento e la raccolta di informazioni ha colpito anche il campo biomedico. Esempi rilevanti sono rappresentati dall'evoluzione delle tecniche diagnostiche tramite immagini e dalla nascita della biologia molecolare, soprattutto nei primi anni '90 con la produzione dei microarray, vetrini delle dimensioni di un francobollo, su cui viene depositato il DNA di moltissimi geni di organismi esposti a diverse condizioni oppure di organismi sani o malati e che quindi permettono di estrarre data set contenenti l'intero corredo genico di un organismo.
Il trattamento analitico di dati biomedici può essere effettuato in due direzioni: interrogando gli elementi tramite test di verifica delle ipotesi e quindi tramite la statistica parametrica e/o non parametrica oppure effettuando un'analisi esplorativa dei dati.
In questo scenario le fattorizzazioni matriciali (o tensoriali quando si vogliono mettere in relazione molteplici parametri) ricoprono un ruolo essenziale per scoprire i fattori latenti nella struttura del data set analizzato. Nello specifico quindi, le tecniche di approssimazione low rank, permettono di proiettare i dati in uno spazio di dimensionalità ridotta aumentando notevolmente la capacità interpretativa e la comprensione reale che questi celano nelle fittissime reti in cui sono intrecciati, dando quindi la possibilità agli esperti del dominio di leggere tutto il fenotipo rappresentato dallo studio in esame in chiave di una manciata di elementi.

5 Stato dell'arte

L'evoluzione tecnologica per la raccolta e la conservazione dei dati provenienti da un esperimento scientifico, come già descritto nel punto 4, sta influenzando notevolmente sia l'ambiente quotidiano che quello di ricerca delle ultime generazioni.
Questo fenomeno ha colpito svariati campi dal telerilevamento alle indagini statistiche, dalla analisi di pagine web alle elaborazioni di immagini fino anche al campo biomedico. In questo scenario infatti, nell'ultimo ventennio, si sono sviluppate sia tecniche di biologia molecolare, in grado di raccogliere una grandissima quantità di dati numerici legati ai fattori genici, che tecniche (come il compressive sensing) per le esplorazioni di immagini provenienti da risonanze magnetiche e/o TAC.
In generale l'ammontare di grandi quantità di dati è sfociata nella nascita di una disciplina, la Knowledge Discovery in Databases (KDD), intenta a fornire un processo automatico di esplorazioni dei dati allo scopo di identificare pattern validi, utili e apparentemente ignoti. Nello specifico gli algoritmi e le tecniche per esplorare grandi quantità di dati alla ricerca di tendenze consistenti e/o relazioni sistematiche rappresenta il cuore del processo del KDD ed è noto in letteratura con il nome di Data Mining.
La maggior parte dei dati biomedici sono dati strutturati pertanto, al fine di studiarli matematicamente è possibile rappresentarli come matrici, o come tensori se si vogliono studiarne le relazioni in base a più parametri, con lo scopo di ridurne la dimensionalità e di estrarre informazioni significative [48].
Finora le tecniche di apprendimento non supervisionato più utilizzate per le decomposizioni matriciali si dividono in metodi analitici (Decomposizione a valori singolari, SVD o Analisi delle componenti principali, PCA) e in metodi iterativi (come l' Analisi delle Componenti indipendenti ICA, la Network Component Analysis e le Nonnegative Matrix Factorization NMF) [20], [18].
L'obiettivo comune di queste tecniche è quello di cambiare lo spazio in cui lavorare, catturando dei vettori che costituiranno la base del nuovo sottospazio, contentente tutte le informazioni più significative ai fini dell'analisi del problema. Questa prima fase di estrapolazione dei vettori della base prende il nome di fase di apprendimento off-line ed è seguita dalla fase di riconoscimento on-line, nella quale viene ricostruito il modello proiettandolo nel sottospazio trovato in precedenza. Le tecniche di riduzione di dimensionalità determinano in generale un'approssimazione V della matrice di dati X

$$X = V + E$$

ove con E si è denotata la matrice errore che tiene conto dell'approssimazione che si vuole calcolare.
In letteratura sono stati studiati vari metodi di riduzione low-rank per l'analisi di dati biomedici soprattutto per quanto riguarda l'analisi di microarray [31].
La SVD e la PCA furono applicate per la prima volta ai microarray in [3], per analizzare i dati sul ciclo cellulare del lievito [50]. La SVD genera due basi ortonormali, una definita dai vettori singolari destri e l'altra dai vettori singolari sinistri e dei valori singolari in grado di pesare l'importanza della decomposizione e dell'eventuale troncamento [21]. La PCA lavora in maniera simile alla SVD, ma si basa sulla matrice di covarianza supportando l'idea che a autovalori maggiori corrispondono maggiori informazioni trasmesse dal dataset iniziale. Sebbene entrambi i metodi possiedano da un lato una forte base statistica e analitica in grado di determinare a priori la dimensionalità del problema, dall'altro lato vi è però la richiesta dell'ortogonalità della base che in realtà non ha nessun significato biologico, il che tende a produrre ulteriori errori oltre a quelli di misura.

L'ICA nel costesto dell'analisi dei microarray fu introdotta indipendentemente da Lin [37] e da Liebermeister [36]. Quest'ultimo confrontò la ICA con la PCA, dimostrando che l'eliminazione dell'ortogonalità della base poteva sicuramente rappresentare una migliore interpretazione biologica. Tuttavia veniva introdotta l'ipotesi di indipendenza statistica tra le componenti, perdendo così la perfetta applicabilità ai dati di espressione genica.
La NMF fu introdotta per la prima volta da Paatero e Tapper [42] con il nome di Positive matrix factorization (PMF) nel 1994 nel contesto dei dati ecologici e fu successivamente sviluppata da Lee e Seung [33] per il riconoscimento delle immagini. Tuttavia fu solo all'inizio del XXI secolo che iniziarono a essere sviluppate varie applicazioni della NMF allo studio delle espressioni geniche, da Kim e Tidor [28] e Brunet [6] nel contesto dei dati biomedici.
Questa tecnica dal punto di vista matematico si presenta come un problema di ottimizzazione vincolata in cui si considera una specifica funzione d'errore. La maggior parte degli algoritmi NMF utilizza due funzioni errore: la distanza Euclidea e la divergenza di Kullback-Leibler. Tuttavia vi sono esempi di algoritmi che utilizzano altri tipi di funzioni obiettivo, come varianti della divergenza di Bergman o varianti studiate in base a vincoli aggiuntivi al fine di controllare la regolarità dei fattori o la sparsità degli stessi [51].
Dal punto di vista pratico la NMF, imponendo il vincolo di non negatività dei fattori, presenta essenzialmente due grandi vantaggi nell'applicazione a dati reali, in primo luogo vi è la possibilità di associare un significato biologico direttamente alle matrici che caratterizzano la fattorizzazione, in secondo luogo vi è la possibilità di ricostruire l'intero data set con tecniche additive, il che si pone nel naturale contesto dell'apprendimento umano, ovvero quello di riconoscere un oggetto esclusivamente come somma di singole parti [34].
Con il tempo sono state sviluppate diverse varianti della NMF per l'ambito biomedico e non, imponendo vincoli aggiuntivi al problema iniziale. Per esempio Carmone-Saez et al. proposero nel 2006 [9] una variante della tecnica NMF per il biclustering di dati di espressioni geniche imponendo il vincolo di ortogonalità dei fattori e applicando in campo biomedico ciò che era stato precedentemente teorizzato con la Block Value Decomposition [39]. Un' altra applicazione al campo biomedico vede protagonista una variante della NMF, la NetNMF [7] . Questa viene utilizzata dal team di Hofree [23] per la stratificazione delle mutazioni dei tumori basata su reti, come naturale strumento per clusterizzare i geni preservando la struttura e le connessioni della rete.
Altre varianti della tecnica applicate a questo campo possono essere ricercate nella bibliografia elencata nel punto 8.

6 Approccio al problema

Operativamente si procederà in primo luogo con un approfondito studio bibliografico degli articoli e delle tecniche proposte in letteratura, focalizzandosi sulle ultime metodologie per l'applicazione al campo di interesse.
Pensando all'ambito applicativo del progetto di ricerca, si cercherà di estendere le tecniche di fattorizzazioni low-rank al caso di reti. Queste possono essere intese come uno strumento matematico espressione della natura, considerando i descrittori e gli oggetti di un data set come nodi e il valore che questi assumono nello stesso come peso del loro collegamento.
Trovandosi in un contesto globale di apprendimento non supervisionato si cercherà inoltre di effettuare una profonda analisi delle diverse tecniche di clustering e dei loro meccanismi di funzionamento, in modo da confrontare e valutare i punti di forza e le eventuali debolezze per una scelta ottimale applicabile al contesto di interesse del progetto di ricerca.

Si analizzeranno anche gli algortimi proposti in letteratura valutandone la sensibilità alle differenti inizializzazioni e cercando misure di valutazione ottimale per la stima a priori della dimensionalità compatibili con l'ambito applicativo biomedico.

7 Ricadute applicative

Lo sviluppo di decomposizioni low-rank per l'analisi di dati biomedici ha molteplici ricadute applicative in dipendenza dalla tipologia del dato analizzato, grazie alla possibilità di ridurre gli insiemi di oggetti e descrittori presi in esame.
Per esempio analizzando le espressioni geniche di particolari cellule tipiche di un determinato tessuto prelevato in diverse condizioni (sano o malato) sarebbe possibile estrapolare geni che sono rappresentativi della malattia (o in generale della situazione oggetto di esame) che, a partire dalle forma di sano, siano già rappresentativi della progressione della malattia.
Un'altra ricaduta applicativa potrebbe essere ottenuta dall'analisi di immagini biomediche provenienti da TAC o risonanze magnetiche, applicando tecniche di compressive sensing per ottenere immagini ad alta risoluzione basandosi su pochi componenti.
Un ulteriore scenario realizzabile potrebbe essere l'analisi di profili di espressione genica di tessuti tumorali relazionati a particolari farmaci, grazie alla quale si potrebbero estrapolare i geni più influenzati da un particolare set di farmaci somministrato per quella particolare patologia. In questo contesto la tecnica di decomposizione sarebbe un punto di partenza per uno screening sulle attività dei farmaci selezionati, ponendosi come una promettente strategia chimica genomica per selezionare modulatori di complesse malattie per determinarne il meccanismo di azione.

8 Riferimenti bibliografici

[1] Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, 2011.

[2] Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications*. CRC Press, 2013.

[3] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[4] Åke Björck. The calculation of linear least squares problems. *Acta Numerica*, 13:1–53, 2004.

[5] Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[6] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

[7] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 63–72. IEEE, 2008.

[8] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[9] Pedro Carmona-Saez, Roberto D Pascual-Marqui, Francisco Tirado, Jose M Carazo, and Alberto Pascual-Montano. Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC bioinformatics*, 7(1):1, 2006.

[10] G Casalino, N Del Buono, and C Mencar. Nonnegative matrix factorizations for intelligent data analysis. In *Non-negative Matrix Factorization Techniques*, pages 49–74. Springer, 2016.

[11] Gabriella Casalino, N Del Buono, and Massimo Minervini. Nonnegative matrix factorizations performing object detection and localization. *Applied Computational Intelligence and Soft Computing*, 2012:15, 2012.

[12] Eric C Chi and Tamara G Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

[13] Moody Chu and RJ Plemmons. Nonnegative matrix factorization and applications. *Image*, 34:1–5, 2005.

[14] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[15] Nicoletta Del Buono and Gianvito Pio. Non-negative matrix tri-factorization for co-clustering: An analysis of the block matrix. *Information Sciences*, 301:13–26, 2015.

[16] Karthik Devarajan. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 4(7):e1000029, 2008.

[17] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, page None, 2003.

[18] Lars Eldén. *Matrix methods in data mining and pattern recognition*, volume 4. SIAM, 2007.

[19] Yuan Gao and George Church. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, 21(21):3970–3975, 2005.

[20] GH Golub and CF Van Loan. Matrix computations, 4th. *Johns Hopkins*, 2013.

[21] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[22] Kevin E Heinrich, Michael W Berry, and Ramin Homayouni. Gene tree labeling using nonnegative matrix factorization on biomedical literature. *Computational intelligence and neuroscience*, 2008:2, 2008.

[23] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–1115, 2013.

[24] Lucie N Hutchins, Sean M Murphy, Priyam Singh, and Joel H Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 24(23):2684–2690, 2008.

[25] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[26] Hyunsoo Kim and Haesun Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.

[27] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[28] Philip M Kim and Bruce Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome research*, 13(7):1706–1718, 2003.

[29] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[30] Wei Kong, Charles R Vanderburg, Hiromi Gunshin, Jack T Rogers, and Xudong Huang. A review of independent component analysis application to microarray gene expression data. *Biotechniques*, 45(5):501, 2008.

[31] Andrew V Kossenkov and Michael F Ochs. Matrix factorisation methods applied in microarray data analysis. *International journal of data mining and bioinformatics*, 4(1):72–90, 2010.

[32] Da Kuang, Haesun Park, and Chris HQ Ding. Symmetric nonnegative matrix factorization for graph clustering. In *SDM*, volume 12, pages 106–117. SIAM, 2012.

[33] Daniel D Lee and H Sebastian Seung. Unsupervised learning by convex and conic coding. *Advances in neural information processing systems*, pages 515–521, 1997.

[34] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[35] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[36] Wolfram Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[37] Simon M Lin, Xuejun Liao, Patrick McConnell, Korkut Vata, Lawrence Carin, and Pascal Goldschmidt. Using functional genomic units to corroborate user experiments with the rosetta compendium. In *Methods of Microarray Data Analysis II*, pages 123–137. Springer, 2002.

[38] Weixiang Liu, Kehong Yuan, and Datian Ye. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. *Journal of biomedical informatics*, 41(4):602–606, 2008.

[39] Bo Long, Zhongfei Mark Zhang, and Philip S Yu. Co-clustering by block value decomposition. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 635–640. ACM, 2005.

[40] Wing-Kin Ma, José M Bioucas-Dias, Tsung-Han Chan, Nicolas Gillis, Paul Gader, Antonio J Plaza, ArulMurugan Ambikapathi, and Chong-Yung Chi. A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. *Signal Processing Magazine, IEEE*, 31(1):67–81, 2014.

[41] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.

[42] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[43] Alberto Pascual-Montano, Jose Maria Carazo, Kieko Kochi, Dietrich Lehmann, and Roberto D Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):403–415, 2006.

[44] Xiaoyan A Qu and Deepak K Rajpal. Applications of connectivity map in drug discovery and development. *Drug discovery today*, 17(23):1289–1298, 2012.

[45] Alex Sánchez and M Carme Ruíz de Villa. A tutorial review of microarray data analysis. *Bioinfotmatics Tutorial, Universitat de Barcelona*, 2008.

[46] Berkant Savas, Inderjit S Dhillon, et al. Clustered low rank approximation of graphs in information science applications. In *SDM*, pages 164–175. SIAM, 2011.

[47] Reinhard Schachtner, Dominik Lutter, P Knollmüller, Ana Maria Tomé, Fabian J Theis, Gerd Schmitz, Martin Stetter, P Gómez Vilda, and Elmar Wolfgang Lang. Knowledge-based gene expression classification via matrix factorization. *Bioinformatics*, 24(15):1688–1697, 2008.

[48] David Skillicorn. *Understanding complex datasets: data mining with matrix decompositions*. CRC press, 2007.

[49] Zullini A. Sparvoli A., Sparvoli F. *Biochimica*. Istituto Italiano, Edizioni Atlas, 2014.

[50] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

[51] Suvrit Sra and Inderjit S Dhillon. *Nonnegative matrix approximation: Algorithms and applications*. Computer Science Department, University of Texas at Austin, 2006.

[52] Dov Stekel. *Microarray bioinformatics*. Cambridge University Press, 2003.

[53] Erick Suárez, Ana Burguete, and Geoffrey J McLachlan. Microarray data analysis for differential expression: a tutorial. *Puerto Rico health sciences journal*, 28(2), 2009.

[54] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[55] Jiho Yoo and Seungjin Choi. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information processing & management*, 46(5):559–570, 2010.

[56] Zhong-Yuan Zhang, Tao Li, and Chris Ding. Non-negative tri-factor tensor decomposition with applications. *Knowledge and information systems*, 34(2):243–265, 2013.

9 Fasi del progetto

Il progetto di ricerca si svilupperà principalmente in 3 fasi. Di seguito viene riportato un elenco in cui si specificano nel dettaglio lo sviluppo e il tempo (dove è possibile quantificarlo) di ogni fase.

1. Studio e analisi del problema tramite la bibliografia già esistente riguardo l'argomento. Si analizzeranno quindi le varie tecniche proposte in letteratura per le riduzioni low-rank di dati strutturati a carattere biomedico. Questa prima fase verrà effettuata durante il primo anno e parte del secondo anno del dottorato di ricerca;

2. Sviluppo di tecniche legate alla metodologia proposta e di strumenti in grado di valutarne la bontà dei risultati;

3. Sperimentazione della tecnica proposta su data set disponibili in letteratura o da eventuali collaborazioni con esperti del campo di applicazione biomedicale.

In tutto il periodo e in particolare nella prima e seconda fase, sono necessari, oltre che un'intensa attività di studio, colloqui sia con il tutor che con esperti di entrambi i domini di applicazione (matematico e biomedico), è inoltre previsto un periodo di minimo tre mesi presso almeno un'altra struttura di ricerca specializzata nella ricerca delle tecniche di decomposizione low-rank e operante anche nell'analisi di dati biomedici.
I risultati miliari che si vogliono raggiungere sono i seguenti:

- Al termine del primo anno un tecnical report/survey inerente la letteratura disponibile che sarà studiata nella fase 1 del progetto;

- Tra il secondo e parte del terzo anno si cercherà di sviluppare una tecnica e/o uno strumento in grado di facilitare la lettura di analisi di riduzione low-rank anche ai non esperti del campo matematico;

- Scrittura della tesi di dottorato e di un articolo che raccolga i risultati ottenuti.

10 Valutazione dei risultati

Verranno effettuate varie sperimentazioni su data set sintetici e/o reali provenienti da eventuali collaborazioni con esperti del campo di applicazione biomedico al fine di verificare la capacità dei metodi utilizzati, di estrarre fattori latenti attesi. Verranno utilizzate rappresentazioni grafiche dei risultati che meglio evidenzino la presenza di eventuali raggruppamenti nei dati, in modo da rendere i risultati più interpretabili anche dagli esperti del settore di applicazione del progetto di ricerca e esterni quindi all'ambito matematico. In particolare verranno applicate le misure di validazione interne ed esterne tipiche degli algoritmi di apprendimento non supervisionato.

11 Eventuali referenti esterni al Dipartimento

Nicolas Gillis, Associate Professor at Department of Mathematics and Operational Research Faculté polytechnique, Université de Mons
Angelina Boccarelli, Ricercatore presso il Dipartimento di Farmacia, Univeristà degli Studi di Bari Aldo Moro
Enrico Blanzieri, Professore Associato presso il Dipartimento di Ingegneria e Scienza dell'Informazione, Università degli Studi di Trento