Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Relazione Finale

Metodi di biclustering per l'identificazione di sorgenti astronomiche diffuse

Relatrice: Prof.ssa Alessandra Rosalba Brazzale
Dipartimento di Scienze Statistiche
Correlatore: Dott. Andrea Sottosanti
Dipartimento di Medicina

Laureando: Antonio Flotta
Matricola N 2020599

Anno Accademico 2022/2023

Indice

1 Il contesto astrofisico: quasar e getti di raggi X **13**
1.1 Il contesto astrofisico . . . 13
1.2 Dal contesto astrofisico a quello statistico . . . 14
1.3 Stato dell'arte e nuova proposta . . . 16
1.3.1 Stato dell'arte: un esempio . . . 17

2 I modelli per biclustering **21**
2.1 Introduzione . . . 21
2.2 Spike and Slab Lasso . . . 22
2.2.1 Base teorica . . . 22
2.2.2 Applicazione al biclustering . . . 24
2.2.3 Scelta del miglior adattamento . . . 27
2.3 SparseBC . . . 28
2.3.1 Il modello di biclustering . . . 28
2.3.2 Algoritmo di stima . . . 29
2.3.3 Scelta di G, R, λ . . . 30
2.4 Sparse Singular Value Decomposition (SSVD) . . . 30
2.4.1 Il modello . . . 31
2.4.2 Algoritmo di stima . . . 33
2.4.3 Scelta del numero di *bicluster* K . . . 34
2.5 Riassumendo . . . 35

3 Studi di simulazione **39**
3.1 Disegno di simulazione . . . 39
3.2 Caso A: nessun getto . . . 40
3.3 Caso B: getto distante dal quasar . . . 41
3.4 Caso C: getto contiguo . . . 42
3.5 Caso D: getto allungato . . . 44
3.6 Confronto tra modelli . . . 45

4 Applicazioni **47**
4.1 Considerazioni iniziali . . . 47
4.1.1 Inizializzazione dell'algoritmo di stima del modello SSLB 47
4.1.2 Contestualizzazione dei metodi . . . 48
4.1.3 Note di interpretazione delle figure . . . 49
4.2 Risultati . . . 51
4.2.1 Quasar senza getto (A1) . . . 51

4.2.2 Due getti forti lontani dal quasar (B4) 57
4.2.3 Quasar con getto debole ed esteso (D1) 63
4.2.4 Quasar con getto medio contiguo (C2) 67
4.3 Riassumendo 73

5 Un esempio con dati reali **75**
5.1 I dati 75
5.2 Applicazione dei metodi 76
5.3 Risultati 77

6 Conclusioni **81**

A Ulteriori risultati **83**
A.1 Quasar variabile senza getto (A2) 83
A.2 Quasar con getto debole lontano (B1) 88
A.3 Quasar con getto forte lontano (B2) 93
A.4 Quasar con due getti deboli lontani (B3) 98
A.5 Quasar con getto debole contiguo (C1) 103
A.6 Quasar con getto forte contiguo (C3) 107
A.7 Quasar con getto medio esteso (D2) 112
A.8 Quasar con getto forte esteso (D3) 117

B Listato in linguaggio R **123**

Bibliografia **131**

Sitografia **133**

Elenco delle figure

1.1 Due esempi di immagini di quasar con getto di raggi X. 14

2.1 Esempio regola del gomito. 35

3.1 Immagini simulate per il contesto di simulazione A: quasar senza getto. 41
3.2 Immagini simulate per il contesto di simulazione B: quasar con getti di raggi X lontani da esso. 42
3.3 Immagini simulate per il contesto di simulazione C: quasar con getto di raggi X contiguo. 43
3.4 Immagini simulate per il contesto di simulazione D: quasar con getto di raggi X esteso. 44

4.1 Esempio delle differenze tra immagine nella rappresentazione originale e immagine nella rappresentazione in negativo. 49
4.2 Immagine ricostruite dai tre metodi e immagine obiettivo, per lo studio di simulazione A1. 50
4.3 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione A1. 52
4.4 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione A1. 53
4.5 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione A1. 54
4.6 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione A1. 55
4.7 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi, per lo studio di simulazione B4. 56
4.8 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione B4. 58
4.9 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione B4. 59
4.10 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione B4. 60
4.11 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione B4. 61

4.12 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi, per lo studio di simulazione D1. 63
4.13 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione D1. 64
4.14 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione D1. 65
4.15 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione D1. 66
4.16 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi, per lo studio di simulazione C2. 67
4.17 Grafico degli autovalori, al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione C2. 69
4.18 Rappresentazione in negativo della composizione degli strati stimati da SSVD per lo studio di simulazione C2. 70
4.19 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione C2. 71
4.20 Rappresentazione in negativo della composizione dei *bicluster*, stimati da SparseBC, per lo studio di simulazione C2. 72

5.1 Immagine della pulsar IGR J11014-6103 nella banda elettromagnetica dei raggi X. 76
5.2 Ingrandimento dell'immagine in Figura 5.1 nella zona di interesse della pulsar. 77
5.3 Immagine originale e immagini ricostruite dai tre metodi, per l'applicazione al caso reale della pulsar. 78
5.4 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per l'applicazione al caso reale della pulsar. 79
5.5 Rappresentazione della composizione degli strati stimati da SSVD, per l'applicazione al caso reale della pulsar. 80

A.1 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre modelli, per lo studio di simulazione A2. 83
A.2 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione A2. 84
A.3 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione A2. 85
A.4 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione A2. 86
A.5 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione A2. 87
A.6 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre modelli, per lo studio di simulazione B1. 88

A.7 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione B1. 89
A.8 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione B1. 90
A.9 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione B1. 91
A.10 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione B1. 92
A.11 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre modelli, per lo studio di simulazione B2. 93
A.12 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione B2. 94
A.13 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione B2. 95
A.14 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione B2. 96
A.15 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione B2. 97
A.16 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi, per lo studio di simulazione B3. 98
A.17 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione B3. 99
A.18 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione B3. 100
A.19 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione B3. 101
A.20 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione B3. 102
A.21 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi per lo studio di simulazione C1. 103
A.22 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione C1. 104
A.23 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione C1. 105
A.24 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione C1. 106
A.25 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi, per lo studio di simulazione C3. 107
A.26 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione C3. 108
A.27 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione C3. 109

A.28 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione C3. 110
A.29 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione C3. 111
A.30 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi, per lo studio di simulazione D2. 112
A.31 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione D2. 113
A.32 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione D2. 114
A.33 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione D2. 115
A.34 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione D2. 116
A.35 Rappresentazione in negativo dell'immagine obiettivo e delle immagini ricostruite dai tre metodi, per lo studio di simulazione D3. 117
A.36 Grafico degli autovalori al variare del numero di strati e corrispettivo ingrandimento, per lo studio di simulazione D3. 118
A.37 Rappresentazione in negativo della composizione degli strati stimati da SSVD, per lo studio di simulazione D3. 119
A.38 Rappresentazione in negativo della composizione dei fattori stimati da SSLB, per lo studio di simulazione D3. 120
A.39 Rappresentazione in negativo della composizione dei *bicluster* stimati da SparseBC, per lo studio di simulazione D3. 121

Elenco delle tabelle

4.1 Indici di bontà di ricostruzione dell'immagine, per lo studio di simulazione A1. 51
4.2 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione B4. 57
4.3 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione D1. 62
4.4 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione C2. 68

A.1 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione A2. 84
A.2 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione B1. 89
A.3 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione B2. 94
A.4 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione B3. 99
A.5 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione C1. 104
A.6 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione C3. 108
A.7 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione D2. 113
A.8 Indici di bontà di ricostruzione dell'immagine per lo studio di simulazione D3. 118

Introduzione

I corpi celesti nel nostro Universo eludono l'umano intuito. Generalmente si tratta di oggetti immensi, come le galassie, oppure estremamente densi, come i quasar, fino ad arrivare ad oggetti per lo più misteriosi come i buchi neri. Nell'ultima decade i dati astrofisici sono aumentati esponenzialmente. Tuttavia, essendo gli oggetti di studio enormemente distanti, e data l'estrema precisione richiesta dagli strumenti astronomici, questa grande mole di dati è permeata da un elevato rumore di fondo che spesso nasconde il segnale. Lo strumento di rilevazione per eccellenza, il telescopio, oltre a richiedere un'elevata precisione nella costruzione delle sue lenti, è facilmente suscettibile a fattori esterni come temperatura e condizioni climatiche. Si pensi, come esempio della fragilità di questi oggetti, che il telescopio spaziale *Hubble*, con uno specchio principale dal diametro di due metri e mezzo ha richiesto un perfezionamento per un difetto dell'ordine di alcuni micron (10^{-6} metri). Si deduce quindi che la statistica è fondamentale in questo contesto.

In questa tesi si porrà l'attenzione sui quasar, nuclei galattici attivi che ospitano al centro di essi un buco nero. Tramite lo studio di questi corpi celesti e le loro emissioni di raggi X, è possibile ricavare informazioni sulla loro formazione ed evoluzione e indirettamente sulle proprietà che caratterizzano un buco nero. Il punto di partenza dell'analisi è un'immagine, suddivisa in una griglia di pixel che descrivono una porzione di essa e dunque dello spazio. In corrispondenza di ciascun pixel viene rilevato dal telescopio un numero di fotoni. Dal punto di vista matematico l'immagine viene trattata come una matrice, dove ciascuna entrata coincide con un pixel e contiene il conteggio dei fotoni rilevati in corrispondenza di quel pixel. Lo scopo finale dell'analisi è identificare zone della matrice, e dunque dell'immagine osservata, che contengono segnale. In particolare, è di interesse comprendere se siano presenti dei getti di raggi X e che forma abbiano, al fine di trarre delle conclusioni sulla natura e sulle proprietà fisiche dell'oggetto. L'ostacolo principale è la netta differenza di luminosità (quantità di fotoni rilevati) tra il quasar e il suo ipotetico getto di raggi X, sorgente diffusa e poco luminosa. Inoltre, data la colossale distanza di questi oggetti celesti dal nostro punto di osservazione, vi sono delle sorgenti luminose di disturbo che si frappongono tra noi e il quasar e compongono il rumore di fondo. Spesso la luminosità del getto è più simile a quella del rumore di fondo piuttosto che alla luminosità del quasar, aspetto che complica ulteriormente l'analisi.

In letteratura il problema viene tradizionalmente affrontato tramite approcci di verifica d'ipotesi. Ci sono metodi che saggiano la presenza di segnale

pixel per pixel e altri, come quello presentato in [11], che si servono di una procedura Bayesiana per verificare se vi sia evidenza a favore di una componente aggiuntiva oltre al quasar nell'immagine. In questa tesi vogliamo invece sperimentare un approccio modellistico per la risoluzione del problema. Inizialmente abbiamo considerato un modello Bayesiano per la stima dei contorni fisici degli oggetti, presentato in [2]. L'idea è stata tuttavia abbandonata perché si tratta di un approccio supervisionato. In particolare, il modello prevede l'utilizzo di un campione di immagini al fine di ottenere la stima dei contorni fisici. Tuttavia, in astrofisica si ha a disposizione un'unica immagine con la quale lavorare. La scelta è infine ricaduta sui metodi per il *biclustering*, comunemente utilizzati nel contesto della genomica per raggruppare in gruppi bidimensionali i soggetti biologici (campioni o righe) e i geni (variabili o colonne). In ambito astrofisico raggruppare simultaneamente righe e colonne della matrice dei dati significa sostanzialmente identificare zone della matrice composte da pixel in corrispondenza dei quali abbiamo registrato un conteggio di fotoni simile. L'auspicio è di individuare la zona dell'immagine dove si colloca il quasar e quella dove si colloca il suo eventuale getto di raggi X. Questa classe di metodi, mai utilizzata prima d'ora in questo contesto, ci ha particolarmente affascinato poiché si tratta di un approccio non supervisionato, che permette di identificare in modo automatico le zone dell'immagine osservata contenenti segnale, senza bisogno di definire un sistema di verifica di ipotesi e dunque una statistica test.

L'elaborato è così suddiviso: nel primo capitolo viene presentato in dettaglio il problema astrofisico, la transizione da esso al problema statistico e lo stato dell'arte. Si procede, nel secondo capitolo, alla presentazione di 3 metodi per il *biclusterng* che saranno applicati agli studi di simulazione descritti nel capitolo terzo; Il quarto capitolo è invece dedicato alla presentazione dei risultati dell'applicazione dei 3 metodi agli studi di simulazione, il quinto ad un caso di studio reale mentre il sesto alle conclusioni.

Capitolo 1

Il contesto astrofisico: quasar e getti di raggi X

1.1 Il contesto astrofisico

I recenti progressi tecnologici, come l'utilizzo di sofisticati telescopi a diverse frequenze, dalle onde radio ai raggi X, e con elevata risoluzione angolare, che permettono di distinguere con maggior precisione sorgenti di luce contigue, hanno permesso l'osservazione di oggetti celesti ad una precisione prima impensabile. Come spesso accade, l'innovazione tecnologica ha spinto la statistica a compiere un salto di qualità, in questo caso nel contesto della modellazione spaziale delle immagini.

L'oggetto celeste che ha ispirato questa tesi è il quasar, un nucleo galattico attivo, lontano miliardi di anni luce, al cui centro è presente un buco nero super massiccio (SMBH) che inghiotte la materia circostante. I quasar sono gli oggetti più luminosi del nostro Universo, addirittura molto più della galassia che li ospita, pur essendo nettamente più piccoli. Attorno al buco nero centrale si crea il cosiddetto disco di accrescimento, composto dalla materia che sta per essere inghiottita dal buco nero stesso. Tuttavia, parte di questa materia riesce a scappare dall'enorme attrazione gravitazionale e viene espulsa sotto forma di getti di elettroni che si estendono a distanze incredibili, anche di milioni di anni luce. L'interazione di questi getti con le radiazioni del disco producono emissioni di raggi X. Queste ultime, se paragonate al quasar, sono molto più deboli. Inoltre, mentre il quasar è una sorgente puntiforme, i getti di raggi X sono una sorgente diffusa, ovvero di forma estesa, dove per sorgente, in abito astrofisico, si intende un qualsiasi corpo celeste che emette radiazioni. I quasar sono molto variabili: ad osservazioni in momenti diversi potrebbero presentare o non presentare un getto di raggi X o presentarne di diverse forme. Dato che osservare oggetti molto distanti corrisponde ad osservare indietro nel tempo, lo studio di quasar e delle loro emissioni di raggi X, che influenzano l'ambiente circostante con la loro energia, fornisce importanti informazioni sulla formazione delle strutture del giovane Universo e sull'attività dei buchi neri che li ospitano. È di interesse dunque, data l'immagine di un quasar, determinare se vi sia un getto di raggi X e che forma abbia.

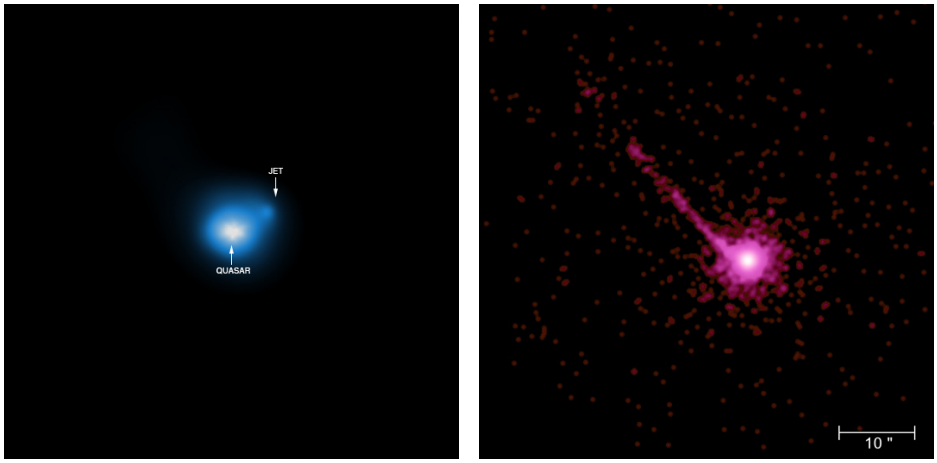

Figura 1.1: *A sinistra: immagine a raggi X del quasar GB 1428+4217, distante dalla terra* 12.4 *miliardi di anni luce con un getto di raggi X che si estende a circa* 230000 *anni luce dal quasar. A destra: immagine a raggi X del quasar PKS 1127-145, oggetto distante* 10 *miliardi di anni luce dalla terra. Presenta un getto di raggi X che si estende fino ad almeno un milione di anni luce dal quasar. Fonte: Chandra X-Ray Observatory.*

Il punto di partenza delle analisi che verranno svolte in questa tesi è dunque un'immagine come quelle in Figura 1.1. Le maggiori problematiche dell'identificazione di getti di raggi X sono:

- la presenza di elementi di disturbo all'interno dell'immagine. Quasi mai si ottengono immagini pulite come quella in Figura 1.1 che raffigura il quasar GB 1428+4217 (a sinistra). Più comune è invece avere a che fare con immagini permeate da rumore di fondo, ovvero punti luminosi che non sono attribuibili né al quasar né al getto di raggi X, come per il quasar PKS 1127-145 (Figura 1.1, a destra). Questo fenomeno si verifica per due principali motivi. Innanzitutto, gli strumenti utilizzati per la rilevazione sono difettosi per meccanismi intrinsechi alle leggi dell'ottica. In secondo luogo, gli oggetti che vengono osservati sono molto distanti ed è dunque semplice che altre sorgenti, più o meno luminose, si frappongano tra il telescopio e l'oggetto celeste stesso.

- la forma dei getti di raggi X stessi, diffusa ed irregolare, li fa confondere con il rumore di fondo.

- nel caso in cui i getti siano relativamente vicini al quasar che li ha generati, potrebbe essere difficile distinguere le due sorgenti l'una dall'altra.

1.2 Dal contesto astrofisico a quello statistico

L'immagine di un quasar e del suo eventuale getto di raggi X non è direttamente utilizzabile come insieme di dati per effettuare l'analisi. Per meglio comprendere come si passa dall'immagine, alla matrice dei dati, è necessario

definire un sistema di coordinate. L'immagine, analoga a quelle in Figura 1.1, rappresenta una porzione di spazio, o, in altre parole una porzione della cupola celeste. Questo spazio, che nella realtà è tridimensionale, una volta fotografato ci appare bidimensionale, poiché non presenta profondità, e dunque necessita di due dimensioni per essere descritto. In astronomia sono molteplici i sistemi di coordinate che possono essere utilizzati. In questa tesi verrà utilizzato il sistema equatoriale, di seguito descritto. La declinazione, analogo astronomico della latitudine, rappresenta l'angolo tra l'astro e l'equatore celeste, che coincide con quello terrestre. L'ascensione retta invece, controparte celeste della longitudine, rappresenta l'angolo tra l'astro ed un meridiano di riferimento. Declinazione e ascensione retta sono dunque le coordinate che definiscono il sistema equatoriale e vengono utilizzate per descrivere la porzione di sfera celeste illustrata dall'immagine. Ad esempio, l'immagine in Figura 1.1 (a destra) copre una porzione di sfera celeste di 60×60 arco secondi, ovvero si estende per 60 arco secondi in ascensione retta (dimensione orizzontale) e altrettanti in declinazione (dimensione verticale).

Per essere descritta in termini matriciali, l'immagine deve essere *pixellizzata*, ovvero frammentata in porzioni di spazio. Ciascun elemento della griglia con la quale viene partizionata l'immagine si chiama pixel. Di solito, la grandezza dei pixel varia da applicazione ad applicazione e viene descritta in termini di arco secondi di angolo. Per esempio, l'immagine in Figura 1.1 è stata divisa in una griglia di 430×430 pixel, di grandezza 0.14×0.14 arco secondi circa. In corrispondenza di ciascun elemento della griglia, e dunque di ciascun pixel, vengono conteggiati il numero di fotoni rilevati dallo strumento. Viene considerato il numero di fotoni poiché questi sono particelle senza massa che compongono le onde elettromagnetiche. Nel nostro caso, lo strumento di rilevazione è il *Chandra X-Ray Telescope*, un telescopio a raggi X in orbita attorno alla terra che registra i fotoni corrispondenti alla banda elettromagnetica dei raggi X, di interesse nella nostra analisi. Il risultato di questa fase iniziale di creazione dell'insieme di dati è dunque una matrice le cui righe e colonne indicizzano la griglia di pixel con la quale è stata partizionata l'immagine. Ciascuna entrata della matrice contiene il conteggio del numero di fotoni, nella banda elettromagnetica dei raggi X, rilevato nella porzione di spazio corrispondente a quel pixel. Più elevato è il numero di fotoni rilevato in corrispondenza di un pixel, più intensa è la sorgente che li ha generati e dunque più intenso è il colore dell'immagine in quella porzione.

Nella maggior parte dei casi, la matrice dei dati sarà quadrata, tuttavia nel proseguo della tesi si considererà di dimensione $n \times p$ per essere più generali possibile. Verrà indicata con $Y = \{y_{ij}\}_{i=1,j=1}^{n,p}$, dove ciascun y_{ij} è il conteggio del numero di fotoni osservato nella porzione di spazio corrispondente al pixel di posizione (i, j).

Quasar ed eventuale getto compongono solamente una piccola parte dell'immagine osservata. Quindi, per quanto detto in precedenza sulla corrispondenza tra immagine e matrice dei dati, ci aspettiamo che quest'ultima presenti un'elevata quantità di entrate pari, o molto prossime, a 0. Dal punto di vista statistico questo si traduce con la presenza di sparsità nei dati. I metodi per

l'analisi dunque dovranno tenere conto di questo aspetto, come si vedrà nel Capitolo 2. Inoltre, per identificare il segnale presente nell'immagine e quindi il quasar ed un eventuale getto i metodi dovranno essere molto sensibili. Infatti, se il quasar è facilmente distinguibile, per luminosità e dunque numero di conteggi per pixel, lo stesso non vale per il getto di raggi X, che è paragonabile al rumore di fondo piuttosto che al quasar, date le sue forme diffuse e varie e la sua luminosità tenue. Si deduce che la presenza di una sorgente molto luminosa nell'immagine non aiuta l'identificazione del getto poiché rispetto ad esso risalta e ne potrebbe quindi nascondere il segnale.

1.3 Stato dell'arte e nuova proposta

Il problema di identificazione del segnale all'interno di immagini viene affrontato in numerosi ambiti della letteratura scientifica, dall'ambito medico, per analizzare le risonanze magnetiche, quello di sicurezza, per analizzare i *frame* delle telecamere, fino a quello astrostatistico. Un approccio di verifica di ipotesi viene comunemente utilizzato per determinare se in una determinata zona dell' immagine vi sia segnale. In particolare, vi sono metodi che, per ogni pixel di cui è composta la matrice dei dati costruiscono un test di verifica di ipotesi come il seguente,

$$\begin{aligned} H_0 &: \text{ Non c'è segnale nel pixel} \\ H_1 &: \text{ C'è segnale nel pixel.} \end{aligned}$$

Il *p-value* risultante viene poi aggiustato per la molteplicità dei test. Un esempio, in ambito medico, può essere trovato in [4], che utilizza un approccio basato sui modelli lineari generalizzati.

Una alternativa è rappresentata da approcci, sempre di verifica di ipotesi, dove si saggia la presenza di segnale in particolari zone della matrice dei dati. Data una zona P dell'immagine dove si è interessati a capire se c'è segnale o meno, una struttura tipica del test, in questo contesto, è

$$\begin{aligned} H_0 &: \text{ Non c'è segnale nella zona } P \\ H_1 &: \text{ C'è segnale nella zona } P. \end{aligned} \tag{1.1}$$

Come esempio, proprio nell'ambito astrofisico di interesse, si cita l'articolo [11], punto di partenza di questa tesi. L'idea di base è ottenere una misura di significatività che non sia pixel-specifica bensì riferita ad un'intera area dell'immagine e di conseguenza della matrice dei dati. Il metodo si inserisce in un contesto Bayesiano, ma utilizza, per saggiare il test di ipotesi in equazione (1.1), una vera e propria statistica test, nel senso di una funzione dei dati la cui distribuzione è nota e non dipende dai parametri. Nel seguito verranno presentati i passi principali del metodo, al fine di evidenziare le differenze con la proposta innovativa di questa tesi, ovvero un approccio modellistico basato su metodi per il *biclustering*.

1.3.1 Stato dell'arte: un esempio

In [11] viene fatta una distinzione fondamentale tra modello base e componente aggiuntiva. Il primo è presente per rappresentare il quasar e il rumore di fondo, mentre la seconda dovrebbe catturare la parte di segnale che va oltre il modello di base ovvero quello associato all'eventuale getto di raggi X. Ci si riferirà al modello completo intendendo l'unione di queste due componenti. Il sistema di ipotesi in equazione (1.1) deve essere dunque modificato come segue:

$$\begin{aligned} H_0 &: \text{La componente di base spiega la totalità dell'immagine osservata} \\ H_1 &: \text{C'è una componente aggiuntiva che va oltre il modello di base.} \end{aligned} \tag{1.2}$$

In questo caso non si lavora in termini matriciali, bensì vettoriali. Si immagini dunque di allineare in un unico vettore tutte le righe della matrice dei dati Y in modo da ottenere un vettore di lunghezza $M = np$ di conteggi osservati $\boldsymbol{y}_{oss} = (y_1, ..., y_M) = \{y_m\}_{m=1}^{M}$. Si ipotizza che i valori dei conteggi siano stati generati indipendentemente da una distribuzione di Poisson,

$$y_m \sim Poisson\left(\sum_{l=1}^{M} P_{ml} A_l (\mu_{0l} + \mu_{1l})\right), \tag{1.3}$$

dove P_{ml} è la *point spread function* (PSF) dello strumento utilizzato, il telescopio, ovvero la probabilità che un fotone rilevato nella porzione di spazio corrispondente al pixel di posizione m provenga invece dalla porzione di spazio corrispondente al pixel l, $A = (A_1, ..., A_M)$ è l'efficienza della rilevazione con A_l probabilità che un fotone proveniente dal pixel l venga effettivamente rilevato. Inoltre, $\mu_0 = (\mu_{01}, ..., \mu_{0M})$ e $\mu_1 = (\mu_{11}, ..., \mu_{1M})$ sono rispettivamente l'intensità della componente di base e della componente aggiuntiva, ovvero il conteggio di fotoni atteso dalle due componenti. Si noti che P_{ml} e A sono componenti note, caratteristiche dello strumento di rilevazione. Il parametro μ_1 è di interesse in questa analisi poiché descrive l'intensità della componente aggiuntiva, parte del modello presente per catturare il segnale proveniente dall'eventuale getto di raggi X.

Si consideri la seguente parametrizzazione. Per $k = 0, 1$ si ha

$$\mu_k = \tau_k \Lambda_k, \qquad \tau_k = \sum_{m=1}^{M} \mu_{km}, \qquad \Lambda_k = (\Lambda_{k1}, ..., \Lambda_{kM}) = \frac{\mu_k}{\tau_k}.$$

Si nota che τ_k e Λ_k possono essere interpretati rispettivamente come l'intensità totale della sorgente k e la proporzione dell'intensità totale della sorgente k in ciascun pixel m. Si ha dunque come parametro del modello $\theta = (\theta_0, \theta_1)$ con $\theta_0 = (\tau_0, \nu_0)$ e $\theta_1 = (\tau_1, \Lambda_1)$, dove ν_0 è una riparametrizzazione di Λ_0 e può essere considerato noto nel proseguo. Come menzionato in precedenza, l'inferenza viene condotta in un contesto Bayesiano. Al fine di ottenere la distribuzione a posteriori dei parametri, e dunque una loro stima, sono necessarie la funzione di verosimiglianza, associata al modello (1.3) e la distribuzione a priori dei parametri. Tralasciando la scelta delle distribuzioni a priori, è sufficiente

notare che un'approssimazione della distribuzione a posteriori dei parametri viene ottenuta tramite un algoritmo di Markov Chain Monte Carlo (MCMC) che sfrutta le catene di Markov per simulare valori dalla posteriori.

A questo punto dell'analisi è necessario definire una statistica test, e dunque una funzione dei dati la cui distribuzione è nota e non dipende dai parametri, per saggiare il sistema di ipotesi in equazione (1.2). Vorremmo che a valori grandi della statistica test $T(y_{obs})$ corrisponda evidenza a favore della presenza di componente aggiuntiva (ipotesi alternativa) e, viceversa, a valori piccoli, evidenza a favore del modello di base (ipotesi nulla). A questo scopo viene definita la seguente funzione dei parametri:

$$\xi = \frac{\tau_1}{(\tau_1 + \tau_0)}. \tag{1.4}$$

Il parametro ξ rappresenta la proporzione di intensità totale dovuta alla componente aggiuntiva. Dato che siamo in un contesto Bayesiano, la distribuzione a posteriori di ξ è disponibile e può essere usata nella formulazione della statistica test. Data una soglia c, si definisce

$$T_c(y_{obs}) = Pr(\xi \geq c \ |y_{obs}), \tag{1.5}$$

dove la probabilità è calcolata rispetto alla posteriori $\pi(\theta \ |y_{obs})$ sotto l'ipotesi alternativa. L'interpretazione di questa quantità è immediata: se l'ipotesi alternativa è vera ci si aspetta che valori alti di ξ siano più verosimili rispetto a valori bassi e quindi che la probabilità a posteriori che ξ ecceda un'opportuna soglia c sia alta. Spesso si è interessati a verificare se c'è evidenza di una componente aggiuntiva in regioni particolari dell'immagine, a questo scopo si può riscrivere la statistica test nel seguente modo:

$$T_{R,c}(y_{obs}) = Pr(\xi_R \geq c \ |y_{obs}),$$

con

$$\xi_R = \frac{\sum_{j \in R} \tau_1 \Lambda_{1j}}{\sum_{j \in R} (\tau_1 \Lambda_{1j} + \tau_0 \Lambda_{0j})},$$

dove R definisce l'insieme di pixel da cui la regione di interesse è formata.

Al fine di stimare il valore osservato della statistica test viene utilizzato un algoritmo MCMC per ottenere un campione simulato dalla distribuzione a posteriori di ξ. In particolare nel pacchetto *LIRA* proposto in [11] viene implementato un algoritmo basato sul Gibbs sampling che porta ad ottenere un campione di lunghezza S, $\theta_{obs}^{(1)}, ..., \theta_{obs}^{(S)}$, dalla distribuzione a posteriori completa $\pi(\theta \ |y_{obs})$. Tramite una semplice trasformazione $\xi_{obs}^{(s)} = h(\theta_{obs}^{(s)})$ è possibile ottenere i corrispettivi valori simulati a posteriori per ξ, dove $h(\cdot)$ in questo caso è la relazione in equazione (1.4). Per calcolare la statistica test è sufficiente prendere il corrispettivo empirico della quantità in equazione (1.5), ovvero

$$\hat{T}_c(y_{obs}) = \frac{1}{S} \sum_{s=1}^{S} I(\xi_{obs}^{(s)} \geq c),$$

Algoritmo 1 Stima del limite superiore del *p-value* in ([11])

1: **for** $j = 1, ..., J$:
2: Simulare $y_0^{(j)} \sim \mathcal{L}_0(y)$.
3: Adattare il modello completo a $y_0^{(j)}$ usando il pacchetto `LIRA` per ottenere S valori simulati $\xi^{(j,s)}, ..., \xi^{(j,S)}$ da $\pi(\xi|y_0^{(j)})$.
4: **end for**
5: Stimare c, il quantile $(1-\phi)$ di $g(\xi)$, con $\hat{c}$: il valore $(SJ\phi)$-esimo più grande tra tutti gli $\xi^{(j,s)}$.
7: Calcolare la statistica test tramite l'equazione (1.5) con $\hat{c}$ al posto di c.
8: Stimare il limite superiore del *p-value* tramite $\hat{u} = \phi / \hat{T}_{\hat{c}}(y_{obs})$.

dove $I(\cdot)$ rappresenta la funzione indicatrice che vale 0 se l'espressione logica in argomento è falsa e 1 altrimenti.

Una delle maggiori innovazioni che questo articolo ha portato al panorama astrostatistico è una strategia di calcolo del *p-value* alternativa. Infatti, non viene proposto il metodo diretto che, utilizzando un bootstrap parametrico prevede la simulazione sotto l'ipotesi nulla di J immagini, l'adattamento del modello completo ad esse, il successivo calcolo della statistica test associata e infine il calcolo della stima Monte Carlo del *p-value*. Questa procedura sarebbe troppo onerosa dal punto di vista computazionale, poiché J dovrebbe essere davvero elevato per ottenere livelli di significatività elevati. Viene invece proposto di stimare un limite superiore del *p-value*. Si dimostra (Appendice in [11]) che:

$$p \leq \frac{\phi}{T_c(y_{obs})} = u, \tag{1.6}$$

dove $\phi = Pr(\xi \geq c)$ sotto la distribuzione

$$g(\xi) = E_0\left[\pi(\xi|y_0)\right] = \sum_{y_0} \pi(\xi|y_0)\mathcal{L}_0(y_0),$$

dove $\pi(\xi|y_0)$ è la distribuzione a posteriori di ξ, $E_0[\cdot]$ indica che il valore atteso della distribuzione a posteriori sotto l'ipotesi alternativa, è calcolato rispetto alla distribuzione a posteriori sotto l'ipotesi nulla, dove y_0 è un'immagine simulata sotto l'ipotesi nulla tramite bootstrap parametrico e $\mathcal{L}_0$ è la verosimiglianza del modello sotto l'ipotesi nulla. Si noti che $1-\phi$ può essere stimato con il quantile empirico di $g(\xi)$. Il limite superiore del *p-value* in equazione (1.6) può essere stimato seguendo l'Algoritmo 1.

L'approccio appena descritto ha numerosi punti di forza. Innanzitutto aggira il problema di specificazione di una struttura parametrica per la componente aggiuntiva, che potendo assumere varie ed irregolari forme, non si presta ad essere parametrizzata. In altre parole, un solo modello che cercasse di parametrizzare la struttura aggiuntiva del getto di raggi X non riuscirebbe ad essere abbastanza generale da comprendere tutti i modi in cui questi getti si presentano. Utilizzando invece un modello Bayesiano, flessibile ad allontanamenti dalla componente di base, si considerano tutte le diverse configurazioni

che i getti di raggi X potrebbero avere. In secondo luogo, il metodo, invece che quantificare l'incertezza dei risultati di stima, comunque ottenibile esplorando la posteriori dei parametri ottenuta tramite Gibbs sampling, si preoccupa di fare identificazione di componenti aggiuntive nel loro insieme e non pixel per pixel. Con questa formulazione si guadagna in flessibilità e semplicità di calcolo. È infatti sufficiente stimare un unico *p-value* e non un numero pari a quello dei pixel. Infine, il limite superiore sul *p-value* permette di abbattere i costi computazionali se ce n'è il bisogno, senza perdere in significatività, soprattutto per quelle situazioni in cui i getti sono particolarmente evidenti. Nel caso in cui si avesse disponibilità computazionali illimitate è sempre preferibile utilizzare il metodo di stima diretta del *p-value* poiché la strategia del limite superiore potrebbe risultare troppo conservativa, soprattutto in presenza di getti di raggi X deboli. Infine, il modello è in grado includere elementi di disturbo noti come per esempio l'efficacia di rilevazione del telescopio e la *Point Spread Function*.

Sia in [11] che in un articolo gemello ([7]), il modello presentato viene utilizzato per esaminare immagini di quasar, tra le quali alcune provenienti dal *Chandra X-Ray telescope*. Gli autori, molti dei quali autori anche in [11], utilizzano immagini radio degli oggetti per identificare le regioni di interesse sulle quali verificare se la componente di base è sufficiente per descrivere i dati o, al contrario, è presente una componente aggiuntiva. Questa strategia semplifica il problema perché restringendo il campo dove cercare segnale è più facile identificarlo, a parità di forza del segnale. In altre parole, se la regione di interesse è molto circoscritta, il segnale risalta di più che non nell'immagine completa che contiene più rumore di fondo. Inoltre, non è detto che esistano immagini radio degli oggetti a cui si è interessati e quindi questa strada potrebbe non essere sempre percorribile. Infine, non è detto che ad emissioni radio corrispondano emissioni di raggi X. I metodi di *biclustering*, che verranno presentati nel Capitolo 2, sono stati applicati in questa tesi al contesto astrofisico di identificazione delle sorgenti diffuse per la necessità di superare questo vincolo. Infatti, raggruppando le righe e le colonne della matrice dei dati in gruppi bidimensionali, dovrebbero essere in grado di identificare automaticamente le regioni dello spazio significativamente diverse dal rumore di fondo, rendendo le osservazioni radio superflue.

Capitolo 2

I modelli per biclustering

2.1 Introduzione

In questo capitolo verranno presentati 3 diversi metodi di *biclustering*: *Spike and Slab Lasso Biclustering* (SSLB, [8]), un modello Bayesiano che utilizza una penalizzazione di tipo *spike and slab* per ottenere una scomposizione in fattori latenti della matrice dei dati osservati; *SparseBC* ([12]), che utilizza un approccio di penalizzazione della verosimiglianza per identificare i *bicluster*, e *Sparse Singular Value Decomposition* (SSVD, [6]), ovvero un metodo che utilizza un adattamento della scomposizione a valori singolari al caso di dati con elevata sparsità, per scomporre la matrice dei dati in strati. Lo scopo della nostra analisi è identificare un getto di raggi X, che è una sorgente diffusa e debole, dal quasar, che è una sorgente luminosa e puntiforme, in presenza di rumore di fondo. Ricordando che la nostra matrice dei dati è la rappresentazione matematica dell'immagine osservata, si può tradurre lo scopo dell'analisi in una ricerca di zone della matrice dei dati, composte da un insieme di celle della matrice, significativamente diverse dal resto della matrice o immagine.

Una classica procedura di *clustering* non è dunque appropriata, poiché queste tecniche partizionano le righe o le colonne in gruppi, chiamati *cluster*, composti da elementi simili tra loro. Noi, invece, siamo interessati a partizionare righe e colonne simultaneamente poiché entrambe sono di interesse scientifico, per questo vengono utilizzate procedure di *biclustering* in questa tesi. Esse, infatti, identificano insiemi di entrate della matrice che sono simili tra loro, chiamati *bicluster*, che coincide con l'identificazione di zone dell'immagine significativamente diverse dal resto e dal rumore di fondo. Tipicamente, solamente un piccolo sottoinsieme di colonne è responsabile della differenza tra gruppi di righe, tuttavia, in una procedura di *clustering* delle righe, vengono utilizzate tutte le colonne. Al contrario, il *biclustering*, raggruppando simultaneamente entrambe le dimensioni della matrice dei dati, permette di escludere quelle colonne che non sono importanti nell'identificazione dei vari gruppi di righe. In questo modo possono essere identificati gruppi che altrimenti non sarebbe possibile ottenere dal *clustering*. Si noti che il medesimo ragionamento può essere fatto per il caso di *clustering* delle colonne. Inoltre, mentre nel *clustering* una riga può appartenere ad uno ed un solo gruppo e non può non

appartenere a nessun gruppo, nei metodi per il *biclustering* questo è possibile. La caratteristica di non dover raggruppare obbligatoriamente tutte le righe e le colonne è la motivazione che ci ha fatto scegliere il *biclustering* invece del *co-clustering*, un'altra procedura di raggruppamento simultaneo di righe e colonne che, tuttavia, prevede che la totalità di esse venga utilizzata per il raggruppamento. Dato che nel contesto astrofisico, ciascuna entrata della matrice dei dati è un pixel, in corrispondenza del quale abbiamo a disposizione il conteggio dei fotoni rilevati in quella porzione di spazio, i *bicluster* risultanti dall'analisi corrisponderanno ad un insieme di pixel, in corrispondenza dei quali è stato rilevato un conteggio di fotoni simile. Sperabilmente, uno di questi bicluster corrisponderà al quasar e altri corrisponderanno agli eventuali getti di raggi X. Essendo la matrice dei dati tipicamente sparsa, verranno implementati metodi per il *biclustering* capaci di considerare la sparsità che caratterizza i dati astrofisici.

La differenza con i metodi descritti nel Capitolo 1 sta nell'approccio intrapreso. Mentre in letteratura viene utilizzato un sistema di ipotesi per saggiare la presenza di segnale nell'immagine (paragrafo 1.3), i metodi per il *biclustering* includono questo passo direttamente nell'aspetto di modellazione, senza bisogno di identificare prima dell'analisi aree dove cercare il segnale, e senza la necessità di definire un sistema di ipotesi.

2.2 Spike and Slab Lasso

2.2.1 Base teorica

In [10] viene presentato lo *Spike and Slab Lasso* come un tentativo di applicare, in ambito frequentista, una tecnica generalmente utilizzata in ambito Bayesiano, per contemporaneamente stimare e selezionare i parametri di un modello lineare penalizzato.

Sia

$$\boldsymbol{z} = X\boldsymbol{\eta} + \boldsymbol{\epsilon}$$

il classico modello lineare con $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I_n)$, $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ matrice di disegno, $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$ vettore dei parametri e $\boldsymbol{z} = (z_1, ..., z_n)$ vettore risposta n-dimensionale. In un contesto dove i dati sono sparsi, è utile considerare una penalizzazione della verosimiglianza. Si può scrivere, in termini generali

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta} \in \mathbb{R}}{\arg\max} \left(-\frac{1}{2} ||\boldsymbol{z} - X\boldsymbol{\eta}||^2 + pen_\lambda(\boldsymbol{\eta}) \right), \tag{2.1}$$

dove per esempio per $pen_\lambda(\boldsymbol{\eta}) = \sum_{j=1}^{p} -\lambda|\eta_j|$ si ha la penalità del lasso ([13]). Qualsiasi problema di questo tipo può essere visto in un'ottica Bayesiana semplicemente definendo $pen_\lambda(\boldsymbol{\eta}) = \log(\pi(\boldsymbol{\eta}|\lambda))$, dove $\pi(\boldsymbol{\eta}|\lambda)$ indica una qualche distribuzione a priori di $\boldsymbol{\eta}$. Così facendo, $\hat{\boldsymbol{\eta}}$ in (2.1) è il valore di $\boldsymbol{\eta}$ che massimizza la distribuzione a posteriori del modello, ovvero la moda. Definendo una distribuzione a priori del tipo *spike and slab* sugli elementi di $\boldsymbol{\eta}$, è possibile

sfruttarne l'effetto di schiacciamento verso zero, ideale in presenza di sparsità. La generica forma di una priori *spike and slab* è la seguente:

$$\pi(\boldsymbol{\eta}|\lambda) = \prod_{j=1}^{p} \left[\gamma_j \psi_1(\eta_j) + (1-\gamma_j)\psi_0(\eta_j)\right],$$

con $\psi_1(\eta_j)$, parte di *slab*, che controlla gli effetti grandi e $\psi_0(\eta_j)$, parte di *spike*, che modella invece gli effetti piccoli ed è infatti concentrata in 0. Vengono proposte come priori delle Laplace di parametri λ_1 e λ_0 rispettivamente,

$$\psi_1(\eta_j) = \frac{\lambda_1}{2}\exp(-\lambda_1|\eta_j|) \qquad \psi_0(\eta_j) = \frac{\lambda_0}{2}\exp(-\lambda_0|\eta_j|),$$

dove λ_0 è un valore grande mentre λ_1 è un valore piccolo. A $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)^T$ viene associata una distribuzione della forma

$$\pi(\boldsymbol{\gamma}|\theta) = \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j},$$

dove $\theta = P(\gamma_j = 1|\theta)$ è la frazione attesa a priori di η_j diversi da zero. Condizionatamente a θ si ha

$$\pi(\boldsymbol{\eta}|\theta) = \prod_{j=1}^{p} \left[\theta\psi_1(\eta_j) + (1-\theta)\psi_0(\eta_j)\right],$$

ovvero una mistura tra una parte di *spike* che schiaccia gli effetti piccoli verso zero e una parte di *slab* che evita che anche valori grandi dei parametri vengano schiacciati. Utilizzando questa a priori, trovare la moda a posteriori porta sia alla stima dei parametri che alla selezione degli stessi. Se si tratta anche θ come una variabile casuale, imponendo su di esso una distribuzione a priori, si rende la componente di penalizzazione non separabile. In altre parole, i η_j, con $j = 1, ..., p$, non sono più indipendenti e si possono sfruttare le informazioni provenienti da $\boldsymbol{\eta}_{/j}$ ($\boldsymbol{\eta}$, senza la j-esima componente) per stimare la j-esima componente di $\boldsymbol{\eta}$. Questo permette al livello di sparsità θ di adattarsi a seconda della coordinata j che si sta stimando. È possibile notare che la stima dei parametri nel caso di penalizzazione non separabile, viene ottenuta, similmente a quanto accade nel lasso, tramite un algoritmo che itera sulle coordinate del vettore dei parametri. A differenza di ciò che accade nella stima di tipo lasso, il livello di sparsità è specifico per ogni coordinata, a causa della distribuzione a priori che viene imposta su θ. Inoltre, la stima di $\boldsymbol{\eta}$ viene ottenuta per una serie di N λ_0, ($\lambda_0^1 < ... < \lambda_0^N$), con l'effetto di eliminare gradualmente i coefficienti non significativi che risultano diversi da zero solo a causa del rumore di fondo, e contemporaneamente mantenere quelli che risultano essere diversi da zero. Vengono dunque ottenute N stime di $\boldsymbol{\eta}$, ($\hat{\boldsymbol{\eta}}^1, ..., \hat{\boldsymbol{\eta}}^N$), e quando un ulteriore incremento di λ_0^N non influenza la stima di $\boldsymbol{\eta}$, si riporta $\hat{\boldsymbol{\eta}}^N$ come stima. Per ulteriore approfondimento riguardo l'algoritmo di stima si fa riferimento a [10].

2.2.2 Applicazione al biclustering

Ora che il concetto di *Spike and Slab Lasso* è stato introdotto, viene presentato il metodo di *biclustering* che si basa su di esso.

Nel contesto astrofisico ci si attende che pochi pixel compongano il segnale presente nei dati mentre la maggior parte siano attribuibili al rumore di fondo. Per questo motivo, metodi di raggruppamento classici, che utilizzano tutte le osservazioni e tutte le variabili, o nel nostro caso, tutte le righe e tutte le colonne della matrice dei dati, potrebbero fallire nella definizione dei gruppi. Il vantaggio dei metodi di *biclustering* con penalizzazione per la sparsità è che non tutti gli elementi devono appartenere ad un gruppo, dando così voce solamente a quelli maggiormente influenti. Lo *Spike and slab lasso biclustering* si serve di una scomposizione in fattori latenti per approssimare la matrice dei dati. Vengono stimati i fattori ed i coefficienti della scomposizione tramite un approccio Bayesiano che prevede l'utilizzo di priori di tipo *spike and slab*. Ciascun fattore identifica un *bicluster*, in quanto essi sono espressione della relazione tra un insieme di righe ed un insieme di colonne. Idealmente vorremmo che il primo fattore identifichi il quasar e i successivi gli eventuali getti di raggi X presenti nell'immagine.

Si definisce la matrice dei dati,

$$Y = \{y_{ij}\}_{i,j}^{n,p} \in \mathbb{R}^{n\times p},$$

dove nel caso in esame, y_{ij} è il conteggio di fotoni osservato nel pixel di posizione (i, j). Sia

$$Y = \sum_{k=1}^{K} \boldsymbol{x}_F^k (\boldsymbol{\beta}^k)^T + \mathbf{E}, \tag{2.2}$$

dove $\mathbf{X}_F = [\boldsymbol{x}_F^1, ..., \boldsymbol{x}_F^K] \in \mathbb{R}^{n\times K}$ è una matrice di fattori, con $\boldsymbol{x}_F^k$ k-esima colonna di $\mathbf{X}_F$, $\mathbf{B} = [\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^K] \in \mathbb{R}^{p\times K}$ è la matrice dei coefficienti, con $\boldsymbol{\beta}^k$ k-esima colonna di $\mathbf{B}$ e $\mathbf{E} = [\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n]^T \in \mathbb{R}^{n\times p}$ è una matrice di errore Gaussiana con riga i-esima $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(0, \Sigma)$ con $\Sigma = \text{diag}\{\sigma_j^2\}_{j=1}^p$. Dal punto di vista statistico non è ideale imporre una distribuzione per valori continui a dati di conteggio, ma si vedrà, analizzando i risultati degli studi di simulazione, che il modello funziona adeguatamente anche in questo contesto, come presentato anche dagli autori in [8]. L'utilizzo di un modello di tipo fattoriale è giustificato dal fatto che in tante applicazioni vengono riscontrati effetti moltiplicativi tra le righe e le colonne della matrice dei dati.

Si vuole fornire una struttura del tipo *spike and slab*, vista nel paragrafo 2.2.1, alle colonne di $\mathbf{X}_F$ e di $\mathbf{B}$ al fine di penalizzare le stime, tenendo così conto della sparsità dei dati. Si lavora in un contesto Bayesiano dove tutti i parametri vengono trattati come variabili aleatorie e dunque possiedono una distribuzione a priori, che va definita. Neanche K, il numero di fattori, o *bicluster*, viene deciso a monte dell'analisi, infatti si utilizza una a priori *Indian Buffet Process* (IBP) sulle grandezze dei *bicluster* (elementi non zero delle colonne di $\mathbf{X}_F$ e $\mathbf{B}$). L'*Indian Buffet Process* è un processo stocastico, la cui distribuzione limite viene utilizzata come priori per i fattori latenti in modelli a fattori

latenti sparsi, esattamente il contesto in cui ci troviamo ora. In particolare, si basa sulla scomposizione della matrice dei fattori latenti in due matrici, una binaria che descrive quali entrate della matrice dei fattori sono diverse da zero e una che contiene il valore vero e proprio dei fattori. La priori *Indian Buffet Process* si riferisce agli elementi della matrice binaria, i quali regolano la dimensionalità del problema, nel senso che gli elementi non zero di essa definiscono l'effettiva dimensionalità del modello. Dato che non si conosce il numero di fattori, l'idea alla base di questo modello è descrivere la matrice binaria attraverso un numero potenzialmente infinito di fattori $K \to \infty$. Per ulteriori dettagli si rimanda a [5]. Uno dei grandi vantaggi di SSLB è proprio la sua capacità di stimare automaticamente il numero di *bicluster* a partire dai dati. Le stime dei parametri $\mathbf{X}_F$ e $\mathbf{B}$, ovvero la moda delle distribuzioni a posteriori ad essi riferite, vengono ottenute tramite un algoritmo EM. Si rende dunque necessaria un'inizializzazione delle matrici $\mathbf{X}_F$ e $\mathbf{B}$ e per farlo si deve a sua volta inizializzare il numero di *bicluster* K. Si suggerisce di adottare un valore di partenza che sia una sovrastima del valore che ci si aspetta, per esempio $K^* = 50$. Nel caso in cui si trovassero 50 *bicluster* alla fine dell'analisi, sarà necessario aumentare ulteriormente K^*. Il numero di *bicluster* stimato sarà ottenuto eliminando le colonne di $\mathbf{X}_F$ e $\mathbf{B}$ le cui componenti a posteriori vengono stimate tutte uguali a zero.

Per $\mathbf{B}$ si ha la seguente struttura: a ciascuna colonna $\boldsymbol{\beta}^k = \{\beta_{jk}\}_{j=1}^p$ si impone una priori di tipo *spike and slab*, i.e., ciascun elemento della colonna k di $\mathbf{B}$, ovvero ciascun β_{jk} è generato a priori da una distribuzione di Laplace, o di tipo "spike", parametrizzata da un λ_0 piccolo, e quindi trascurabile, o di tipo "slab", parametrizzata da un λ_1 grande, che gli permette di assumere valori grandi. Formalmente si ha

$$\pi(\beta_{jk}|\gamma_{Bjk}, \lambda_0, \lambda_1) = (1 - \gamma_{Bjk})\psi(\beta_{jk}|\lambda_0) + \gamma_{Bjk}\psi(\beta_{jk}|\lambda_1)$$

per $j = 1, ..., p$ e $k = 1, ..., K^*$, dove $\psi(\beta_{jk}|\lambda)$ rappresenta la distribuzione di Laplace e

$$\gamma_{Bjk}|\theta_{Bk} \sim Bernoulli(\theta_{Bk}),$$
$$\theta_{Bk} \sim Beta(a, b).$$

Si noti che θ_{Bk} può essere interpretato come la percentuale di elementi non zero nella colonna $\boldsymbol{\beta}^k$. Questa scelta porta ad ottenere una distribuzione a priori IBP per gli elementi γ_{Bjk}. Se vengono scelti a e b pari a $a \propto 1/K^*$ e $b = 1$ si ottiene invece una approssimazione finita della IBP. Viene suggerito, inoltre, nel caso ci si aspetti sia *bicluster* molto densi che molto sparsi, di adottare dei valori degli iperparametri $a = 1/p$ e $b = 1/p$ che hanno l'effetto di concentrare la distribuzione Beta verso 0 e verso 1. Non essendo il caso in esame, ci si attende infatti che tutti i *bicluster* siano abbastanza sparsi, questa alternativa non viene presa in considerazione e si useranno i valori degli iperparametri di default ($a \propto 1/K^*$ e $b = 1$).

Dato che si vuole indurre sparsità anche sulle colonne della matrice dei fattori $\mathbf{X}_F$, a ciascun elemento x_{Fik} di essa, con $i = 1, ..., n$ e $k = 1, ..., K^*$ si impone una priori *spike and slab* con una leggera modifica rispetto al caso della matrice dei coefficienti $\mathbf{B}$, al fine di rendere trattabile l'algoritmo EM di

stima. Si introducono le variabili ausiliarie $\{\omega_{ik}\}_{i,k=1}^{n,K^*}$ per la varianza di ciascun x_{Fik},

$$x_{Fik}|\omega_{ik} \sim \mathcal{N}(0, \omega_{ik}).$$

Successivamente, ad ogni ω_{ik} viene assegnata una mistura di priori esponenziali, ovvero, a priori, ω_{ik} è generato o da uno *spike* esponenziale parametrizzato da $\tilde{\lambda}_0$, ed assume dunque un valore piccolo, o da uno *slab* esponenziale parametrizzato da $\tilde{\lambda}_1$ ed assume un valore grande,

$$\pi(\omega_{ik}|\gamma_{Fik}) = \gamma_{Fik}\frac{\tilde{\lambda}_1^2}{2}e^{-\tilde{\lambda}_1^2\omega_{ik}/2} + (1-\gamma_{Fik})\frac{\tilde{\lambda}_0^2}{2}e^{-\tilde{\lambda}_0^2\omega_{ik}/2},$$

dove γ_{Fik} è una variabile indicatrice binaria. Esattamente come nel caso di γ_{Bik} per la matrice dei coefficienti $\mathbf{B}$, γ_{Fik} è uguale a 1 se la riga i appartiene al *bicluster* k e zero altrimenti. Su ciascun γ_{Fik} viene imposta una distribuzione bernoulliana

$$\begin{aligned}\gamma_{Fik} &\sim Bernoulli(\theta_{F(k)})\\ \theta_{F(k)} &= \prod_{l=1}^{k}\nu_{(l)}\\ \nu_k &\sim Beta(\tilde{\alpha}+kd, 1-d),\end{aligned}$$

dove $d \in [0,1)$ e $\tilde{\alpha} > -d$. Analogamente a quanto visto per θ_{Bk}, l'iperparametro $\theta_{F(k)}$ può essere interpretato come la proporzione attesa a priori di elementi non zero della colonna $\boldsymbol{x}_F^k$ di $\mathbf{X}_F$. In questo caso, per $d = 0$ si ottiene la distribuzione a priori IBP per i parametri $\{\gamma_{Fik}\}_{i,k=1}^{n,K^*}$, mentre per $0 < d < 1$ si ottiene una variante della distribuzione a priori IBP chiamata Pitman-Yor. Per ulteriori dettagli su questo aspetto si rimanda a [8]. Infine, per completare il modello rimane da definire la distribuzione a priori per σ_j^2, gli elementi diagonali della matrice di varianze e covarianze di $\mathbf{E}$ in (2.2),

$$\sigma_j^2 \sim IG\left(\frac{\alpha}{2}, \frac{\alpha\delta}{2}\right),$$

dove $IG(a,b)$ indica la distribuzione gamma inversa di parametri (a,b).

L'algoritmo EM è suddiviso in due passi che si alternano iterativamente. Per descriverli adeguatamente si utilizza la seguente notazione: $\Omega = \{\omega_{ik}\}_{i,k=1}^{N,K^*} \in \mathbb{R}^{n\times K^*}$ e $\Gamma_F = \{\gamma_{Fik}\}_{i,k=1}^{N,K^*}$. Nel primo passo si calcola il valore atteso di $\mathbf{X}_F$ e Γ_F condizionatamente ai dati e al valore corrente di tutti i restanti parametri. Nel passo di massimizzazione invece si utilizza l'algoritmo accennato nel paragrafo 2.2.1, ampiamente discusso in [10], per ottenere la moda a posteriori di $\mathbf{B}$. Nel calcolo della media condizionata di $\mathbf{X}_F$ si adotta una strategia analoga ma leggermente differente. Se per la stima della moda di $\mathbf{B}$ si utilizza una griglia di valori $(\lambda_0^1 < \lambda_0^2 < ... < \lambda_0^N)$ sempre crescente, per schiacciare valori non significativi verso zero il medesimo approccio per $\mathbf{X}_F$ non è praticabile. Se per esempio si ha un $x_{Fik} = 0.005$, il contributo della riga i al *bicluster* k sembra non essere significativo. Tuttavia per un $\tilde{\lambda}_0 = 200$ è improbabile che x_{Fik} appartenga allo *spike*. Con l'aumentare di $\tilde{\lambda}_0$ dunque potrebbe accadere che valori di x_{Fik} precedentemente schiacciati a zero rientrino nel modello. Questo

problema non si presenta nella stima di $\mathbf{B}$ poiché per essa viene utilizzata la moda e non la media a posteriori. È dunque necessario fermare la griglia per $\tilde{\lambda}_0$ a valori non troppo elevati, per cui si suggerisce un $\tilde{\lambda}_0 = 5$ circa. Infine, viene utilizzata una approssimazione variazionale per trovare un limite superiore della distribuzione a posteriori dei parametri della IBP ($\nu_{(l)}$). Questo è necessario poiché la distribuzione a posteriori di questi parametri è non lineare, e dunque difficile da massimizzare. La strategia porta a passi di aggiornamento di forma chiusa per i parametri. Per ulteriori dettagli riguardanti l'algoritmo EM di stima si rimanda a [8].

Le stime dei parametri hanno una diretta interpretazione: se β_{jk} è diverso da zero, allora la colonna j della matrice dei dati Y contribuisce al *bicluster* k. Per determinare invece l'appartenenza di una riga i ad un *bicluster* k, SSLB ottiene la media a posteriori di Γ_F. L'indicatore γ_{Fik} può infatti essere interpretato come la probabilità che la riga i della matrice dei dati Y appartenga al *bicluster* k. Di conseguenza si avrà

$$\hat{x}_{Fik} = \begin{cases} \hat{x}_{Fik} & se \;\; E[\gamma_{Fik}|Y, \Omega^*, \boldsymbol{\theta}_F^*] > 0.5, \\ 0 & se \;\; E[\gamma_{Fik}|Y, \Omega^*, \boldsymbol{\theta}_F^*] \leq 0.5, \end{cases}$$

per $1 \leq i \leq n$ e $1 \leq k \leq K^*$, dove Ω^* e $\boldsymbol{\theta}_F^*$ sono le soluzioni ottenute per i parametri Ω e $\boldsymbol{\theta}_F = (\theta_{F(1)}, ..., \theta_{F(K^*)})$ dopo la convergenza dell'algoritmo. In altre parole, se la probabilità a posteriori che x_{Fik} appartenga alla parte di tipo *spike* è maggiore di 0.5, la sua stima $\hat{x}_{Fik}$ viene troncata a zero, altrimenti conserva il suo valore $\hat{x}_{Fik}$. Al termine dell'analisi vengono escluse tutte le colonne che hanno elementi tutti pari a zero sia in $\mathbf{X}_F$ che in $\mathbf{B}$, in questo modo si ottiene una scomposizione a $\hat{K} < K^*$ fattori della matrice dei dati Y.

Si noti che il materiale di modellistica di questo paragrafo è stato interamente tratto da [8].

2.2.3 Scelta del miglior adattamento

Abbiamo riscontrato che, se si esegue l'algoritmo di stima più volte, utilizzando la funzione `SSLB` dell'omonimo pacchetto di *R* ([1]) il numero di *bicluster* stimato $\hat{K}$, in maniera automatica dal metodo, varia molto a in base al punto di partenza. Dato che non è possibile impostare un valore K di *bicluster* che devono essere stimati, abbiamo deciso, al fine di ridurre l'instabilità dovuta ad aspetti di calcolo numerico, di stimare il modello 100 volte. Tra questi 100 lanci scegliamo quello che minimizza il BIC (*Bayesian information criterion*). In generale, si ha:

$$BIC = -2 \log \mathcal{L}(y|\hat{\zeta}) + q \log(m),$$

dove y sono i dati, q il numero di parametri da stimare, o gradi di libertà, m è il numero di osservazioni a disposizione, $\mathcal{L}(y|\hat{\zeta})$ è la verosimiglianza del modello e $\hat{\zeta}$ indica la stima di massima verosimiglianza di ζ che parametrizza il modello. In ambito Bayesiano, si dimostra ([1], Capitolo 11) che, quando la numerosità campionaria è grande e dunque l'apporto della distribuzione a priori è ininfluente paragonato a quello della verosimiglianza, e la stima della

moda a posteriori $\tilde{\zeta}$ è circa pari a quella di massima verosimiglianza $\hat{\zeta}$, il BIC può essere approssimato dalla seguente quantità

$$BIC \doteq -2 \log \mathcal{L}(y|\tilde{\zeta}) + q \log(m),$$

e può essere utilizzato per scegliere, tra una serie di modelli, quello che meglio si adatta ai dati.

Tradotto in termini del modello SSLB, il numero di osservazioni è pari al numero degli elementi della matrice dei dati Y, ovvero $n \times p$, la funzione di verosimiglianza è descritta in equazione (2.2), la moda a posteriori dei parametri viene ottenuta tramite un algoritmo EM e il numero di parametri da stimare è la somma tra numero di elementi delle stime a posteriori delle matrici $\mathbf{X}_F$ e $\mathbf{B}$ stimati diversi da zero.

2.3 SparseBC

2.3.1 Il modello di biclustering

Il secondo modello di *biclustering* che viene preso in considerazione in questa tesi è nettamente più semplice del precedente. Di nuovo, l'obiettivo dell'analisi è raggruppare sia le colonne che le righe della matrice dei dati. Data una matrice di dati $Y \in \mathbb{R}^{n \times p}$ l'approccio che va sotto il nome di *SparseBC* ([12]) assume che le sue righe appartengano a G classi ignote $C_1, ..., C_G$ e le sue colonne a R classi ignote, $D_1, ..., D_R$. Inoltre, ciascuna entrata della matrice y_{ij} si assume essere indipendentemente distribuita come una Gaussiana con valore atteso specifico del *bicluster* a cui appartiene, o, in altre parole, $y_{ij} \sim \mathcal{N}(\mu_{gr}, \sigma^2)$, per $i \in C_g$ e $j \in D_r$. Il modello non prevede sovrapposizioni tra *bicluster*, al contrario di *SSLB* ([8]), ciascuna entrata della matrice dei dati può appartenere ad un solo *bicluster*. Si vuole stimare C_g, D_r e μ_{gr} per $g = 1, ..., G$ e $r = 1, ..., R$. Sotto queste assunzioni, massimizzare la verosimiglianza del modello, data la matrice dei dati Y corrisponde a minimizzare la seguente quantità,

$$\underset{C_1,...,C_G,D_1,...,D_R,\boldsymbol{\mu} \in \mathbb{R}^{G \times R}}{\text{minimize}} \Big\{ \sum_{g=1}^{G} \sum_{r=1}^{R} \sum_{i \in C_g} \sum_{j \in D_r} (y_{ij} - \mu_{gr})^2 \Big\}, \tag{2.3}$$

dove $\boldsymbol{\mu} = \{\mu_{gr}\}_{g,r=1}^{G,R}$. Dato che siamo interessati a schiacciare verso 0 il conteggio di tutti quei pixel (entrate della matrice dei dati) che non contengono segnale, ovvero che appartengono al rumore di fondo, è naturale considerare la variante di questo modello che tiene conto della sparsità dei dati. In particolare, in [12] propongono l'utilizzo di una penalità lasso,

$$\underset{C_1,...,C_G,D_1,...,D_R,\boldsymbol{\mu} \in \mathbb{R}^{G \times R}}{\text{minimize}} \Big\{ \frac{1}{2} \sum_{g=1}^{G} \sum_{r=1}^{R} \sum_{i \in C_g} \sum_{j \in D_r} (y_{ij} - \mu_{gr})^2 + \lambda \sum_{g=1}^{G} \sum_{r=1}^{R} |\mu_{gr}| \Big\}, \tag{2.4}$$

dove λ è un parametro di lisciamento maggiore di zero. Al crescere di λ ad un maggior numero di *bicluster* sarà associata una media μ_{gr} pari a zero. Se

Algoritmo 2 *Sparse biclustering*

1: Inizializzare $C_1, ..., C_G$ e $D_1, ..., D_R$.
2: Iterare fino a convergenza i seguenti passi:

a) Fissati $C_1, ..., C_G$ e $D_1, ..., D_R$ risolvere (2.4) rispetto a $\boldsymbol{\mu}$, ovvero

$$\mu_{gr} = \frac{S(\sum_{i \in C_g} \sum_{j \in D_r} y_{ij}, \lambda)}{|C_g||D_r|} \tag{2.5}$$

dove $S(a, b) = \text{sign}(a)(|a| - b)_+$ è l'operatore di *soft-thresholding* e $|C_g|$ e $|D_r|$ sono rispettivamente le cardinalità di C_g e D_r.

b) Tenendo $D_1, ..., D_R$ e $\boldsymbol{\mu}$ fissati, risolvere (2.4) rispetto a $C_1, ..., C_G$ assegnando l'i-esima riga al gruppo di riga C_g per il quale $\sum_{r=1}^{R} \sum_{j \in D_r} (y_{ij} - \mu_{gr})^2$ è minimo.

c) Ripetere il passo 2 a)

d) Tenendo $C_1, ..., C_G$ e $\boldsymbol{\mu}$ fissati, risolvere (2.4) rispetto a $D_1, ..., D_R$ assegnando la j-esima colonna al gruppo di colonna D_r per il quale $\sum_{g=1}^{G} \sum_{i \in C_g} (y_{ij} - \mu_{gr})^2$ è minimo.

$\hat{\mu}_{gr} = 0$ significa che il *bicluster* (C_g, D_r) ha una media non significativamente diversa da zero. Al contrario se $\hat{\mu}_{gr} \neq 0$ vuol dire che il *bicluster* (C_g, D_r), e di conseguenza la zona della matrice ad esso associata, contiene segnale. In questo modo, il metodo conduce all'identificazione di zone dell'immagine osservata che contengono segnale, azzerando tutti i conteggi delle zone che contengono invece unicamente rumore di fondo. La speranza è di identificare un *bicluster* (C_g, D_r) i cui elementi corrispondano ai pixel che definiscono la regione di spazio dove si trova il quasar, e altri *bicluster* che corrispondano ai getti di raggi X. Per trovare un ottimo del problema di minimizzazione in equazione (2.4) è sufficiente seguire i passi dell'algoritmo 2. Per l'inizializzazione di $C_1, ..., C_G$ e $D_1, ..., D_R$ si suggerisce di utilizzare l'algoritmo di k-medie prima sulle righe e poi sulle colonne della matrice dei dati centrata.

2.3.2 Algoritmo di stima

L'algoritmo di stima prevede che siano noti il parametro di lisciamento λ e il numero di gruppi di riga e di colonna, G e R rispettivamente. In [12] viene proposto un metodo per selezionare automaticamente G e R che si basa su un approccio simile alla convalida incrociata. In particolare, per un numero elevato di volte, si esclude dalla matrice dei dati Y una certa porzione di elementi e al posto di essi, come valore viene utilizzata la media globale di Y. Si applica l'algoritmo 2 sulla matrice dei dati così rinnovata, per una serie di coppie di valori (G, R). Si utilizza come metrica di bontà di stima l'errore quadratico medio tra il valore stimato per gli elementi esclusi dalla matrice dei dati e quello

vero. Il parametro di lisciamento λ, che in questo procedimento per la scelta di G ed R è assunto noto, viene scelto, fissati G ed R, tramite il criterio di informazione Bayesiano (BIC). Combinando le due tecniche, si ottiene un algoritmo che itera i due passi fino a che le scelte di G, R, e λ non cambiano da una iterazione alla successiva. Non avremmo avuto motivo di non adottare questa tecnica, dato che nell'articolo di presentazione del metodo, l'algoritmo sembra funzionare egregiamente. Tuttavia, gli esempi presenti nell'articolo sono di applicazioni ad insiemi di dati tratti dall'ambito genomico, molto differente dal contesto astrofisico di interesse in questa tesi. L'algoritmo va adattato, poiché si è riscontrato empiricamente che l'implementazione di questa strategia porta ad una cancellazione di tutto il rumore di fondo presente nei dati ma anche del segnale associato al getto di raggi X. Questo è dovuto ad una tendenza del modello a scegliere un valore per λ che schiaccia la stima delle medie μ_{gr} fortemente verso lo zero. Di seguito presentiamo una strategia per scegliere (G, R, λ) alternativa, che mitiga questo effetto.

Si noti che tutto il materiale di modellistica presentato in questo paragrafo è stato tratto da [12].

2.3.3 Scelta di G, R, λ

Per scegliere i valori ottimali di G, R e λ utilizziamo un criterio che valuta contemporaneamente la tripletta di valori, e non prima (G, R) e poi λ iterativamente, come viene proposto in [12]. Trattandosi di un approccio frequentista, è immediato pensare al criterio di informazione BIC per scegliere tra modelli, si ha

$$BIC = -2 \log \mathcal{L}(y|\hat{\zeta}) + q \log(m).$$

Nel caso specifico del modello SparseBC che stiamo considerando, la log-verosimiglianza è la somma del logaritmo delle funzioni di densità delle singole y_{ij}, che sono normalmente distribuite, meno il termine di penalizzazione, i gradi di libertà sono pari a $G \times R$, $\hat{\zeta}$ è la stima dei parametri del modello $\{\mu_{gr}\}_{g,r=1}^{G,R}$ ottenuta utilizzando l'Algoritmo 2 e il numero di osservazioni è il numero di entrate della matrice dei dati Y ovvero $n \times p$.

Viene stimato il modello per una griglia di (G, R, λ) e per ciascuna tripletta si calcola il corrispondente BIC, si sceglie infine la tripletta $(\hat{G}, \hat{R}, \hat{\lambda})$ corrispondente al BIC minimo. Questa operazione viene ripetuta per 10 volte, al fine di mitigare fluttuazioni nelle stime dovute ad aspetti numerici. Si seleziona la tripletta coincidente con il BIC minimo tra tutte le 10 ripetizioni.

2.4 Sparse Singular Value Decomposition (SSVD)

L'ultimo metodo di *biclustering* che viene presentato in questa tesi si basa sulla ben nota scomposizione a valori singolari, in particolare su una sua versione modificata per matrici sparse ([6]).

2.4.1 Il modello

Sia Y la nostra matrice dei dati, di dimensione $n \times p$, contenente in ciascuna entrata y_{ij} il conteggio del numero di fotoni rilevati in corrispondenza del pixel di posizione (i, j), con $i = 1, ..., n$ e $j = 1, ...p$. La scomposizione a valori singolari (SVD) di Y si può formalizzare come segue:

$$Y = UDV^T = \sum_{k=1}^{r} s_k \boldsymbol{u}_k \boldsymbol{v}_k^T, \tag{2.6}$$

dove r è il rango di Y, $U = (\boldsymbol{u}_1, ..., \boldsymbol{u}_r)$ è una matrice di vettori singolari sinistri ortonormali, $V = (\boldsymbol{v}_1, ..., \boldsymbol{v}_r)$ è una matrice di vettori singolari destri ortonormali e $D = \text{diag}(s_1, ..., s_r)$ è una matrice diagonale contenente i valori singolari, positivi e tali che $s_1 \geq s_2 \geq ... \geq s_r$. SVD decompone dunque la matrice Y in una somma di matrici $s_k \boldsymbol{u}_k \boldsymbol{v}_k^T$ di rango 1. Normalmente si è interessati a considerare i primi K elementi, o strati, della somma, ovvero quelli associati a valori $s_1, ..., s_K$ grandi. Al contrario, è possibile non considerare tutti gli strati con s_k piccolo perché probabilmente rappresentano unicamente rumore di fondo. Come stima della matrice osservata si ottiene dunque una sua approssimazione di rango K:

$$Y \approx Y^{(K)} \equiv \sum_{k=1}^{K} s_k \boldsymbol{u}_k \boldsymbol{v}_k^T,$$

che si dimostra essere la miglior approssimazione di Y tramite una matrice di rango K, nel senso che minimizza il quadrato della norma di Frobenius,

$$Y^{(K)} = \underset{Y^* \in \mathcal{A}_K}{\arg\min} ||Y - Y^*||_F^2, \tag{2.7}$$

dove $\mathcal{A}_K$ rappresenta l'insieme di tutte le matrici di rango K. Per considerare la natura sparsa dei dati, si impone che $\boldsymbol{v}_k$ e $\boldsymbol{u}_k$ debbano essere sparsi, aggiungendo una penalità alla funzione obiettivo della minimizzazione in equazione (2.7). I K strati ottenuti tramite questa strategia identificheranno dunque K *bicluster* o zone della matrice dei dati contenenti segnale. Ad esempio, per il primo strato, gli elementi non zero della matrice di rango uno $s_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T$, identificheranno le zone della matrice dei dati associate al *bicluster* uno. Tramite il troncamento della sommatoria ad un numero K di strati sarà possibile tenere solo quei *bicluster* che descrivono segnale nei dati e tralasciare quelli che invece sono associabili al rumore di fondo. Sperabilmente il primo strato avrà elementi non zero in entrate della matrice dei dati corrispondenti ai pixel dove è collocato il quasar, mentre i successivi strati identificheranno i raggi X. Una tecnica per la scelta di K verrà descritta in seguito.

Si consideri di voler approssimare Y tramite una matrice di rango unitario, si cerca dunque un unico strato $s_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T$ che sia soluzione di

$$\underset{s, \boldsymbol{u}, \boldsymbol{v}}{\arg\min} ||Y - s\boldsymbol{u}\boldsymbol{v}^T||_F^2,$$

dove s è uno scalare positivo, $\boldsymbol{u}$ è un vettore n dimensionale con componenti $(u_1, ..., u_n)$ e $\boldsymbol{v}$ è un vettore p dimensionale con componenti $(v_1, ..., v_p)$. Una volta aggiunte delle penalità sia per $\boldsymbol{u}$ che per $\boldsymbol{v}$ il problema di minimizzazione diventa:

$$\arg\min_{s,\boldsymbol{u},\boldsymbol{v}} \Big\{ ||Y - s\boldsymbol{u}\boldsymbol{v}^T||_F^2 + \lambda_u P_1(s\boldsymbol{u}) + \lambda_v P_2(s\boldsymbol{v}) \Big\}, \tag{2.8}$$

dove $P_1(s\boldsymbol{u})$ e $P_2(s\boldsymbol{v})$ sono delle penalità che inducono sparsità su $\boldsymbol{u}$ e $\boldsymbol{v}$ rispettivamente e λ_u e λ_v sono dei parametri di lisciamento non negativi che controllano la magnitudine della penalità. Si ha che per $\boldsymbol{u}$ fissato la minimizzazione in equazione (2.8) rispetto a $(s, \boldsymbol{v})$ è equivalente alla minimizzazione rispetto a $\tilde{\boldsymbol{v}} = s\boldsymbol{v}$ di

$$||Y - \boldsymbol{u}\tilde{\boldsymbol{v}}^T||_F^2 + \lambda_v P_2(\tilde{\boldsymbol{v}}) = ||Q - (I_p \otimes \boldsymbol{u})\tilde{\boldsymbol{v}}||^2 + \lambda_v P_2(\tilde{\boldsymbol{v}}), \tag{2.9}$$

dove $Q = (\boldsymbol{y}_1^T, ..., \boldsymbol{y}_p^T) \in \mathbb{R}^{np}$ con $\boldsymbol{y}_j$ la j-esima colonna di Y e $\otimes$ indica il prodotto di Kronecker. Il membro destro di (2.9) corrisponde al criterio di minimizzazione di una regressione penalizzata con variabile risposta Q, matrice di disegno $I_d \otimes \boldsymbol{u}$ e coefficienti di regressione $\tilde{\boldsymbol{v}}$. Questa connessione suggerisce, per il facile ottenimento della stima di Y, di utilizzare la penalità di tipo lasso $P_2(\tilde{\boldsymbol{v}}) = \sum_{j=1}^p |\tilde{v}_j|$. Un ragionamento analogo si può fare per $\boldsymbol{u}$; infatti, fissando $\boldsymbol{v}$, la minimizzazione in (2.8) è equivalente alla minimizzazione rispetto a $\tilde{\boldsymbol{u}} = s\boldsymbol{u}$ di

$$||Y - \tilde{\boldsymbol{u}}\boldsymbol{v}^T||_F^2 + \lambda_u P_1(\tilde{\boldsymbol{u}}) = ||Z - (I_n \otimes \boldsymbol{v})\tilde{\boldsymbol{u}}||^2 + \lambda_u P_1(\tilde{\boldsymbol{u}}), \tag{2.10}$$

dove $Z = (\boldsymbol{y}_{(1)}, ..., \boldsymbol{y}_{(n)}) \in \mathbb{R}^{np}$, con $\boldsymbol{y}_{(i)}^T$ l'i-esima riga di Y. In questo caso Z veste i panni della variabile risposta e $(I_n \otimes \boldsymbol{v})$ della matrice di disegno. Data questa similitudine con il contesto della regressione penalizzata anche per $P_1(\tilde{\boldsymbol{u}})$ si protende verso la scelta di una penalità di tipo lasso $P_1(\tilde{\boldsymbol{u}}) = \sum_{i=1}^n |\tilde{u}_i|$. Viene in realtà fatto un passo ulteriore ovvero considerare una classe di penalità più ampia che va in letteratura sotto il nome di penalità lasso adattiva ([14]), per la quale avremo

$$P_1(s\boldsymbol{u}) = s\sum_{i=1}^n w_{1,i}|u_i|, \qquad P_2(s\boldsymbol{v}) = s\sum_{j=1}^p w_{2,j}|v_j|,$$

dove $\{w_{1,i}\}_{i=1}^n$ e $\{w_{2,j}\}_{j=1}^p$ sono un sistema di pesi che deve tenere conto della grandezza degli u_i e v_j rispettivamente, al fine di schiacciare maggiormente verso zero gli elementi più vicini a zero e viceversa non schiacciare quelli lontani da zero. Un modo semplice per definire questi pesi è utilizzare la stima ai minimi quadrati di $\tilde{\boldsymbol{v}}$ e $\tilde{\boldsymbol{u}}$ ottenibile minimizzando (2.9) e (2.10) rispetto a $\tilde{\boldsymbol{v}}$ e $\tilde{\boldsymbol{u}}$ rispettivamente, senza considerare il termine di penalità. Il sistema di pesi viene dunque definito come segue

$$\boldsymbol{w}_1 = (w_{1,i}, ..., w_{1,n}) = |\hat{\tilde{\boldsymbol{u}}}|^{-\gamma_1} \qquad \boldsymbol{w}_2 = (w_{2,i}, ..., w_{2,p}) = |\hat{\tilde{\boldsymbol{v}}}|^{-\gamma_2},$$

dove $\hat{\tilde{\boldsymbol{v}}}$ e $\hat{\tilde{\boldsymbol{u}}}$ sono gli stimatori ai minimi quadrati di $\tilde{\boldsymbol{v}}$ e $\tilde{\boldsymbol{u}}$ rispettivamente, l'operatore $|\cdot|$ è da intendere come applicato elemento per elemento al vettore

che ha come argomento e γ_1 e γ_2 sono dei parametri non negativi che regolano il sistema di pesistica. In questo modo, ad un $\tilde{u}_i$ piccolo corrisponderà dunque un peso grande e di conseguenza una maggior penalizzazione, proprio come intuitivamente vorremmo che accada.

2.4.2 Algoritmo di stima

Con la penalità lasso adattiva, la funzione obiettivo da minimizzare ha forma

$$||Y - s\boldsymbol{u}\boldsymbol{v}^T||_F^2 + s\lambda_u \sum_{i=1}^{n} w_{1,i}|u_i| + s\lambda_v \sum_{j=1}^{p} w_{2,j}|v_j|. \tag{2.11}$$

L'algoritmo di stima, riportato in Algoritmo 3, itera due passi, fissando alternativamente $\boldsymbol{v}$ ed $\boldsymbol{u}$ e si basa su semplici regole di *soft-thresholding* componente per componente, le medesime utilizzate nella stima dei parametri nelle regressioni con penalizzazione lasso. Per $\boldsymbol{u}$ fissato, si ha che minimizzare (2.11) è equivalente a minimizzare la quantità

$$||Y-\boldsymbol{u}\tilde{\boldsymbol{v}}^T||_F^2+\lambda_v \sum_{j=1}^{p} w_{2,j}|\tilde{v}_j| = ||Y||_F^2+\sum_{j=1}^{p}\{\tilde{v}_j^2-2\tilde{v}_j(Y^T\boldsymbol{u})_j+\lambda_v w_{2,j}|\tilde{v}_j|\}, \tag{2.12}$$

dove tra il primo e il secondo membro è stato svolto il quadrato ed è stata esplicitata la sommatoria della norma di Frobenius e $(Y^T\boldsymbol{u})_j$ indica l'elemento j-esimo del vettore tra parentesi. Si noti che $||Y||_F^2$ è una costante e può essere dunque tralasciata ed inoltre si può procedere alla minimizzazione della sommatoria elemento per elemento. Utilizzando la regola di *soft-thresholding* si ottiene che i $\tilde{v}_j$ che minimizzano (2.12) sono pari a

$$\text{sign}\{(Y^T\boldsymbol{u})_j\}(|(Y^T\boldsymbol{u})_j| - \lambda_v w_{2,j}/2)_+,$$

con $j = 1, ..., p$. Analogamente, fissando $\boldsymbol{v}$ e minimizzando rispetto a $\tilde{\boldsymbol{u}}$ si ottiene come stima per gli elementi $\tilde{u}_i$

$$\text{sign}\{(Y\boldsymbol{v})_i\}(|(Y\boldsymbol{v})_i| - \lambda_u w_{1,i}/2)_+,$$

con $i = 1, ..., n$. Vengono successivamente scalate le due stime ottenendo prima le costanti $c_1 = ||\tilde{\boldsymbol{v}}||$ e $c_2 = ||\tilde{\boldsymbol{u}}||$ e poi definendo $\boldsymbol{v} = \tilde{\boldsymbol{v}}/c_1$ e $\boldsymbol{u} = \tilde{\boldsymbol{u}}/c_2$. I due passi vengono iterati fino a convergenza, ovvero fino a quando le stime non cambiano più di una piccola soglia di tolleranza. Una volta raggiunta la convergenza, si pone $s = \boldsymbol{u}^T Y \boldsymbol{v}$, e il primo strato, migliore approssimazione possibile di Y tramite una matrice di rango 1, è dato da $s\boldsymbol{u}\boldsymbol{v}^T$. Per ottenere i successivi strati è sufficiente applicare la medesima procedura sulla matrice residua $Y - s\boldsymbol{u}\boldsymbol{v}^T$.

A questa procedura manca un passaggio fondamentale, ovvero la stima dei parametri di lisciamento, λ_u e λ_v, che regolano la quantità di penalizzazione. Dato che questi due parametri controllano quanto gli elementi dei vettori $\boldsymbol{u}$ e $\boldsymbol{v}$ vengono schiacciati verso lo zero è lecito affermare che sono strettamente legati al grado di sparsità dei due vettori $\boldsymbol{u}$ e $\boldsymbol{v}$. Si può sfruttare il fatto che

Algoritmo 3 *Sparse Singular Value Decomposition*

1: Si applica la scomposizione SVD classica ad X e si denota con $\{s_{old}, \boldsymbol{u}_{old}, \boldsymbol{v}_{old}\}$ il primo strato di tale scomposizione.

2: Iterare fino a convergenza i seguenti passi:

a) Sia $\tilde{v}_j = \text{sign}\{(Y^T\boldsymbol{u}_{old})_j\}(|(Y^T\boldsymbol{u}_{old})_j| - \lambda_v w_{2,j}/2)_+$, per $j = 1, ..., p$ e con λ_v pari al valore che minimizza $BIC(\lambda_v)$ in equazione (2.14). Sia $\tilde{\boldsymbol{v}} = (\tilde{v}_1, ..., \tilde{v}_p)^T$. Porre $c_1 = ||\tilde{\boldsymbol{v}}||$ e successivamente $\boldsymbol{v}_{new} = \tilde{\boldsymbol{v}}/c_1$

b) Sia $\tilde{u}_j = \text{sign}\{(Y\boldsymbol{v}_{new})_i\}(|(Y\boldsymbol{v}_{new})_i| - \lambda_u w_{1,i}/2)_+$, per $i = 1, ..., n$ e con λ_u pari al valore che minimizza $BIC(\lambda_u)$ in equazione (2.13). Sia $\tilde{\boldsymbol{u}} = (\tilde{u}_1, ..., \tilde{u}_n)^T$. Porre $c_2 = ||\tilde{\boldsymbol{u}}||$ e successivamente $\boldsymbol{u}_{new} = \tilde{\boldsymbol{u}}/c_2$

c) Porre $\boldsymbol{u}_{old} = \boldsymbol{u}_{new}$ e ripetere i passi 2a) e 2b) fino a convergenza.

3: Raggiunta la convergenza porre $\boldsymbol{u} = \boldsymbol{u}_{new}$, $\boldsymbol{v} = \boldsymbol{v}_{new}$, $s = \boldsymbol{u}_{new}^T Y \boldsymbol{v}_{new}$

in una regressione con penalizzazione lasso il numero di coefficienti diversi da zero è uno stimatore non distorto dei gradi di libertà del modello, come viene dimostrato in [15]. Si può dunque utilizzare il criterio di informazione Bayesiano (BIC) per scegliere i valori di λ_v e λ_u ottimali. Per la regressione penalizzata in (2.9) con $\boldsymbol{u}$ fissato si definisce

$$BIC(\lambda_v) = \frac{||Q - \hat{Q}||^2}{np \cdot \hat{\sigma}_1^2} + \frac{log(np)}{np}\hat{df}(\lambda_v), \tag{2.13}$$

dove $\hat{\sigma}_1^2$ è la stima ai minimi quadrati della varianza del modello di regressione e $\hat{df}(\lambda_v)$ sono i gradi di libertà associati alla scelta di λ_v come parametro di penalizzazione del modello. Analogamente, per la regressione in equazione (2.10), per $\boldsymbol{v}$ fissato si definisce

$$BIC(\lambda_u) = \frac{||Z - \hat{Z}||^2}{np \cdot \hat{\sigma}_2^2} + \frac{log(np)}{np}\hat{df}(\lambda_u), \tag{2.14}$$

dove $\hat{\sigma}_2^2$ è la stima ai minimi quadrati della varianza del modello di regressione e $\hat{df}(\lambda_u)$ sono i gradi di libertà associati alla scelta di λ_u come parametro di penalizzazione del modello. La selezione dei parametri di penalizzazione si basa sulla minimizzazione di queste due quantità che avviene all'interno dei passi iterativi dell'algoritmo. In questo modo si evita di selezionare i due parametri contemporaneamente, processo che richiede un costo computazionale maggiore, rispetto alla selezione di un parametro alla volta.

Si noti che tutto il materiale di modellistica presentato in questo paragrafo è stato tratto da [6].

2.4.3 Scelta del numero di *bicluster* K

Per ottenere una stima della matrice dei dati, e dunque dell'immagine osservata, è necessario scegliere il numero di strati, o *bicluster* K da considerare. Non

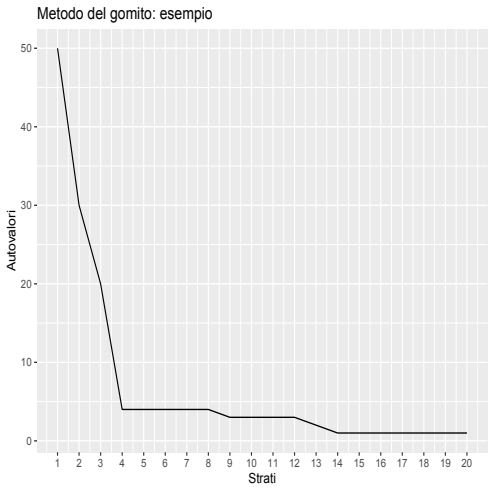

Figura 2.1: *Esempio di grafico per utilizzare la regola del gomito nella scelta del numero di strati, o bicluster, K da considerare. In ascissa il numero di strati, in ordinata, i corrispondenti valori degli autovalori.*

vorremmo utilizzare troppi strati poiché si rischierebbe di includere nella stima rumore di fondo, ma neanche troppo pochi poiché si perderebbe del segnale. Dobbiamo quindi trovare una strategia per decidere dove fermarci nell'inclusione. A questo fine, sfruttiamo le proprietà dei valori singolari e in particolare di una loro trasformazione. Consideriamo una scomposizione a valori singolari classica e non sparsa, come in equazione (2.6), allora i valori singolari s_k per $k = 1, ..., K$ di Y e i suoi autovalori ρ_k sono legati dalla relazione:

$$\rho_k = s_k^2.$$

Dato che l'autovalore più grande di una matrice descrive la direzione, data dal corrispondente autovettore, di maggior variabilità, può essere utilizzato come una metrica di importanza degli strati per decidere quanti tenerne. In analisi delle componenti principali, uno dei metodi per decidere quante componenti utilizzare per descrivere i dati è la regola del gomito. Si illustra graficamente la curva che descrive il valore degli autovalori ρ_k al variare di k e si sceglie un numero di componenti pari al punto in cui la curva si appiattisce, ovvero dove si registra l'ultima rapida decrescita del valore di ρ_k. Per esempio, in Figura 2.1 si sceglierebbe 4 come valore, poiché dal quinto valore in poi la curva si appiattisce. Anche per scegliere il numero di strati da considerare in SSVD abbiamo deciso di utilizzare questo metodo, con la consapevolezza che la relazione tra valori singolari e autovalori è solamente approssimata in questo caso, poiché applicando SSVD, si considera una variante sparsa della scomposizione a valori singolari.

2.5 Riassumendo

In questo capitolo sono stati descritti tre metodi per il *biclustering*, ovvero metodi per raggruppare simultaneamente righe e colonne della matrice dei

dati. Il *biclustering* è particolarmente utile in questo contesto poiché sia le righe che le colonne della matrice dei dati sono di interesse scientifico. L'obiettivo dell'analisi è infatti individuare zone dell'immagine, e dunque della matrice dei dati, contenenti segnale. I tre metodi sono

- Lo *Spike and Slab Lasso Biclustering*, [8], che si serve di una scomposizione in fattori latenti per approssimare la matrice dei dati. Vengono stimati i fattori ed i coefficienti della scomposizione tramite un approccio Bayesiano che prevede l'utilizzo di priori di tipo *spike and slab* che hanno l'effetto di indurre sparsità sui fattori e sui coefficienti. Ciascun fattore identifica un bicluster, in quanto essi sono espressione della relazione tra un insieme di righe ed un insieme di colonne. Idealmente vorremmo che il primo fattore identifichi il quasar e i successivi gli eventuali getti di raggi X presenti nell'immagine. Per selezionare il numero di fattori da considerare lanciamo l'algoritmo di stima 100 volte e scegliamo il miglior adattamento tramite un'approssimazione in ambito Bayesiano del criterio di informazione BIC.

- Il modello *SparseBC* ([12]), assume che le righe della matrice dei dati appartengano a G classi ignote $C_1, ..., C_G$ mentre le colonne a R classi ignote, $D_1, ..., D_R$. Inoltre, ciascuna entrata y_{ij} della matrice dei dati Y si assume essere indipendentemente distribuita come una Gaussiana con valore atteso specifico del *bicluster* a cui appartiene, o, in altre parole, $y_{ij} \sim \mathcal{N}(\mu_{gr}, \sigma^2)$, per $i \in C_g$ e $j \in D_r$. Per trovare la stima delle medie di ciascun *bicluster* il metodo minimizza la quantità:

$$\underset{C_1,...,C_G,D_1,...,D_R,\boldsymbol{\mu}\in\mathbb{R}^{G\times R}}{\text{minimize}} \Big\{\frac{1}{2}\sum_{g=1}^{G}\sum_{r=1}^{R}\sum_{i\in C_g}\sum_{j\in D_r}(y_{ij}-\mu_{gr})^2 + \lambda\sum_{g=1}^{G}\sum_{r=1}^{R}|\mu_{gr}|\Big\},$$

  dove $\boldsymbol{\mu} = \{\mu_{gr}\}_{g,r=1}^{G,R}$ e λ è un parametro di lisciamento maggiore di zero. Minimizzare questa quantità corrisponde a massimizzare la funzione di verosimiglianza a cui viene aggiunta una penalità di tipo lasso. Per scegliere la tripletta (G, R, λ) che porta il miglior adattamento ai dati viene utilizzato il criterio di informazione BIC. La speranza è di identificare un *bicluster* (C_g, D_r) i cui elementi corrispondano ai pixel che definiscono la regione di spazio dove si trova il quasar, e altri *bicluster* che corrispondano ai getti di raggi X.

- Il modello *Sparse Singular Value Decomposition* ([6]), si basa su una versione sparsa della scomposizione a valori singolari per scomporre la matrice dei dati Y in K matrici di rango 1, dette anche strati. I K strati ottenuti tramite questa strategia identificano K *bicluster* o zone contenenti segnale della matrice dei dati e quindi dell'immagine. Ad esempio, per il primo strato, gli elementi non zero della matrice di rango uno che lo descrive, corrispondono alle entrate della matrice dei dati associate al *bicluster* uno. Sperabilmente il primo strato avrà elementi non zero in entrate della matrice dei dati corrispondenti ai pixel dove è collocato il

quasar, mentre i successivi strati identificheranno i raggi X. Per scegliere il numero di strati da considerare viene sfruttata la relazione tra valori singolari e autovalori che permette di utilizzare la regola del gomito.

Capitolo 3

Studi di simulazione

3.1 Disegno di simulazione

Gli studi di simulazione descritti in questo capitolo hanno la finalità di saggiare i limiti dei metodi per il *biclustering* presentati nel Capitolo 2 in contesti particolarmente complessi. Per la loro costruzione si è fatto riferimento a [11].

Le immagini che simuliamo sono frammentate in una griglia di 64×64 pixel e possono essere dunque descritte da una matrice di dimensione 64×64. Coerentemente con quanto visto fino ad ora, definiamo con Y la matrice dei dati simulati e con y_{ij}, $i, j = 1, ..., 64$, ciascuna entrata di Y. Nel caso più generale possibile, l'immagine rappresenterà un quasar centrale, getti di raggi X e il rumore di fondo. Siamo interessati a simulare la matrice di conteggi Y in 3 fasi, prima vengono simulati i conteggi associati al quasar, poi quelli ai getti di raggi X e infine quelli associati al rumore di fondo, le tre componenti vengono poi sommate per comporre l'immagine simulata. La griglia con la quale è stata partizionata l'immagine definisce anche la grandezza di ciascun pixel, al quale, si ricorda, è associata una porzione vera e propria di spazio. Tralasciando la proporzione di superficie che i pixel rappresentano, supponiamo che ognuno di essi ricopra un'area di 1×1. In questo modo, l'intera immagine ricopre un'area di 64×64 unità al quadrato e il pixel di posizione $(1, 1)$ ricopre l'area compresa nei punti di coordinate $A = (0, 0)$, $B = (0, 1)$, $C = (1, 0)$ e $D = (1, 1)$. L'entrata y_{11} esprimerà dunque il conteggio dei fotoni simulati nell'area dell'immagine compresa tra i punti A, B, C, D ovvero in corrispondenza del pixel di posizione $(1, 1)$.

Per il quasar si suppone di osservare un totale di 500 fotoni, distribuiti nell'immagine secondo una Gaussiana bivariata sferica (a componenti dunque indipendenti) con deviazione standard 0.5 e media centrata nell'immagine. La distribuzione ha dunque valore atteso μ e matrice di varianze e covarianze Σ, con

$$\mu = (32, 32) \qquad \Sigma = \begin{bmatrix} 0.5^2 & 0 \\ 0 & 0.5^2 \end{bmatrix}. \tag{3.1}$$

Questo rispecchia il fatto che vogliamo che il quasar sia centrale nell'immagine ($\mu = (32, 32)$) e copra un'area ristretta (varianza delle singole componenti piccola). Per ciascun pixel, si calcola il punto centrale, per esempio, per il pixel

di posizione $(1,1)$ che ricopre l'area compresa nei punti A, B, C e D il punto centrale sarà $(0.5, 0.5)$. Si calcola il valore della densità Gaussiana in equazione (3.1) in ciascun punto centrale, si normalizzano le quantità dividendole per la loro somma e si ottiene così la percentuale di fotoni provenienti dal quasar che ci aspettiamo in ciascun pixel. Seguendo quanto delineato in [11] moltiplichiamo questa quantità per 500, ovvero il numero totale di fotoni che vogliamo nell'immagine per il quasar, ottenendo l'intensità del quasar per ciascun pixel, che chiameremo t_{ij}, con $i, j = 1, ..., 64$. Successivamente, si simula un valore da una distribuzione Poisson con valore atteso t_{ij} e si ottiene la matrice di conteggi simulati per il quasar.

Un procedimento analogo viene sfruttato per generare i conteggi associati ai getti di raggi X. In questo caso però le componenti della Gaussiana bivariata che distribuisce i fotoni associati ai getti avrà elementi Σ_{getto} e μ_{getto} che cambiano a seconda del conteso di simulazione. Se, per esempio, si desidera un getto vicino al quasar si avrà $\mu_{getto} = (34, 34)$, se invece si desidera un getto distante dal quasar si avrà $\mu_{getto} = (45, 45)$. Oppure, per ottenere un getto concentrato si avrà $\Sigma_{getto} = \begin{bmatrix} 0.5^2 & 0.5^3 \\ 0.5^3 & 0.5^2 \end{bmatrix}$, mentre se si desidera un getto esteso $\Sigma_{getto} = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$. Anche il conteggio totale del numero di fotoni associati al getto varierà da simulazione a simulazione, si parlerà di getto debole quando il totale è 10, di getto medio quando è 20 e di getto forte quando è 35.

Si ipotizza un numero totale di fotoni attribuibili al rumore di fondo pari a 200, distribuiti nell'immagine secondo una uniforme. Per le proporzioni tra numero totale di fotoni provenienti da quasar, getto e rumore di fondo abbiamo ricalcato gli studi di simulazione in [11].

3.2 Caso A: nessun getto

Nel contesto di simulazione A vogliamo indagare la nettezza dei metodi per il *biclustering*, ovvero la capacità di non aggiungere segnale quando non è presente. A questo scopo viene generato solamente un quasar centrale, con numero atteso di fotoni 500 e rumore di fondo uniforme, con numero atteso di fotoni pari a 200. Idealmente, i metodi dovrebbero essere capaci di schiacciare a zero il conteggio dei fotoni nei pixel corrispondenti al rumore di fondo e identificare come segnale i pixel corrispondenti alla porzione di spazio nella quale sono stati simulati i conteggi associati al quasar. Questo primo contesto permette di analizzare la capacità dei metodi di distinguere tra rumore e segnale.

Verranno considerate due tipologie di quasar, il primo analogo a quello descritto sopra e il secondo più diffuso, che corrisponde ad aumentare gli elementi della matrice di varianze e covarianze della distribuzione Gaussiana che ne distribuisce i fotoni nell'immagine. Quest'ultimo quasar è utile per sondare la sensibilità del metodo. Ci aspettiamo che più l'oggetto è diffuso più il modello fatichi ad identificarne correttamente i contorni. Le due immagini simulate, a cui ci riferiremo tramite gli abbreviativi A1 e A2, intendendo che

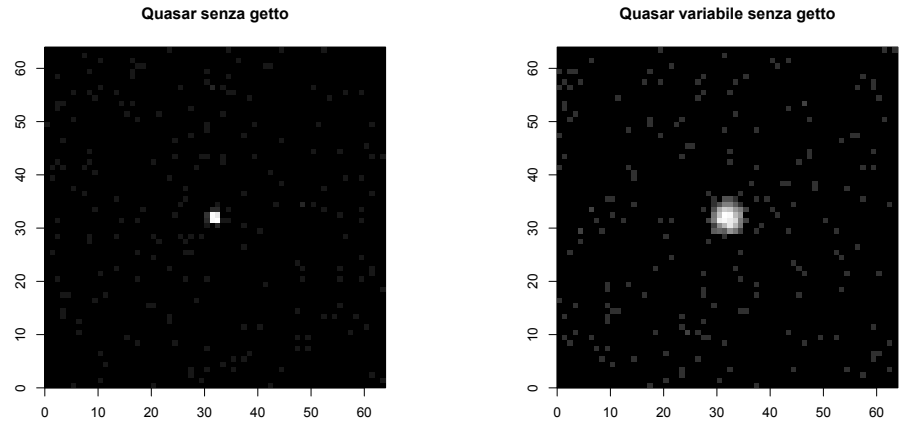

Figura 3.1: *A sinistra: immagine di un quasar centrale con numero atteso di fotoni* 500*, elementi di varianza della Gaussiana che ne distribuisce i fotoni nell'immagine* 0.5^2 *e rumore di fondo con numero atteso di fotoni* 200*, uniformemente distribuiti nell'immagine. A destra: immagine di un quasar centrale con numero atteso di fotoni* 500*, elementi di varianza della Gaussiana che ne distribuisce i fotoni nell'immagine* 2 *e rumore di fondo uniforme con numero atteso di fotoni* 200*, uniformemente distribuiti nell'immagine.*

sono la prima e la seconda immagine simulata nel contesto A, sono raccolte in Figura 3.1, dove a pixel colorati di nero corrispondono porzioni di spazio dove non sono stati simulati fotoni, mentre a pixel colorati di bianco corrispondono zone dove è stato simulato il conteggio di fotoni massimo per l'immagine. Le gradazioni intermedie, invece, indicano un pixel in corrispondenza del quale il conteggio è una via di mezzo tra 0 e il massimo simulato nell'immagine.

3.3 Caso B: getto distante dal quasar

Nel contesto di simulazione B vengono simulate immagini con getti di raggi X distanti dal quasar centrale. Una delle principali difficoltà nell'identificazione delle sorgenti diffuse è la loro debolezza se confrontata con l'alta luminosità del quasar. È logico aspettarsi che più la sorgente debole (getto) è vicina alla sorgente luminosa (quasar) più il modello fatichi a distinguere il segnale del getto dal rumore di fondo. In altre parole, conteggi bassi vicino a conteggi molto alti tendono ad essere categorizzati come rumore e non come segnale. Si ritiene dunque che un getto distante dal quasar sia più facilmente identificabile dai metodi per il *biclustering*.

In particolare, si considerano 4 diverse situazioni: quasar più getto debole (B1), quasar più getto forte (B2), quasar più due getti deboli (B3) e quasar più due getti forti (B4), tutti naturalmente lontani dal quasar. Si noti che la distanza dal quasar è data dalla centratura della normale bivariata tramite la quale vengono distribuiti i fotoni dei getti nell'immagine. Quando è presente

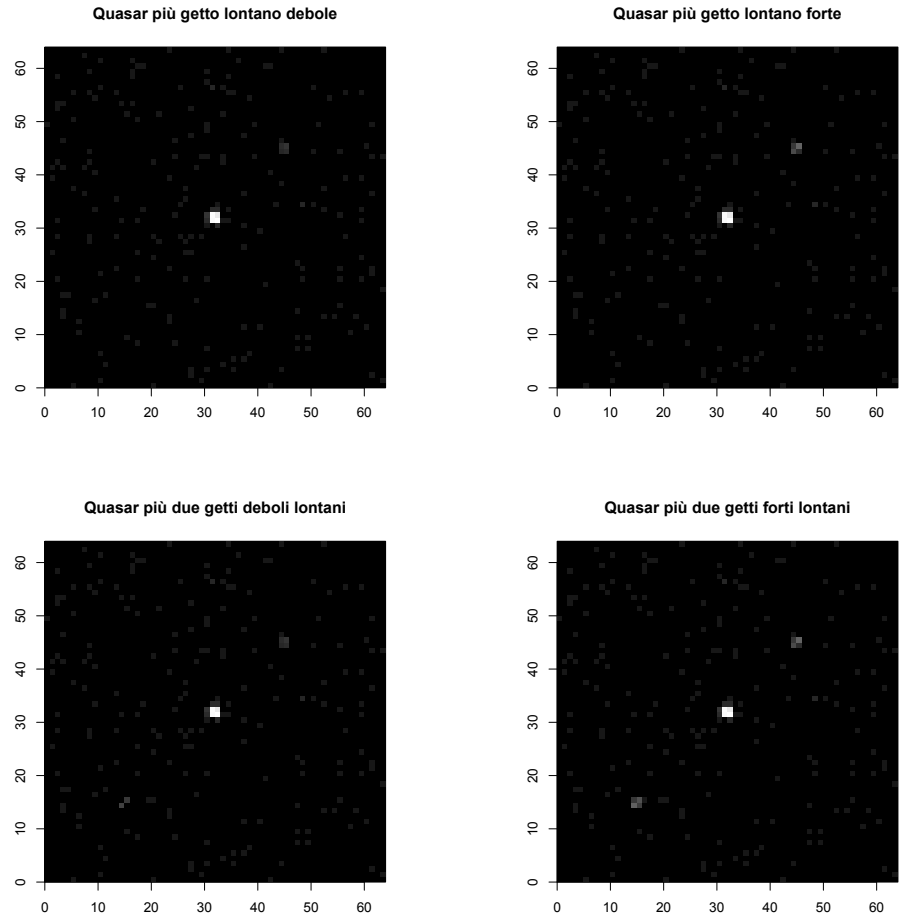

Figura 3.2: *In alto a sinistra: immagine di un quasar centrale con numero atteso di fotoni* 500 *e elementi di varianza della Gaussiana che ne distribuisce i fotoni nell'immagine pari a* 0.5^2*, getto lontano con numero atteso di fotoni* 10*, centrato in* $(45, 45)$*, elementi di varianza e di covarianza della Gaussiana che ne distribuisce i fotoni nell'immagine pari a* 0.5^2 *e* 0.5^3 *rispettivamente, e rumore di fondo con numero atteso di fotoni* 200*, uniformemente distribuiti nell'immagine. In alto a destra: come in alto a sinistra ma con numero atteso di fotoni per il getto pari a* 30*. In basso a sinistra: come in alto a sinistra ma con due getti, il secondo centrato in* $(15, 15)$*. In basso a destra: come in alto a destra ma con due getti, il secondo centrato in* $(15, 15)$*.*

un unico getto, esso è centrato in $(45, 45)$, quando ce ne sono due, il secondo è centrato in $(15, 15)$. Presentiamo le 4 immagini simulate in Figura 3.2.

3.4 Caso C: getto contiguo

Nel contesto di simulazione C simuliamo immagini che presentano un getto di raggi X contiguo al quasar centrale. Sicuramente questo è un contesto più

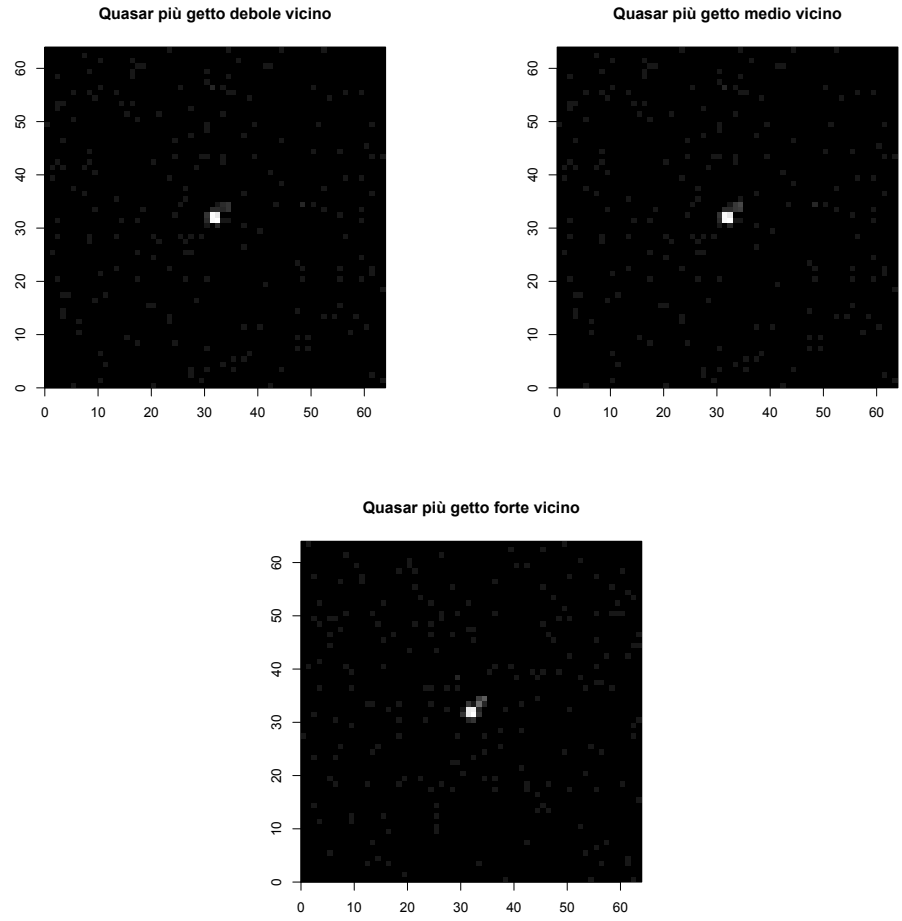

Figura 3.3: *In alto a sinistra: immagine di un quasar centrale con numero atteso di fotoni* 500 *ed elementi di varianza della Gaussiana che ne distribuisce i fotoni nell'immagine pari a* 0.5^2*, getto contiguo con numero atteso di fotoni* 10*, centrato in* $(34, 34)$*, elementi di varianza e di covarianza della Gaussiana che ne distribuisce i fotoni nell'immagine pari a* 0.5^2 *e* 0.5^3 *rispettivamente e rumore di fondo con numero atteso di fotoni* 200*, uniformemente distribuiti nell'immagine. In alto a destra: come in alto a sinistra ma con getto con numero atteso di fotoni* 20*. In basso: come in alto a sinistra ma con getto con numero atteso di fotoni* 35*.*

problematico rispetto ad A e B, per quanto detto nel paragrafo 3.2 sulla difficoltà di identificare sorgenti deboli vicine a sorgenti forti. Vanno inoltre fatte delle considerazioni specifiche ai modelli di *biclustering*. La vicinanza tra getto e quasar potrebbe infatti richiedere che i modelli consentano a due o più bicluster di intersecarsi (*overlapping* in letteratura). Ci si attende, per esempio, che il modello SparseBC non sia il quello più performante in quest'ultima simulazione perché non prevede che i *bicluster* si possano sovrapporre. Al contrario SSLB dovrebbe essere in grado di identificare il getto poiché permette ad un'entrata della matrice di appartenere a più *bicluster* e quindi a due o più *bicluster* di sovrapporsi. Consideriamo 3 diverse combinazioni di quasar

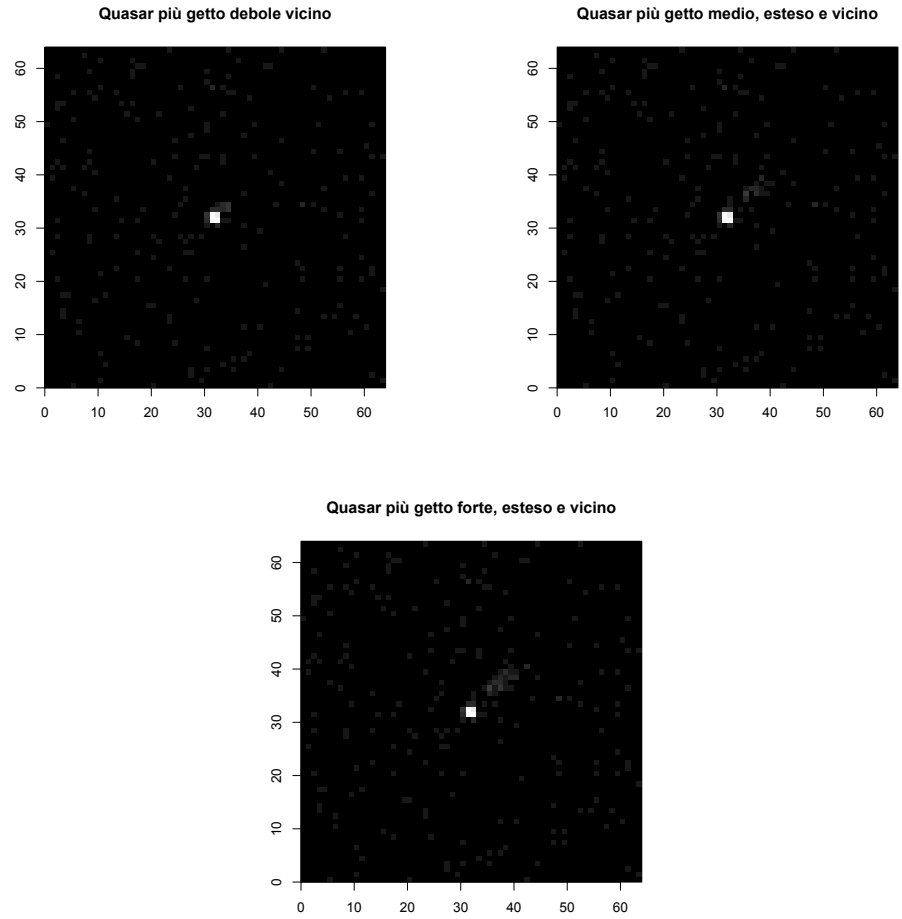

Figura 3.4: *In alto a sinistra: immagine di un quasar centrale con numero atteso di fotoni* 500 *ed elementi di varianza della Gaussiana che ne distribuisce i fotoni nell'immagine pari a* 0.5^2*, getto esteso e vicino con numero atteso di fotoni* 10*, centrato in* $(38, 38)$*, elementi di varianza e di covarianza della Gaussiana che ne distribuisce i fotoni nell'immagine pari a* 4 *e* 3 *rispettivamente, e rumore di fondo con numero atteso di fotoni* 200*, uniformemente distribuiti nell'immagine. In alto a destra: come in alto a sinistra ma con getto con numero atteso di fotoni* 20*. In basso: come in alto a sinistra ma con getto con numero atteso di fotoni* 35*.*

e getto di raggi X contiguo: quasar più getto debole (C1), quasar più getto medio (C2) e quasar più getto forte (C3). Le immagini simulate sono riportate in Figura 3.3.

3.5 Caso D: getto allungato

Nell'ultimo contesto di simulazione D si simulano immagini che presentano un quasar con un getto di raggi X diffuso, abbastanza vicino ad esso. In particolare, aumentiamo le componenti di varianza della matrice di varianze e covarianze

della Gaussiana, che distribuisce i fotoni associati al getto nell'immagine, da 0.5^2 a 4, mentre gli elementi di covarianza da 0.5^3 a 3. Il fine è analizzare come si comportano i metodi per il *biclutering* al variare della diffusione del getto presente nell'immagine simulata. In altre parole, il modello funziona meglio se deve identificare una zona di segnale più estesa, ma meno intensa, o più intensa, ma più compatta? Anche in questo caso si considerano 3 diverse situazioni: quasar più getto debole (D1), quasar più getto medio (D2), quasar più getto forte (D3). Le immagini simulate sono apprezzabili in Figura 3.4.

3.6 Confronto tra modelli

Al fine di confrontare tra loro le immagini ricostruite dai tre metodi per il *biclustering* è necessario definire alcuni indici che permettano di riassumere la bontà dell'adattamento dei modelli ai dati, in questo caso simulati. Definiamo il concetto di immagine ideale. Abbiamo simulato le immagini per il contesto A in due fasi: prima i conteggi associati al quasar e poi quelli al rumore di fondo, mentre per i contesti B, C, e D abbiamo simulato i conteggi del quasar, poi quelli del/dei getto/i e poi quelli del rumore di fondo. Non vogliamo che i metodi includano, nell'immagine ricostruita, i conteggi associati al rumore di fondo. D'altra parte, vogliamo che includano il segnale proveniente dal quasar e dal/dai getto/i di raggi X. L'immagine ideale dunque è composta per il contesto A unicamente dai conteggi associati al quasar, mentre per i contesti B, C, e D da quelli del quasar e del/dei getto/i di raggi X. Utilizziamo questa immagine ideale, o obiettivo, come riferimento per confrontare tra loro le immagini ricostruite dai tre metodi. Ovviamente, come per l'immagine osservata costruiamo la matrice dei dati osservati, in modo analogo per l'immagine obiettivo disponiamo della sua traduzione in formato matriciale, che chiameremo matrice obiettivo.

In particolare, siamo interessati alla percentuale di conteggi pari a zero della matrice obiettivo, correttamente stimati come zero anche dal nostro metodo. In termini formali,

$$\iota_0 = \frac{\sum_{(i,j)\in\mathcal{A}_0^*} I(\hat{y}_{ij} = 0)}{|\mathcal{A}_0^*|}, \qquad \iota_0 \in [0,1],$$

dove $\mathcal{A}_0^*$ indica l'insieme delle entrate y_{ij}^*, per $i, j = 1, ..., 64$ della matrice ideale Y^* che sono uguali a zero, $\hat{y}_{ij}$ indica l'entrata (i, j) della matrice stimata $\hat{Y}$, $I(\cdot)$ è la funzione indicatrice e $|\cdot|$ è l'operatore cardinalità. Questo indice, che assume valori tra zero e uno, ci permettere di comprendere che percentuale del rumore di fondo il nostro metodo è stato in grado di eliminare. È di fondamentale importanza avere un valore prossimo ad 1 per ι_0, poiché l'obiettivo dell'analisi è proprio distinguere il segnale presente nell'immagine dal rumore di fondo, ovvero determinare in corrispondenza di quali pixel il conteggio osservato proviene da una sorgente luminosa di interesse, o da semplici elementi di disturbo.

In secondo luogo, valutiamo la percentuale di conteggi diversi da zero della matrice ideale, correttamente stimati come diversi da zero anche dai nostri

metodi. In termini formali,

$$\iota_{not} = \frac{\sum_{(i,j)\in\mathcal{A}^*_{not}} I(\hat{y}_{ij} \neq 0)}{|\mathcal{A}^*_{not}|}, \qquad \iota_{not} \in [0,1],$$

dove $\mathcal{A}^*_{not}$ indica l'insieme delle entrate della matrice ideale Y^* che sono diverse da zero. L'indice di maggior interesse è ι_0, ovvero la percentuale di volte in cui, nei pixel dell'immagine obiettivo in corrispondenza dei quali è stato rilevato un conteggio pari a zero, è stato stimato un conteggio pari a zero anche dal metodo. Seguendo una strada conservativa, preferiamo infatti che i modelli non rilevino eventuali getti di raggi X quando sono presenti, piuttosto che ne rilevino quando assenti. Naturalmente, a parità di percentuale di zeri correttamente stimati, si preferisce un modello che identifica correttamente una maggiore percentuale di conteggi diversi da zero, ovvero a cui corrisponde un ι_{not} maggiore.

Infine, consideriamo l'errore quadratico medio (MSE) tra gli elementi non zero di Y^* e i corrispettivi elementi della matrice stimata $\hat{Y}$. In altre parole,

$$\text{MSE} = \frac{\sum_{(i,j)\in\mathcal{A}^*_{not}} (y^*_{ij} - \hat{y}_{ij})^2}{|\mathcal{A}^*_{not}|}.$$

In questo modo siamo in grado di valutare quanto accuratamente sia stato ricostruito il segnale dai modelli per il *biclustering*. Ci interessiamo a questa metrica unicamente dopo aver analizzato le precedenti due e ne spieghiamo la motivazione con un esempio. Supponiamo che l'immagine osservata, di 10×10 pixel, sia composta da un quasar che si estende per 4 pixel, con conteggio in ciascun pixel pari a 10, e rumore di fondo uniforme nell'immagine con conteggio per pixel pari ad 1. In questo caso, l'immagine obiettivo è l'immagine che rappresenta unicamente il quasar, senza rumore di fondo. Se un modello ritornasse come stima dell'immagine una contenente unicamente il quasar, ma con conteggi per pixel dimezzati, l'errore quadratico medio sarebbe elevato. Al contrario, se il modello ritornasse invece esattamente l'immagine osservata, l'MSE sarebbe zero. Tuttavia preferiamo il primo scenario, poiché siamo in grado di distinguere il segnale dal rumore di fondo. Per questo motivo utilizziamo l'MSE unicamente per scegliere tra modelli che sono ugualmente performanti sulla base degli altri due indici ι_0 e ι_1. Infine, un ruolo centrale sarà ricoperto dall'interpretazione dei *bicluster* individuati dai vari modelli, come si vedrà nel Capitolo 4, in fase di presentazione dei risultati. Si noti che i due indici presentati sono stati tratti da [6].

Capitolo 4

Applicazioni

4.1 Considerazioni iniziali

4.1.1 Inizializzazione dell'algoritmo di stima del modello SSLB

Il modello *spike and slab lasso biclustering* ([8]) necessita, al fine di utilizzare il suo algoritmo di stima, la specificazione di alcuni iperparametri, quali la griglia di λ_0 e $\tilde{\lambda}_0$ da utilizzare, d che governa la distribuzione a priori IBP, a e b che parametrizzano la distribuzione Beta di θ_{Fk}, $\tilde{\alpha}$ che parametrizza la distribuzione Beta di $\nu_{(k)}$, ed infine gli iperparametri α e δ che definiscono la a priori gamma inversa per gli elementi diagonali di Σ. Si noti che tutti i simboli utilizzati in questo paragrafo sono stati introdotti nel Capitolo 2, paragrafo 2.2.

Al fine di indagare la sensibilità dei risultati al variare di queste impostazioni iniziali, abbiamo stimato più modelli sullo stesso insieme di dati. Si è riscontrato che l'analisi non è assolutamente sensibile ad una modifica della griglia di λ_0 o $\tilde{\lambda}_0$ a patto che quest'ultima sia fermata ad un valore non troppo elevato. Per queste due quantità dunque verranno utilizzati nel seguito i valori di default forniti in [8], ovvero $\lambda_0 \in (1, 5, 10, 50, 100, 500, 10^3, 10^4, 10^5, 10^6, 10^7)$ e $\tilde{\lambda}_0 \in (1, 5, ..., 5)$ con lunghezza pari a quella della sequenza di λ_0. Per lo stesso motivo, per $\tilde{\alpha}$ utilizziamo il valore di default $1/n$, dove n indica il numero di righe della matrice dei dati. Per gli iperparametri a e b i valori di default $a = 1/K^*$ e $b = 1$ non hanno motivo di essere cambiati, in quanto l'analisi non è sensibile ad una loro modifica e l'alternativa $a = 1/p$ e $b = 1/p$, con p numero di colonne della matrice dei dati, non sembra sensata, neanche a priori, dato che non ci aspettiamo sia *bicluster* densi che sparsi. L'iperparametro d, che definisce la distribuzione a priori IBP, viene posto pari a 0 poiché non si è riscontrata empiricamente una sensibilità dei risultati al variare della specificazione di d, né siamo interessati alla specificazione di valori compresi tra $(0, 1)$, che favoreggiano un numero più elevato di *bicluster* con sparsità crescente. L'obiettivo è invece ottenere pochi *bicluster*, uno per il quasar ed un paio per i getti, e tutti di uguale sparsità. Non è possibile, tramite la funzione `SSLB`, dell'omonimo pacchetto di R che implementa il modello ([1]), specificare dei

valori diversi da quelli di default per α e δ che vengono posti pari ad una loro stima ottenuta tramite una strategia Bayesiana empirica. Questa procedura calibra la distribuzione a priori gamma inversa su σ_j^2 verso valori che sono in accordo con la scala osservata dei dati.

Data la natura iterativa dell'algoritmo EM è anche necessario specificare, per ciascun parametro del modello, un valore iniziale. Seguendo quanto suggerito in [8], si procede come segue: ciascuna entrata della matrice dei coefficienti B viene generata da una distribuzione Gaussiana standard, le entrate di Ω vengono impostate pari a 100, implicando una a priori iniziale per X relativamente non informativa, i parametri θ_{Fk}, che rappresentano il livello di sparsità, vengono posti pari a 0.5, i parametri $\boldsymbol{\nu} = (\nu_{(1)}, ..., \nu_{(k)})$ vengono generati indipendentemente da una distribuzione Beta$(1, 1)$ ed infine, per il numero di *bicluster* K^* si usa una sovrastima di 50 che verrà eventualmente aumentata se necessario.

4.1.2 Contestualizzazione dei metodi

Il modello *SparseBC* ([12]) viene applicato utilizzando la funzione `SparseBC` dell'omonima libreria disponibile in *R* ([3]). Si ricorda che, in questo metodo, un *bicluster* (C_g, D_r) viene identificato quando la sua media μ_{gr} è significativamente diversa da 0. Inoltre, il modello assume che le entrate della matrice dei dati osservati y_{ij}, con $i = 1, ..., 64$ e $j = 1, ..., 64$, siano indipendentemente distribuite come una Gaussiana. Tuttavia, nel contesto astrofisico, abbiamo a disposizione conteggi, che sono per definizione maggiori di zero e discreti, non continui; i risultati vanno dunque rimaneggiati per soddisfare questi vincoli. In particolare, si è riscontrato che, in nessuno studio di simulazione è stata stimata una media negativa per un *bicluster*. Tuttavia, spesso troviamo che la grande maggioranza delle medie stimate assume valori in $(0, 0.5]$, che possono essere dunque arrotondate a zero. Seguendo lo stesso principio, vengono arrotondati anche gli altri valori di μ_{gr} al numero intero più vicino e si ottiene così l'immagine ricostruita dal modello.

Il modello SSVD ([6]), basato sulla scomposizione a valori singolari sparsa, viene applicato utilizzando la funzione `biclust` della libreria `s4vd` di `R` ([2]). Come descritto nel paragrafo 2.4.2, ci serviamo della regola del gomito per scegliere il numero di strati da considerare nel comporre l'immagine ricostruita. Si ricorda che gli autovalori, calcolati come il quadrato dei valori singolari s_k, possono essere interpretati come una misura di quanto importanti sono gli strati. Essendo il quasar centrale sempre molto più luminoso del getto di raggi X, e dunque preponderante rispetto ad esso, ci aspettiamo che il primo autovalore sia enormemente più elevato degli altri. Per questo, ad una prima analisi, seguendo pedissequamente la regola del gomito, ci fermeremmo a considerare solo due strati poiché dopo di essi la pendenza della retta che rappresenta gli autovalori al crescere del numero di strati, sembra appiattirsi. Potremmo dunque concludere che i successivi strati siano poco importanti per spiegare i dati osservati, rispetto ai primi due, ed escluderli. Questo effetto è evidente in

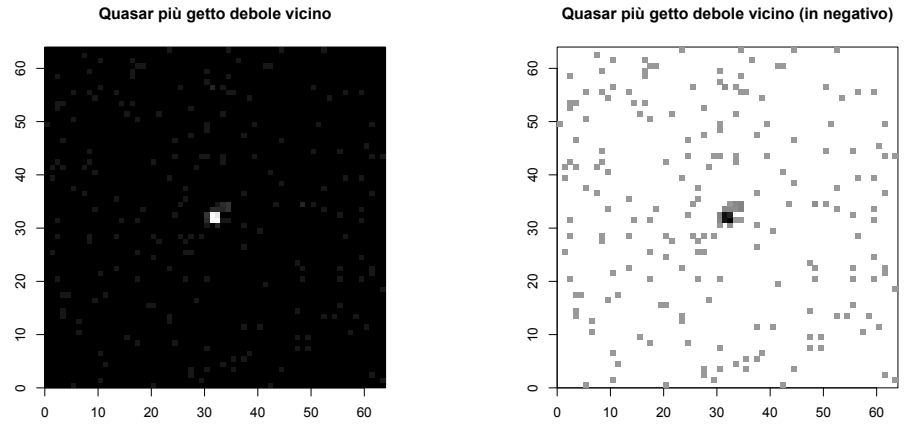

Figura 4.1: *Esempio delle differenze tra immagine nella rappresentazione originale (sinistra) e immagine nella rappresentazione in negativo (destra).*

Figura 4.8 (a sinistra) che rappresenta i primi 20 autovalori risultanti dall'applicazione di SSVD allo studio di simulazione B4. Tuttavia, siamo interessati a distinguere il quasar dal getto di raggi X, e dobbiamo dunque spingerci oltre il primo strato, che con tutta probabilità descriverà il quasar e che influenza pesantemente la nostra decisione. Ingrandendo il grafico e considerando gli autovalori dal secondo strato in poi, si vede distintamente in Figura 4.8 (a destra) che, secondo la regola del gomito, anche il terzo, quarto e quinto strato sembrano essere importanti. Abbiamo inoltre riscontrato che, considerando il grafico ingrandito, fermarsi uno strato prima di quello che suggerisce la regola del gomito garantisce risultati migliori. Procederemo dunque in questo senso per ogni studio di simulazione.

Sia per il modello SSLB ([8]) che per SSVD può capitare che alcune entrate della matrice stimata risultino pari ad un valore leggermente negativo (tra -0.6 e -0.001), poiché nessuno dei due metodi tiene in considerazione la natura di conteggio dei dati. Per questo motivo si decide di porre uguale a 0 il valore delle entrate stimate come negative, interpretando il valore negativo come una sicurezza che il conteggio stimato nel pixel corrispondente sia 0. Infine, per riportarci a valori di conteggio, e quindi discreti e non continui si arrotonda ciascun elemento della matrice stimata all'intero più vicino.

4.1.3 Note di interpretazione delle figure

Nonostante la rappresentazione grafica delle immagini adottata nel Capitolo 3, per la quale ad un pixel di colore nero coincide un conteggio simulato pari a zero in corrispondenza di quel pixel, sia utile per comprendere le luminosità relative tra gli oggetti celesti ed il rumore di fondo, è poco adatta per scorgere i piccoli dettagli che in questo contesto fanno la differenza. Per questo motivo si è deciso di lavorare, in fase di presentazione grafica dei risultati, con quella che chiameremo immagine in negativo. Trasformiamo i conteggi in scala

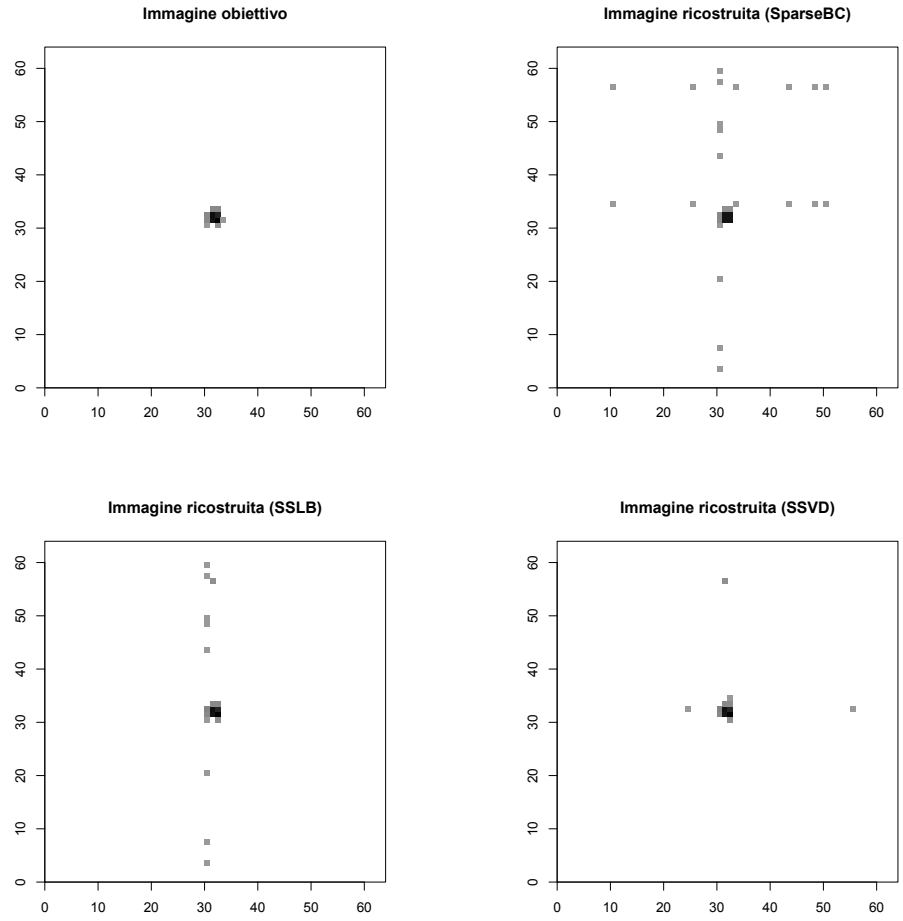

Figura 4.2: *Immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione A1, che considera un'immagine composta solamente da un quasar, senza getto di raggi X, e immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra). Tutte le immagini sono rappresentate in negativo.*

logaritmica e invertiamo il significato dei colori. Il colore bianco rappresenta dunque il valore $-\infty$, ovvero un conteggio simulato pari a 0, mentre il colore nero rappresenta un valore pari al logaritmo del massimo conteggio simulato. A ciascuno studio di simulazione sarà dunque associata una scala diversa, a seconda del valore del massimo conteggio ottenuto. Tutte le gradazioni tra il bianco e il nero indicano una via di mezzo tra i due estremi. Questa trasformazione ha l'effetto di schiacciare le differenze tra conteggi diversi da zero ed evidenziare invece la differenza di questi con quelli uguali a zero che saranno ora pari a $-\infty$. Per apprezzare la differenza tra le due rappresentazioni si può fare riferimento alla Figura 4.1 che riporta l'immagine simulata per lo studio C1 nella rappresentazione classica (sinistra) e in quella in negativo (destra). Per illustrare le immagini ricostruite dai metodi e le composizioni dei *bicluster* trovati useremo dunque la rappresentazione in negativo.

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.995	0.818	30.273	5
SSLB	0.998	0.909	0.182	4
SSVD	0.999	0.818	0.273	2

Tabella 4.1: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione A1.*

4.2 Risultati

In questo paragrafo vengono presentati i risultati più significativi relativi ad alcuni degli studi di simulazione descritti nel Capitolo 3, i restanti risultati sono fruibili in Appendice A. Il campione che abbiamo voluto commentare approfonditamente è rappresentativo della totalità dei risultati ed eventuali similarità e differenze saranno opportunamente evidenziate. Inoltre, in Appendice B viene riportato il codice in linguaggio *R* necessario per la riproduzione dei risultati dello studio di simulazione D1. Per tutte le altre simulazioni, il procedimento è analogo, l'unica modifica da apportare è nel codice per ottenere l'immagine simulata, seguendo quanto visto nel Capitolo 3 sugli studi di simulazione.

4.2.1 Quasar senza getto (A1)

I 3 modelli per il *biclustering*, presentati nel Capitolo 2 vengono in questo paragrafo applicati al primo studio di simulazione (A1). L'immagine che si ipotizza essere osservata non presenta alcun getto di raggi X ma solamente un quasar e rumore di fondo uniforme, ed è disponibile in Figura 3.1 (a sinistra). In questo caso sappiamo che non c'è un getto di raggi X fuoriuscente dal quasar centrale; siamo infatti interessati a verificare che, in assenza di segnale oltre il quasar, i metodi non aggiungano segnale che in realtà non è presente.

Per il modello *sparseBC* ([12]), applicando la procedura di selezione di (G, R, λ) discussa nel paragrafo 2.3.1 si ottiene $(\hat{G}, \hat{R}, \hat{\lambda}) = (7, 7, 1)$ ovvero un totale di 49 *bicluster* identificati. In questo caso 44 su 49 assumono un valore compreso tra $[0, 0.5)$ che può dunque essere approssimato a 0, ottenendo così 5 *bicluster* individuati. Seguendo lo stesso principio, vengono arrotondati anche gli altri valori di μ_{gr} al numero intero più vicino e si ottiene così la matrice stimata dal modello, rappresentata in negativo in Figura 4.2 (in alto a destra).
Per il modello SSVD ([6]), osservando il grafico ingrandito del valore degli autovalori al variare del numero di strati (Figura 4.3, a destra), si conclude che il numero di strati adatto per il primo studio di simulazione è 2. Questo sembra essere in contrasto con l'immagine obiettivo, composta solo dal quasar, ma si vedrà in fase di interpretazione dei *bicluster* ottenuti, che entrambi gli strati sono associati al quasar in questo caso.
Applichiamo il metodo *spike and slab lasso biclustering* ([8]) ai dati simulati

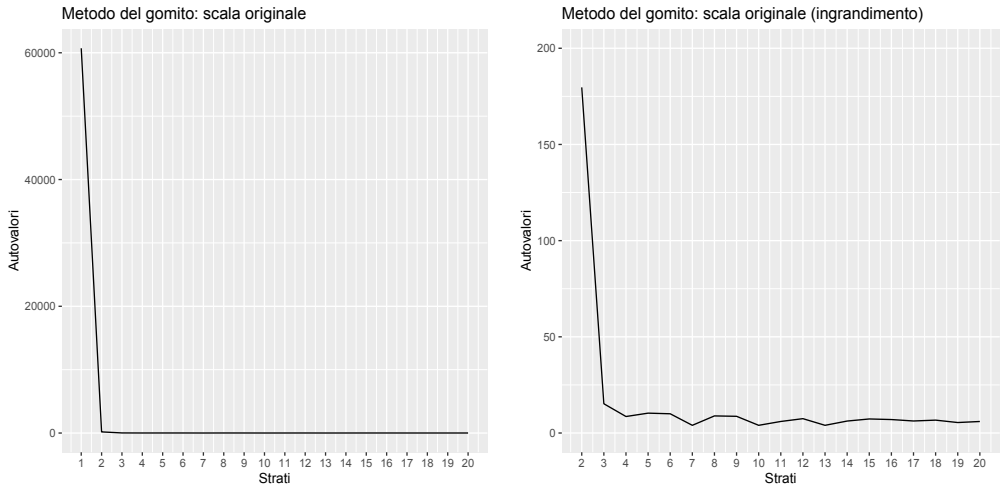

Figura 4.3: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione A1. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

tramite la funzione `SSLB` dell'omonimo pacchetto di `R`. Come descritto nel paragrafo 2.2 scegliamo tra 100 lanci dell'algoritmo di stima quello che minimizza un'approssimazione, in ambito Bayesiano, del BIC. Questa strategia porta ad un numero di fattori o *bicluster* trovati pari a 4.

Innanzitutto, valutiamo la capacità dei metodi di ricostruire il segnale presente nell'immagine osservata, ovvero di schiacciare verso zero il conteggio del numero di fotoni rilevato nei pixel che sono associati al rumore di fondo. Esploriamo quindi, da un punto di vista visivo, le immagini ricostruite dai 3 modelli, in Figura 4.2 (in negativo) e li valutiamo tramite gli indici presentati nel paragrafo 3.6. L'immagine ideale, che vorremmo i nostri metodi ritornassero è quella che presenta solamente il quasar centrale, senza rumore di fondo. Se così fosse, avremmo raggiunto l'obiettivo di estrarre segnale dall'immagine osservata senza includere nella stima il rumore di fondo. L'immagine obiettivo è quindi formata unicamente dai conteggi simulati come provenienti dal quasar e può essere apprezzata in Figura 4.2 (in alto a sinistra, in negativo). Come si può dedurre dalla Tabella 4.1 si preferisce in questo caso il modello basato sulla scomposizione a valori singolari della matrice di dati osservata (SSVD, [6]), poiché la percentuale di zeri correttamente stimata è maggiore rispetto agli altri modelli. Questa decisione, basata sugli indici, viene confermata anche da un punto di vista visivo. Nonostante vi sia ancora del rumore residuo nell'immagine ricostruita da SSVD (Figura 4.2, in basso a destra), è l'immagine che più assomiglia a quella obiettivo (in alto a sinistra).

Un aspetto fondamentale da valutare nell'applicazione di questi modelli nel contesto astrofisico di identificazione di getti di raggi X è l'interpretazione dei *bicluster*. Ricordiamo che in questa tipologia di analisi vogliamo identificare zone della matrice dei dati, e quindi dell'immagine, che contengono segnale. Inoltre, vorremmo che il metodo distingua una zona che contiene il

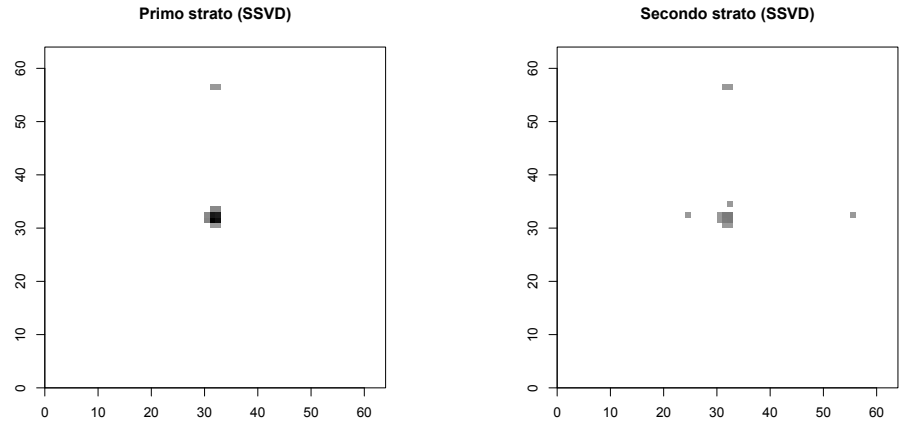

Figura 4.4: *Rappresentazione in negativo della composizione degli strati o bicluster stimati in SSVD per lo studio di simulazione A1, primo strato a sinistra e secondo strato a destra. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

segnale proveniente dal quasar da quelle che invece contengono il segnale degli eventuali getti di raggi X. Le zone contenenti segnale sono rappresentate dai *bicluster* trovati dal modello, che sono infatti insiemi di righe e di colonne della matrice dei dati. Vorremmo ottenere un *bicluster* composto da pixel che corrispondono alla zona dell'immagine dove è presente il quasar e, se presenti getti di raggi X, tanti *bicluster* quanti sono questi getti, che identifichino le zone dell'immagine dove sono stati osservati. In Figura 4.4 si possono osservare i due strati, o *bicluster*, risultanti dall'applicazione di SSVD. L'interpretazione è immediata, per il primo strato (Figura 4.4, a sinistra): per esempio, gli elementi non bianchi dell'immagine sono i pixel in corrispondenza dei quali è stato stimato un valore per il primo strato maggiore di zero in valore assoluto. È possibile infatti che alcuni valori delle entrate appartenenti al primo strato o *bicluster*, vengano stimati anche sensibilmente minori di zero, anche -5 o -10. Facciamo notare, tuttavia, che questa caratteristica, una volta sommati tutti gli stati tra loro per ottenere la matrice stimata dal modello, non porta a valori delle entrate della matrice stimata sensibilmente minori di zero. Come visto precedentemente nel paragrafo 4.1.2, i valori negativi, se presenti, sono nell'intervallo $(-0.6, -0.001)$ e vengono posti pari a zero. Tuttavia, in fase di interpretazione dei *bicluster*, un'entrata appartenente ad uno strato con valore negativo, rappresenta comunque una partecipazione di essa alla composizione dello strato stesso e va dunque considerata. Per questo motivo viene utilizzato il valore assoluto per la rappresentazione grafica delle composizioni dei *bicluster*. Gli elementi non bianchi in Figura 4.4 (a sinistra) definiscono dunque quei pixel che appartengono al primo *bicluster*. Più il colore è verso il nero, più il conteggio corrispondente al pixel è elevato nel primo strato e dunque il pixel è maggiormente presente in esso. Un ragionamento analogo può essere fatto per

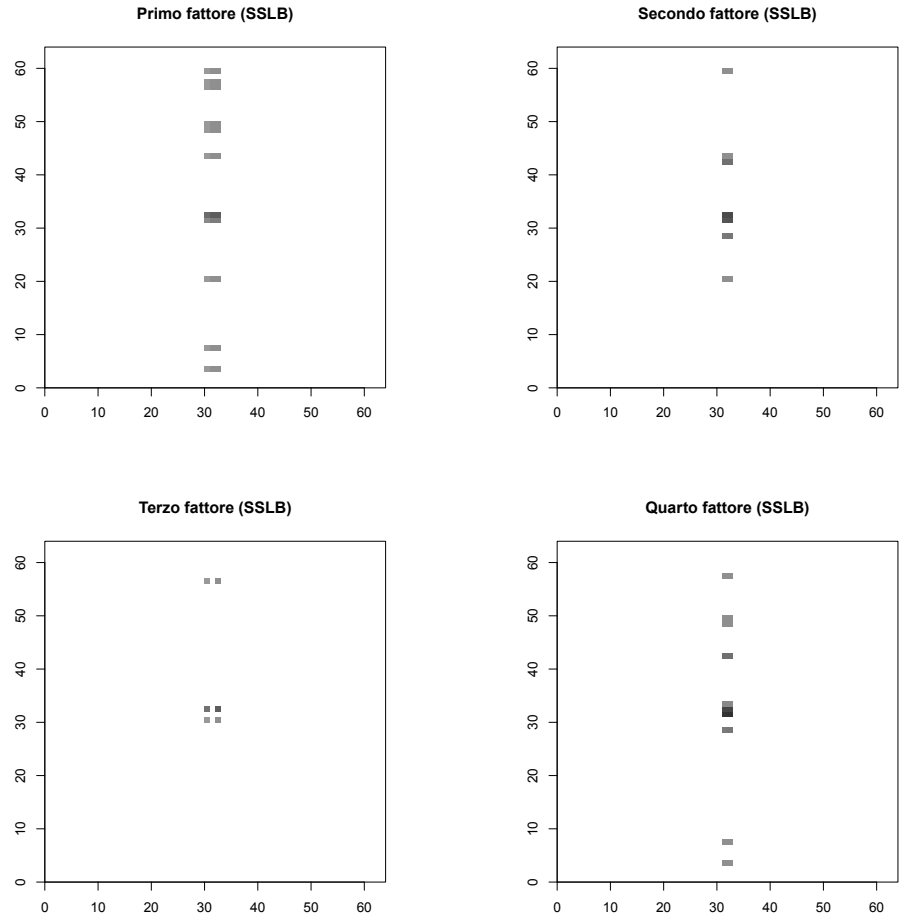

Figura 4.5: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione A1, dal primo fattore al quarto. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

il secondo *bicluster*. Come si vede chiaramente in Figura 4.4, a meno di leggeri rimasugli di rumore di fondo, entrambi gli strati identificano la regione corrispondente al quasar e, come anticipato, il segnale proveniente dal primo strato è nettamente più forte, data la sua colorazione praticamente nera. Anche senza sapere che la zona centrale dell'immagine corrisponde al quasar, si sarebbe in ogni caso concluso che le zone individuate dai due *bicluster* identificano lo stesso oggetto, data la loro sovrapposizione quasi perfetta. Inoltre, dato che la sorgente di segnale più forte in questa tipologia di immagini astrofisiche è sempre il quasar, si sarebbe concluso che vi è un'unica sorgente nell'immagine osservata ed essa corrisponde al quasar.

Un'interpretazione tanto limpida non è disponibile per gli altri due metodi. Per SSLB ([8]), per esempio, si hanno 4 fattori, o *bicluster*, al termine della stima e selezione del modello. Tuttavia, al contrario di quello che accade per

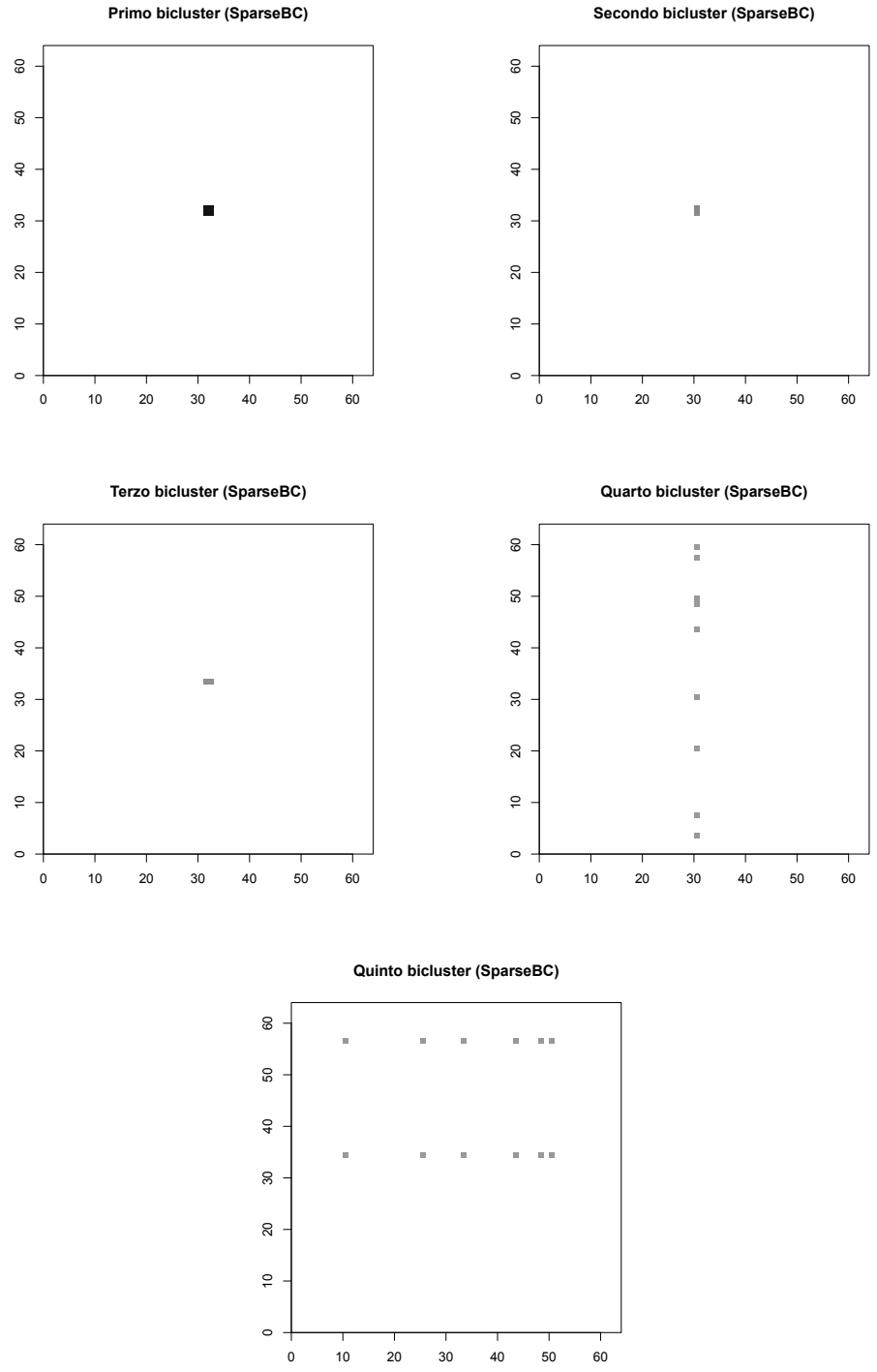

Figura 4.6: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione A1, dal primo al quinto. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

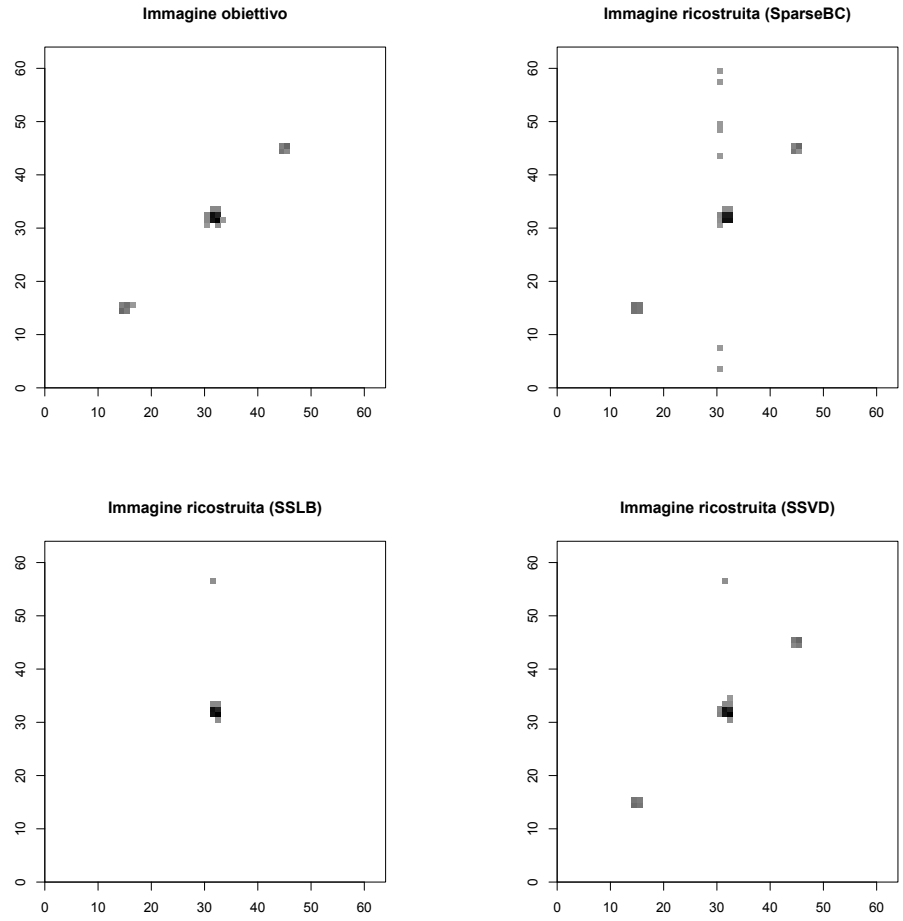

Figura 4.7: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione B4, che considera un'immagine composta da un quasar e due getti di raggi X lontani da esso di luminosità forte, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

SSVD non rappresentano tutti e 4 il quasar. Vediamo infatti, in Figura 4.5, che tutti e 4 i *bicluster* sono composti sia da pixel associati al quasar sia da pixel dove è presente solo rumore di fondo e non segnale. Nonostante nell'immagine ricostruita (Figura 4.2, in basso a sinistra), i valori in corrispondenza di questi pixel si bilancino, nel senso che quando si sommano i 4 fattori, valori positivi e negativi sommano a zero o quasi, la presenza importante di pixel associati al rumore di fondo in ciascun *bicluster* non permette di ottenere un'interpretazione soddisfacente.

Per interpretare i *bicluster* stimati dal metodo SparseBC ([12]), per ciascuno di essi viene raffigurata la consueta immagine in negativo, con tutti i pixel bianchi e dunque conteggi pari a zero, tranne per quei pixel che appartengono al *bicluster*, a cui viene assegnato un conteggio pari alla media del *bicluster*, i valori vengono successivamente trasformati in scala logaritmica per ottenere

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.998	0.850	14.650	12
SSLB	1.000	0.35	43.150	5
SSVD	1.000	0.850	5.400	4

Tabella 4.2: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione B4.*

la rappresentazione in negativo. Consideriamo come primo *bicluster*, ovvero quello che esprime più segnale, quello a cui è associata una media maggiore, sembra intuitivo procedere in questo senso dato che nel contesto astrofisico i dati sono conteggi del numero di fotoni rilevati. I risultati sono visionabili in Figura 4.6, dove al primo *bicluster* è associata una media pari a 123, al secondo, 3, al terzo 2 e al quarto e al quinto 1. Supponendo di non conoscere l'immagine che si vuol stimare, come sarebbe in un'applicazione del metodo ad insiemi di dati reali, si identificherebbe probabilmente il quasar, composto dai pixel appartenenti al primo *bicluster* e riconoscibile grazie ai conteggi molto elevati in corrispondenza di questi pixel rispetto a tutti gli altri. Tuttavia non saremmo in grado di concludere se gli altri gruppi siano o meno associati ad un getto di raggi X, considerando che sono stati individuati come zone differenti dal resto dell'immagine, ma la loro media è di poco maggiore di zero.

Si conclude che se l'immagine simulata presenta solo un quasar e rumore di fondo senza alcun getto di raggi X, il modello SSVD ([6]) è il migliore, sia in termini di ricostruzione dell'immagine che di interpretabilità dei *bicluster* individuati. Risultati analoghi vengono ottenuti dall'applicazione dei 3 metodi all'immagine simulata per lo studio di simulazione A2, uguale a questo, ma con quasar più diffuso (Figura 3.1, a destra). L'unica differenza è che in questo caso tutti e 3 i fattori o *bicluster* individuati da SSLB identificano correttamente il quasar. Inoltre tutti i metodi sembrano ricostruire adeguatamente l'immagine. I grafici che descrivono i risultati possono essere visionati in Appendice A.1.

4.2.2 Due getti forti lontani dal quasar (B4)

In questo studio di simulazione viene presa in considerazione un'immagine composta da un quasar centrale, due getti di raggi X lontani da esso di luminosità forte e rumore di fondo uniforme (Figura 3.2, in baso a destra). A priori ci aspettiamo che i metodi per il *biclustering* riescano ad individuare le tre sorgenti presenti nell'immagine, poiché la lontananza dei getti dal quasar ne dovrebbe facilitare l'identificazione. Intuitivamente, il segnale proveniente dal quasar non dovrebbe mascherare quello dei getti se sono distanti gli uni dall'altro.

Confrontando le immagini ricostruite dai tre metodi in Figura 4.7 con

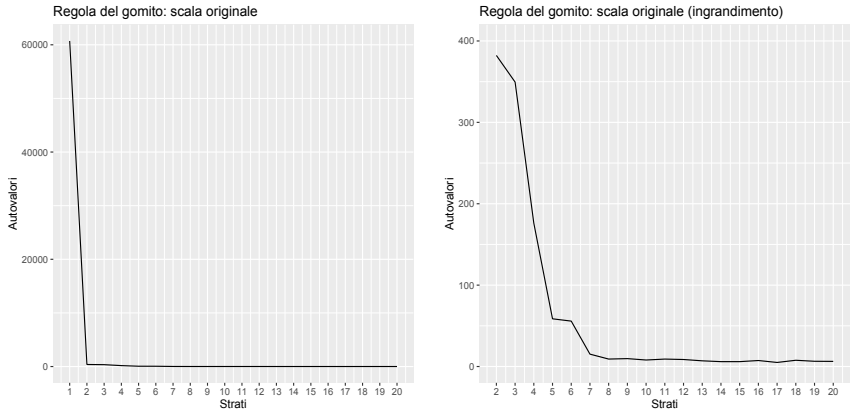

Figura 4.8: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione B4. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

l'immagine obiettivo, sempre in Figura 4.7 (in alto a sinistra), composta dai conteggi associati al quasar e ai due getti di raggi X senza rumore di fondo, notiamo che sparseBC e SSVD hanno correttamente ricostruito le zone contenenti segnale, tuttavia, il modello SSVD è ancora una volta quello che restituisce l'immagine con meno rumore di fondo residuo. SSLB invece non individua i due getti di raggi X. La prima impressione visiva, è confermata dagli indici in Tabella 4.2: SSVD fa registrare un $\iota_0 = 1$, ovvero stima un conteggio pari a zero per tutti i pixel in corrispondenza dei quali abbiamo un conteggio pari a zero nell'immagine obiettivo. Inoltre, SSVD riporta il minor numero di *bicluster* stimati (4), segno che il metodo combina parsimonia e ottimi risultati. Anche SparseBC ricostruisce abbastanza fedelmente l'immagine ma utilizzando una quantità ben più ampia di *bicluster*, 12. Tendenzialmente, più il numero di *bicluster* è alto, più difficoltosa sarà l'interpretazione degli stessi.

La composizione degli strati, o *bicluster*, stimati dal modello SSVD ([6]) possono essere visionati in Figura 4.9. L'interpretazione è immediata, il primo strato, contenendo nettamente maggior segnale degli altri tre (pixel di colore più vicino al nero), individua il quasar. Di conseguenza possiamo concludere che anche il quarto strato, identificando circa la stessa zona dell'immagine, individua il quasar. Il secondo *bicluster* sicuramente contiene il segnale proveniente da un oggetto diverso dal primo, dato che i pixel colorati non bianchi sono in zone differenti. Il medesimo discorso può essere fatto per il terzo strato. Anche non conoscendo l'immagine obiettivo si potrebbe concludere che nelle zone identificate dal secondo e terzo strato, il segnale potrebbe provenire da un getto di raggi X. La facile interpretabilità dei *bicluster* di questo metodo è una qualità che abbiamo osservato essere trasversale a tutti gli studi di simulazione considerati. Non sempre SSVD è il metodo che ricostruisce meglio l'immagine, tuttavia, è sempre quello che meglio si presta a trarre delle conclusioni sulla natura degli oggetti osservati, poiché dispone di un ordinamento

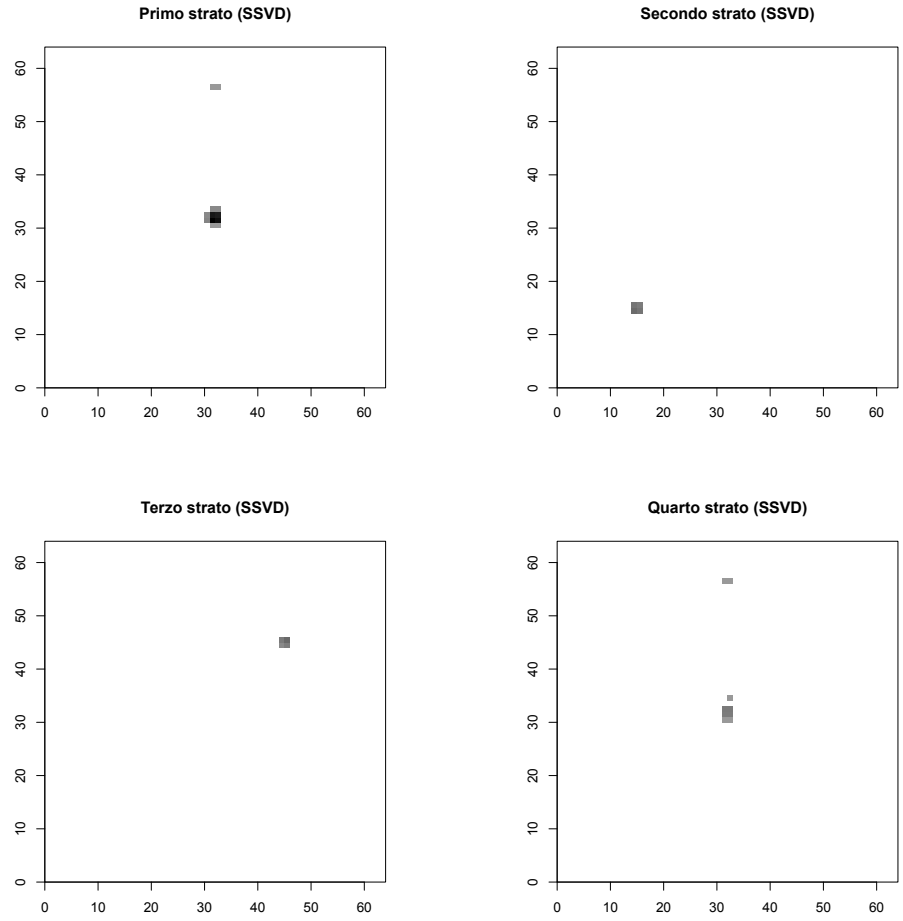

Figura 4.9: *Rappresentazione in negativo della composizione degli strati o bicluster stimati in SSVD per lo studio di simulazione B4, primo strato in alto a sinistra, secondo in alto a destra, terzo in basso a sinistra e quarto strato in basso a destra. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

per importanza degli strati, dato dal valore degli autovalori ad essi associati e permette ai *bicluster* di sovrapporsi, aiutandoci a comprendere quando due strati identificano o meno lo stesso oggetto. Si noti che in Figura 4.8 si può osservare il grafico sulla base del quale, tramite la regola del gomito, è stato possibile scegliere il numero di strati da considerare.

Per il modello sparseBC ([12]), possono essere apprezzate le composizioni dei *bicluster* in Figura 4.10. Conoscendo l'immagine obiettivo che vorremmo fosse stimata, è facile interpretare i *bicluster*. Il primo e il secondo, così come dall'ottavo al decimo, contengono conteggi di fotoni provenienti dal quasar, le zone di colore diverso dal bianco identificate nelle immagini sono infatti quelle centrali.

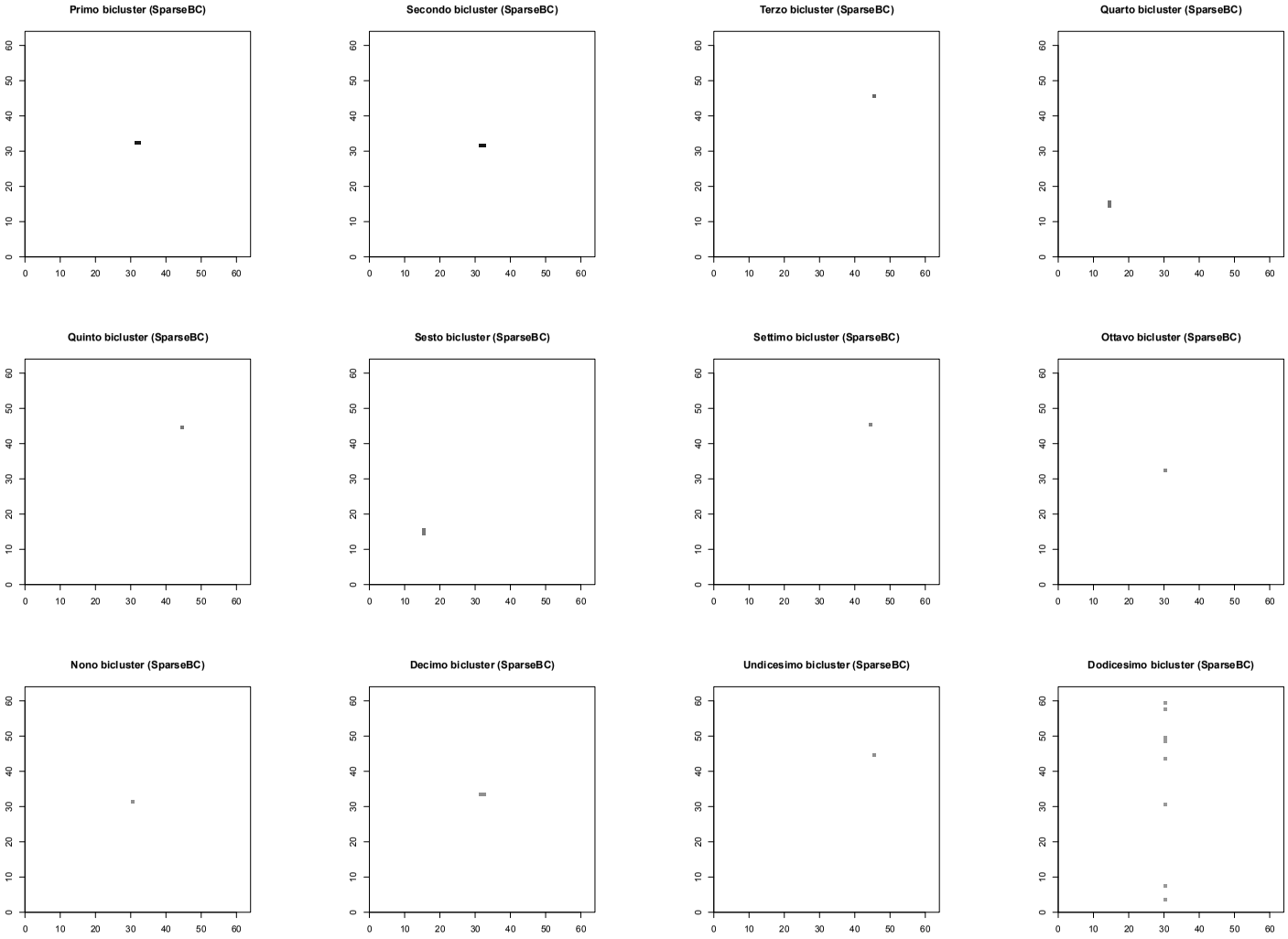

Figura 4.10: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione B4, dal primo al dodicesimo. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

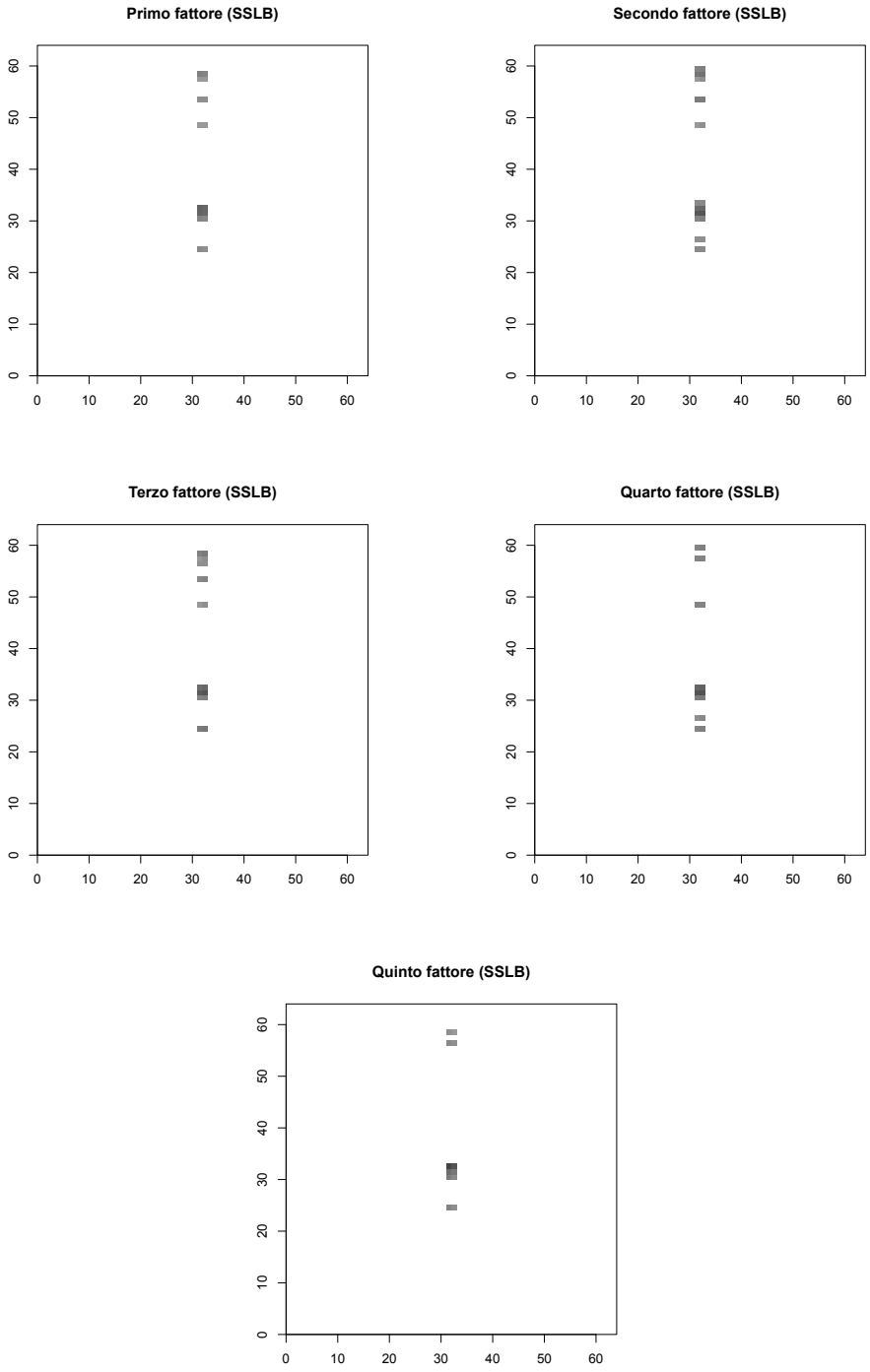

Figura 4.11: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione B4, dal primo fattore al quinto. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	1.000	0.476	10.857	6
SSLB	1.000	0.333	2.190	2
SSVD	0.999	0.476	1.286	2

Tabella 4.3: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione D1.*

Il terzo, quinto, settimo ed undicesimo, identificano il getto di raggi X in alto a destra, mentre quello in basso a sinistra è rappresentato dal quarto e sesto *bicluster*. Infine, il dodicesimo è formato da conteggi attribuibili al rumore di fondo. Tuttavia, questa interpretazione è possibile solo se si conosce l'immagine obiettivo, che, in applicazioni a dati reali chiaramente non è disponibile. Il problema di questo metodo, riscontrato in praticamente tutti gli studi di simulazione, è che trova *bicluster* troppo piccoli, cioè formati da troppi pochi pixel, spesso anche solo uno, come si può notare in tutte le figure che rappresentano le composizioni dei *bicluster* stimati da questo metodo. La conseguenza è la stima di un numero molto elevato di *bicluster* composti da pochi pixel. Inoltre, non potendo essi sovrapporsi in questo modello, non è possibile capire se due *bicluster* che identificano zone vicine, come il primo e il secondo in Figura 4.10, individuino lo stesso oggetto celeste.

Per il modello SSLB ([8]), le composizioni dei fattori, o *bicluster*, stimati sono presentate in Figura 4.11. Come visto anche nella presentazione dei risultati per lo studio di simulazione A1, i *bicluster* stimati da SSLB sono sempre composti da pixel in corrispondenza dei quali sono stati simulati conteggi associati al rumore di fondo. Quindi, oltre a non individuare i getti di raggi X, i *bicluster* stimati non forniscono l'interpretazione che vorremmo, in particolare, non c'è un *bicluster* che individua solo la zona occupata dal quasar e che dunque identifica un oggetto celeste specifico. Risultati di questo tipo sono poco utili in un contesto reale dove non conosciamo l'immagine obiettivo. Il problema del modello è che impiega troppi fattori per eliminare il rumore di fondo e quindi se lo trascina in ognuno dei *bicluster* che stima.

Anche in questo contesto, SSVD ([6]) si conferma dunque il modello più interpretabile, senza sacrificare precisione nella ricostruzione dell'immagine. Quanto emerso in questo paragrafo è confermato anche dagli altri studi di simulazione appartenenti al contesto B (getti lontani dal quasar) e dagli studi D2 e D3 del contesto D (quasar con getto esteso). Per visualizzare i risultati dell'applicazione dei tre metodi per il *biclustering* agli studi sopra citati, si rimanda all'Appendice A.2, A.3, A.4, A.7 e A.8. Due differenze degne di nota sono: nello studio B2, SSLB ricostruisce adeguatamente l'immagine, eliminando il rumore di fondo ed identificando correttamente il getto di raggi X (Figura A.11, in alto a sinistra); nello studio D3, sparseBC ricostruisce quasi

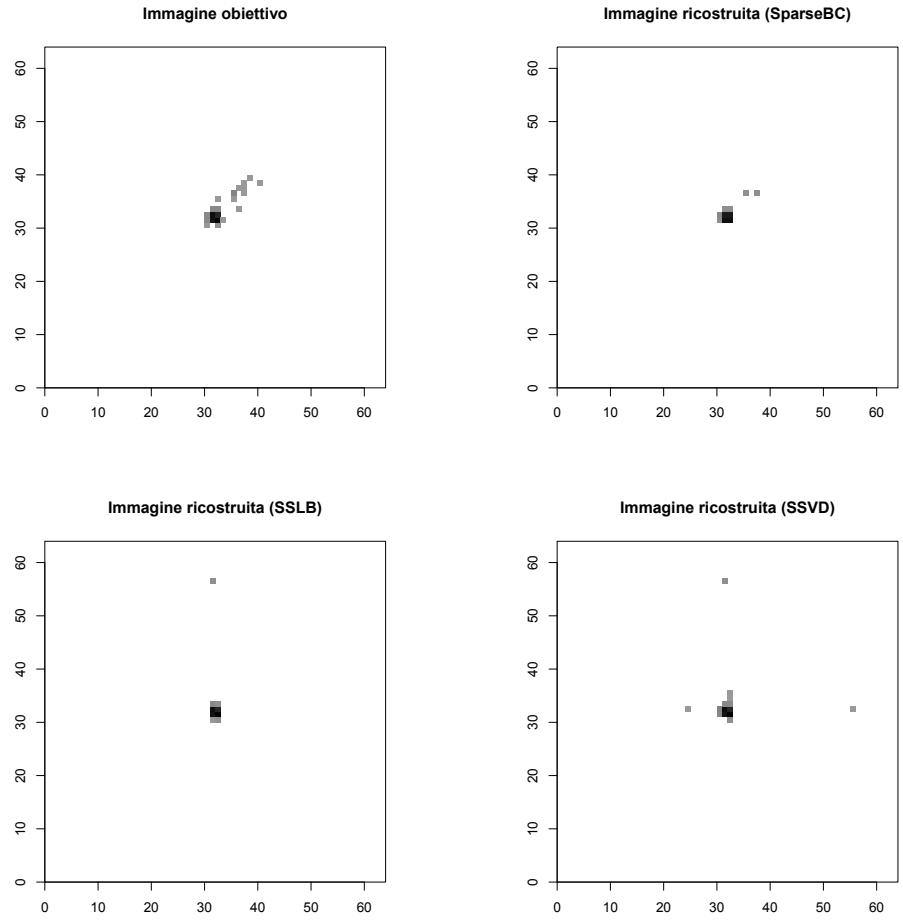

Figura 4.12: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione D1, che considera un'immagine composta da un quasar e un getto di raggi X esteso di luminosità debole, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

perfettamente l'immagine (Figura A.35, in alto a destra). Di seguito verranno invece considerati i risultati per lo studio di simulazione D1, al fine di presentare un'applicazione dei metodi ad un'immagine composta da un quasar e un getto debole.

4.2.3 Quasar con getto debole ed esteso (D1)

In questo paragrafo viene preso in considerazione uno studio di simulazione particolarmente complesso. Siamo nel contesto D1: l'immagine simulata, in Figura 3.4 (in alto a sinistra), è composta dal quasar centrale, da un getto di raggi X debole e diffuso, abbastanza vicino ad esso e rumore di fondo uniforme.

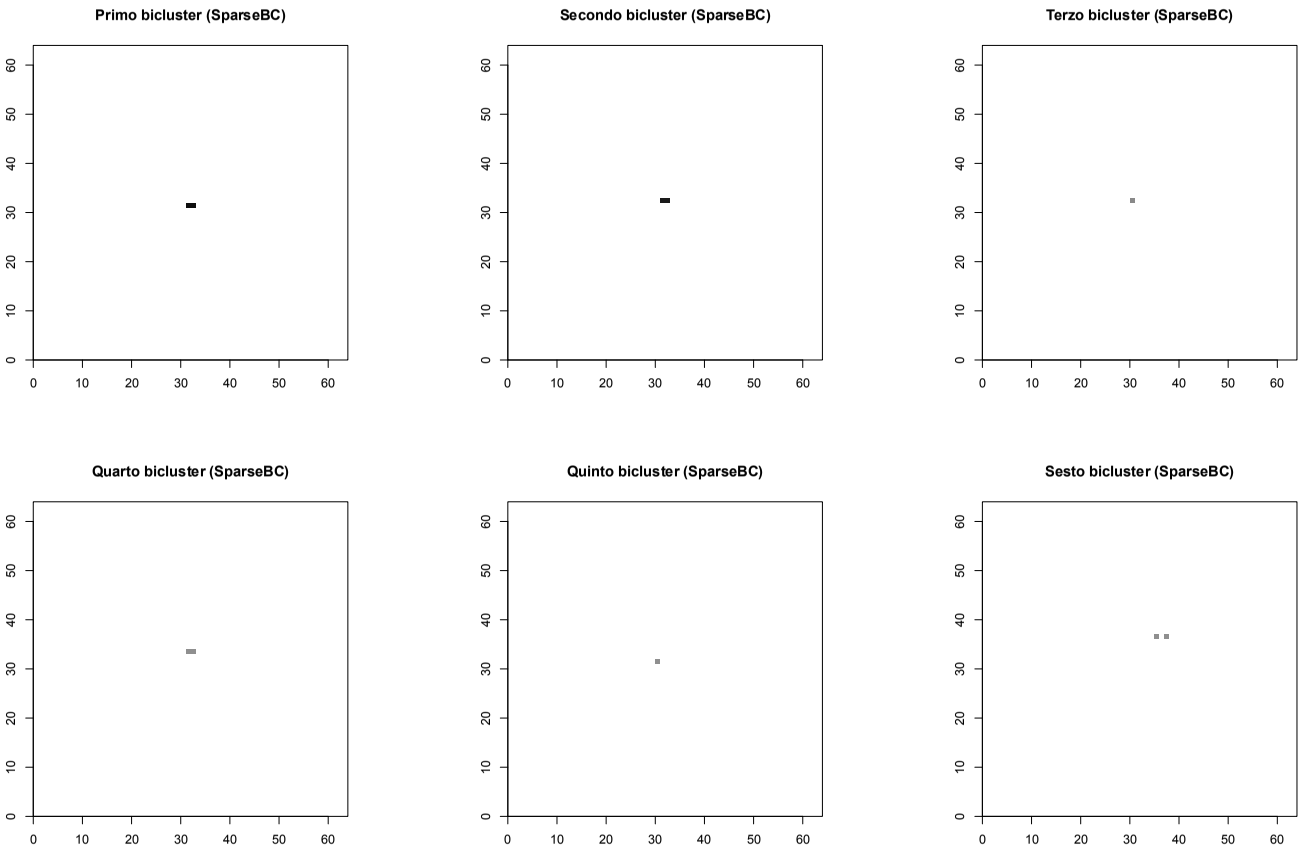

Figura 4.13: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione D1, dal primo al sesto. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

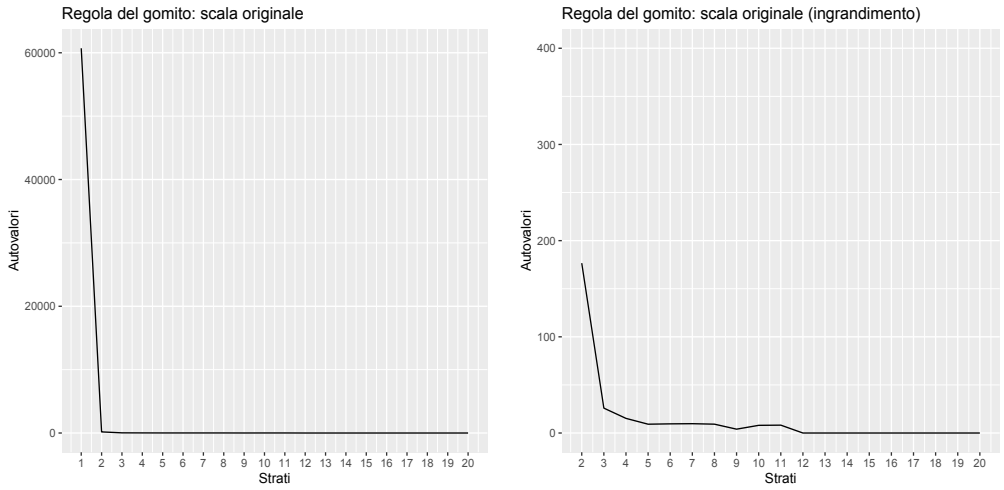

Figura 4.14: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione D1. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

La complessità è data dalla forma diffusa e dalla natura debole del getto a cui è associato un numero tanto esiguo di conteggi da confondersi con il rumore di fondo. Inoltre, essendo vicino al quasar, ci aspettiamo che il segnale proveniente dal getto venga mascherato da quello più preponderante del quasar stesso.

Se confrontiamo le immagini ricostruite dai tre metodi, sparseBC, SSLB e SSVD con l'immagine obiettivo (Figura 4.12), si nota che nessuno dei tre metodi è riuscito ad individuare nettamente il getto di raggi X presente nell'immagine. L'unico modello che, seppur debolmente, individua un qualche tipo di segnale nella zona dove dovrebbe essere il getto di raggi X è sparseBC ([12], Figura 4.12, in alto a destra). Coerentemente, in Tabella 4.3, sparseBC è il modello che meglio ricostruisce l'immagine sia in termini di ι_0 (prima colonna) che in termini di ι_{not} (seconda colonna). Guardando solo la tabella si potrebbe argomentare che SSVD ([6]) riesce a ricostruire praticamente ugualmente bene l'immagine simulata; questo è solo parzialmente vero. Intuitivamente, dall'immagine ricostruita da sparseBC, si evince che potrebbero esserci due sorgenti differenti. SSVD, infatti, stima correttamente la parte di getto attaccata al quasar, e dunque osservando l'immagine ricostruita nel suo insieme si concluderebbe che vi è un'unica sorgente. Al contrario, un dubbio sulla presenza di due sorgenti differenti rimarrebbe guardando l'immagine ricostruita da sparseBC. Il modello SSLB ricostruisce parzialmente l'immagine, identificando solamente il quasar centrale. Per questo motivo evitiamo di seguito di discutere l'interpretazione dei *bicluster* da esso stimati.

Analizzando le immagini raffiguranti le composizioni dei 6 *bicluster* stimati da sparseBC in Figura 4.13 ci accorgiamo subito che il problema di interpretazione riscontrato nella presentazione dei risultati dello studio di simulazione B4, nel paragrafo precedente, persiste. I *bicluster* stimati sono formati da po-

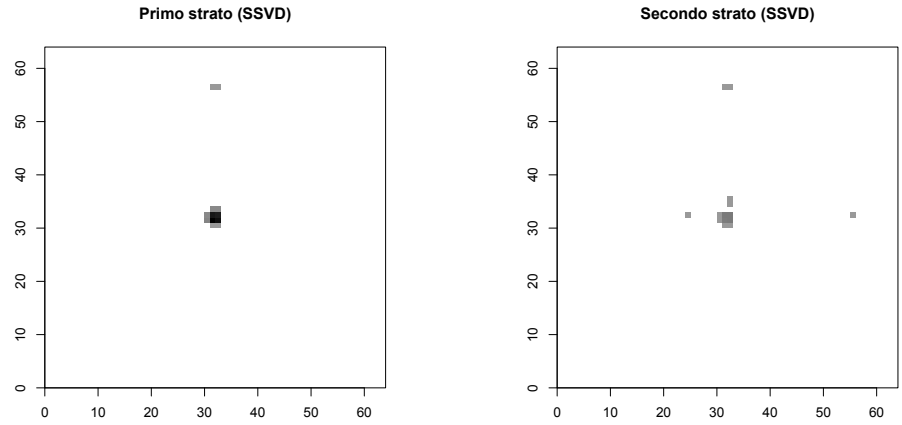

Figura 4.15: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione D1, primo strato in alto a sinistra, secondo in alto a destra, terzo in basso a sinistra e quarto strato in basso a destra. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

chi pixel e non siamo in grado di determinare se diversi *bicluster* individuino la stessa sorgente o sorgenti differenti.

Le composizioni degli strati, o *bicluster* stimati dal modello SSVD ([6]), in Figura 4.15 non sono facilmente interpretabili, al contrario di quanto visto per questo modello fino ad ora. Se si può concludere con una certa sicurezza che il primo strato identifica il quasar, dato il colore nero dei pixel che lo compongono, il secondo strato presenta ancora del rumore di fondo residuo che complica la sua interpretazione. Probabilmente la zona centrale, identificata anche dal primo strato, è associata allo stesso oggetto celeste, ovvero il quasar. Le restanti zone nere che compongono il secondo *bicluster* tuttavia, senza conoscere l'immagine obiettivo, potrebbero essere erroneamente scambiate per getti di raggi X. Inoltre, il vero getto di raggi X non è stato identificato da alcuno strato. I grafici sulla base dei quali applicando la regola del gomito abbiamo deciso di considerare 2 strati vengono proposti in Figura 4.14.

Si conclude che i modelli per il *biclustering* falliscono in questo contesto, sia nella ricostruzione dell'immagine, sia nel fornire una distinzione, tramite l'interpretazione dei *bicluster* da essi stimati, tra segnale associato al quasar e segnale associato al getto di raggi X, che in questo caso non viene proprio identificato. In verità non ci sorprende questa conclusione dato l'elevato grado di complessità dell'analisi dettato dalla struttura dell'immagine simulata di partenza.

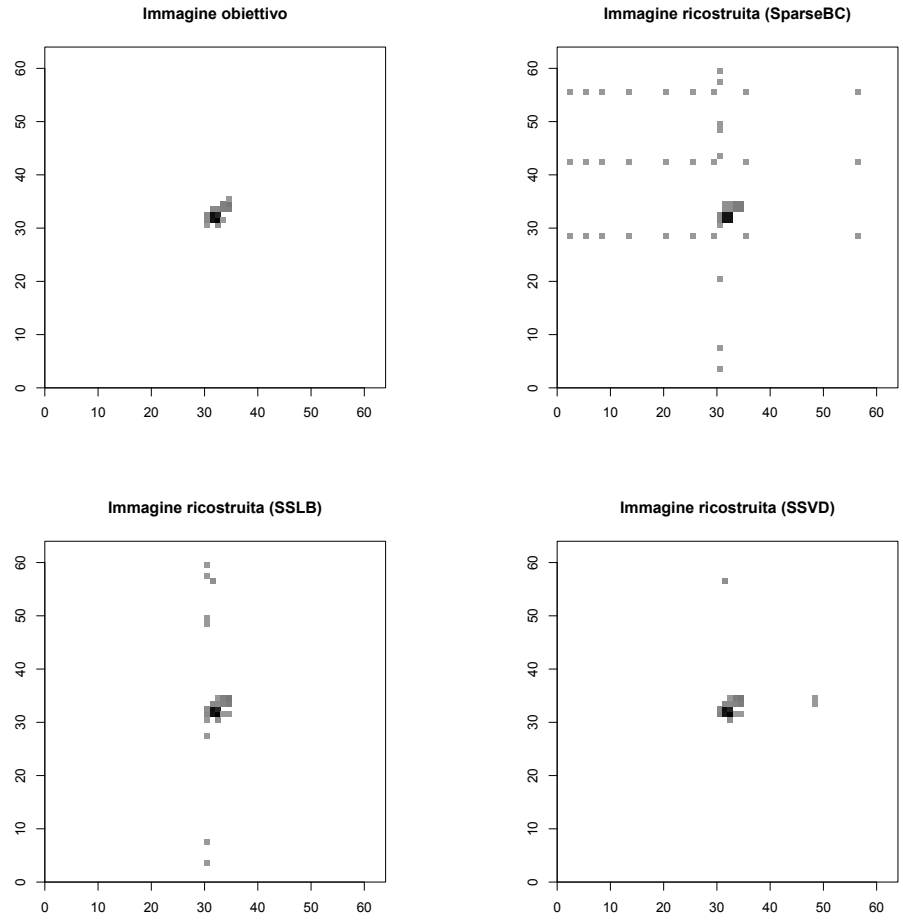

Figura 4.16: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione C2, che considera un'immagine composta da un quasar e da un getto di raggi X contiguo ad esso e di luminosità media, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

4.2.4 Quasar con getto medio contiguo (C2)

Nello studio di simulazione C2 viene considerata un'immagine composta da un quasar, un getto di raggi X di luminosità media contiguo al quasar e rumore di fondo uniforme, l'immagine è disponibile in Figura 3.3 (in alto a destra).

Osserviamo innanzitutto le immagini ricostruite dai tre metodi per il *biclustering* in Figura 4.16. Ancora una volta, l'immagine ricostruita che più assomiglia a quella obiettivo composta dal quasar e il getto di raggi X, e che contiene dunque meno rumore di fondo residuo è quella stimata dal modello SSVD ([6]). La conferma della superiorità del metodo nella ricostruzione dell'immagine proviene dalla Tabella 4.4, che riporta gli indici di bontà di ricostruzione. Per SSVD abbiamo $\iota_0 = 0.999$, $\iota_{not} = 0.875$. Gli altri due modelli forniscono comunque dei risultati buoni a livello di ricostruzione dell'immagi-

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.991	0.750	50.125	6
SSLB	0.998	0.938	0.25	6
SSVD	0.999	0.875	61.625	3

Tabella 4.4: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione C2.*

ne, con SSLB in particolare che ha un $\iota_0 = 0.998$, ovvero rimuove gran parte del rumore di fondo, ed un $\iota_{not} = 0.938$ più elevato di tutti gli altri metodi. In questo caso molto al limite, i modelli SSLB e SSVD possono essere considerati ugualmente performanti, dato anche l'MSE prossimo a zero di SSLB. Infine, come si deduce dall'immagine in Figura 4.16 (in alto a destra), sparseBC include una grande quantità di rumore di fondo nell'immagine ricostruita.

Le composizioni dei 3 strati o *bicluster* stimati dal modello SSVD ([6]) vengono presentate in Figura 4.18, mentre i grafici utilizzati per la scelta del numero di strati da considerare tramite la regola del gomito sono illustrati in Figura 4.17. In questo caso l'interpretazione non è immediata. Il primo strato identifica i pixel associati al quasar mentre il secondo e il terzo identificano entrambi in parte il quasar ed in parte il getto di raggi X. Come concludere in questo caso se l'immagine raffigura due sorgenti diverse o una unica? In realtà non c'è una risposta secca a questa domanda. Dato che il primo *bicluster* identifica chiaramente il quasar e che questo è in generale molto più luminoso delle sorgenti che vengono rilevate assieme ad esso, si potrebbe concludere che se le zone contenti segnale individuate dal secondo e terzo strato fossero riconducibili anch'esse al quasar, il loro segnale sarebbe stato tanto forte da essere individuato già nel primo strato e che dunque nonostante vi sia della sovrapposizione, il secondo e terzo strato identificano una sorgente differente dal primo. Si noti che questo discorso è sensato quando si considerano due *bicluster* che sono solo parzialmente sovrapposti. Infatti se sono invece composti esattamente dagli stessi pixel la sorgente che individuano è la stessa, come visto nei precedenti paragrafi. Questo sembra essere coerente con quanto osservato nel corso del Capitolo per il modello SSVD. Il primo strato identifica sempre il quasar e possiede un livello di importanza molto elevato, dato dal valore dell'autovalore ad esso associato ampiamente più grande degli altri.

Purtroppo ancora una volta i fattori, o *bicluster* stimati da SSLB [8]), la cui composizione è in Figura 4.19, non forniscono un'interpretazione soddisfacente. Solo il secondo *bicluster* identifica nettamente un oggetto celeste, il getto di raggi X, mentre tutti gli altri contengono anche pixel associati al rumore di fondo. Dunque in un'applicazione reale, dove non si conosce l'immagine obiettivo, probabilmente non si concluderebbe che il secondo fattore individua il getto.

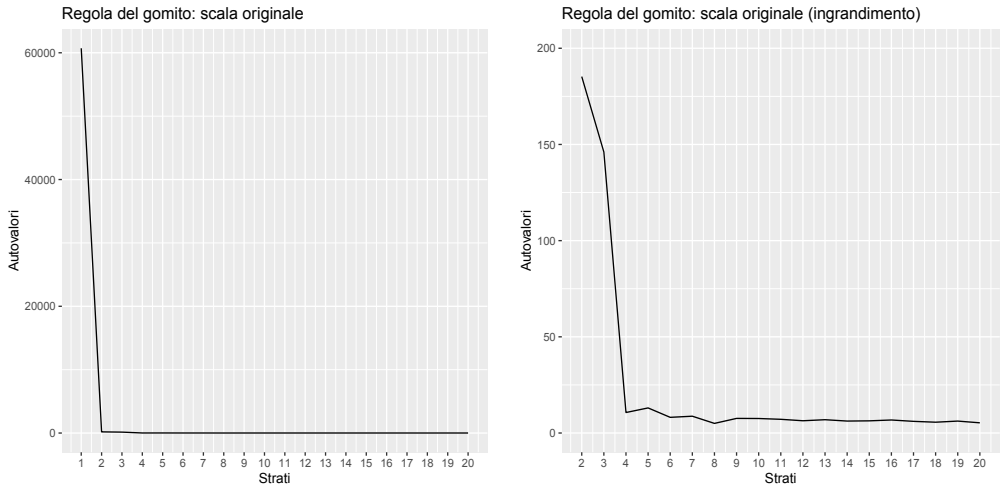

Figura 4.17: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione C2. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

Le composizioni dei 6 *bicluster* stimati per il modello sparseBC ([12]) sono presentate in Figura 4.20. In questo contesto di simulazione, per i primi due *bicluster* si ha l'interpretazione desiderata. Il primo, quello tra i sei con media più alta e che esprime dunque segnale più forte identifica esattamente la zona del quasar. Inoltre, il secondo è composto esattamente dai pixel in corrispondenza dei quali osserviamo il getto di raggi X. Se l'immagine simulata fosse di un vero quasar e volessimo sapere se è presente un getto di raggi X, i primi due *bicluster* verrebbero interpretati esattamente come sopra, poiché essendo le due zone individuate differenti, si concluderebbe che la prima, nettamente più luminosa della seconda, identifica il quasar e la seconda probabilmente il getto di raggi X. Tuttavia, dal secondo *bicluster* in poi si rinnova il problema di interpretazione che abbiamo visto essere presente anche negli altri studi di simulazione precedentemente discussi. Non permettendo ad un pixel di appartenere a più di un *bicluster* non siamo in grado di concludere se per esempio il terzo e il quarto *bicluster*, composti da pixel contigui gli uni agli altri, identifichino o meno lo stesso oggetto. Inoltre, gli ultimi due *bicluster* sono composti per la maggior parte da pixel in corrispondenza dei quali abbiamo simulato i conteggi relativi al rumore di fondo.

Il modello SSVD ([6]) si conferma dunque quello più soddisfacente da un punto di vista di interpretazione dei *bicluster* stimati, caratteristica di fondamentale importanza dati gli obiettivi dell'analisi. Inoltre, questo è l'unico caso in cui il modello sparseBC ([12]) fornisce dei *bicluster* stimati almeno parzialmente interpretabili. I risultati relativi agli studi di simulazione C1 e C3, che differiscono da quello appena trattato per la luminosità del getto di raggi X presente nell'immagine, rispettivamente debole e forte, possono essere visualizzati in Appendice A.5 e A.6. Differenze degne di nota rispetto allo studio C2 sono la tendenza del modello SSVD ad includere una quantità di rumore

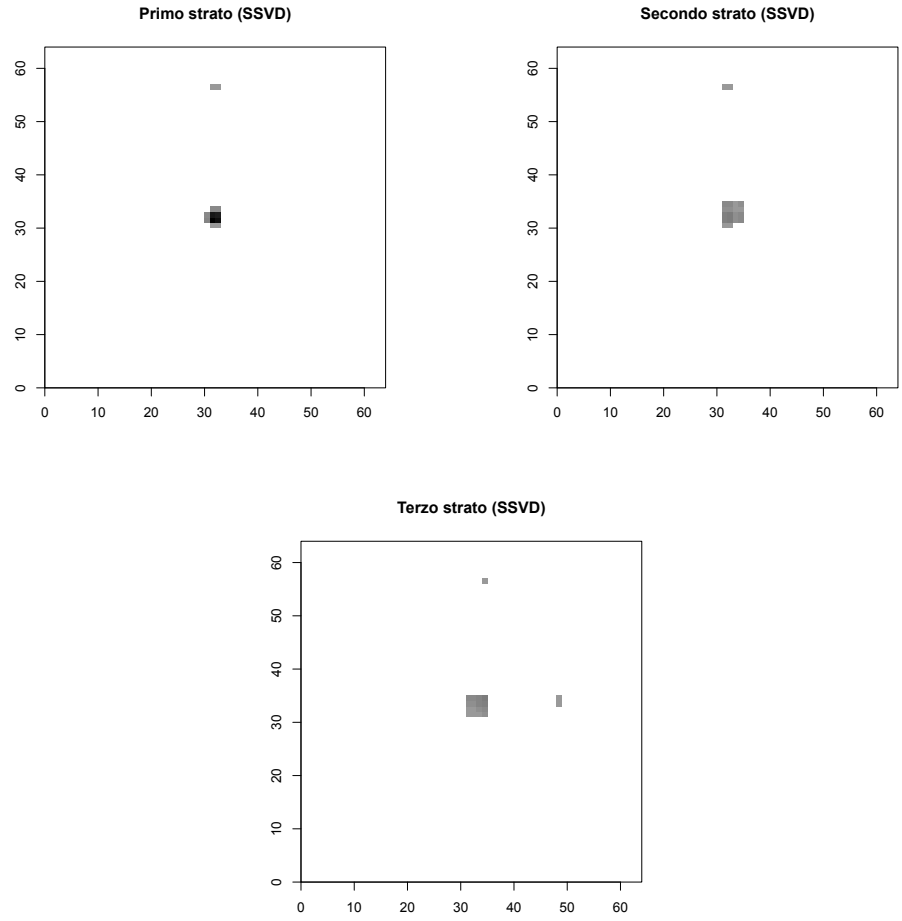

Figura 4.18: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione C2, primo strato in alto a sinistra, secondo in alto a destra, terzo in basso. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio in corrispondenza del pixel è elevato per il bicluster.*

più elevata nell'immagine ricostruita e l'incapacità di SSLB di individuare il getto debole nello studio C1. Tralasciando questo aspetto i risultati ottenuti sono simili a quanto già discusso.

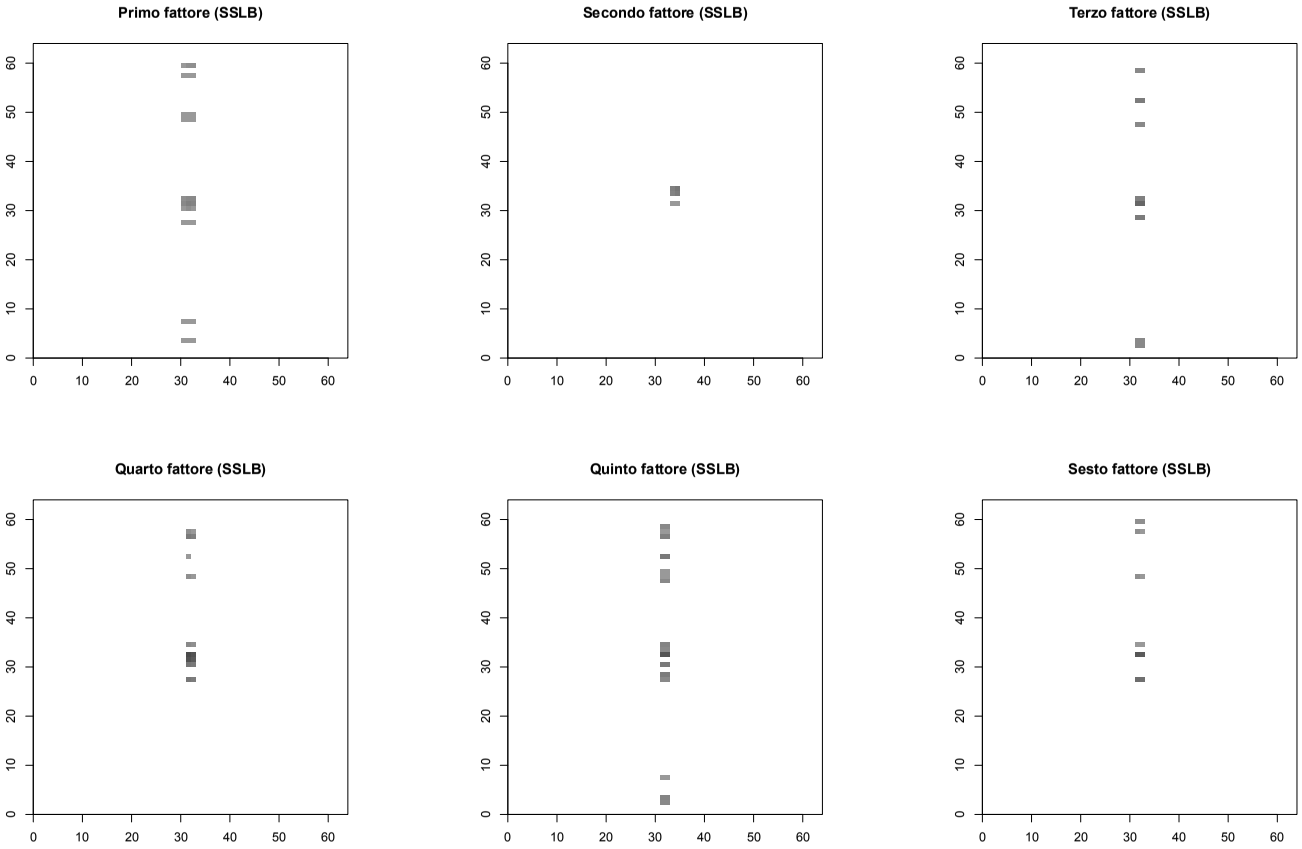

Figura 4.19: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione C2, dal primo fattore al quinto. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

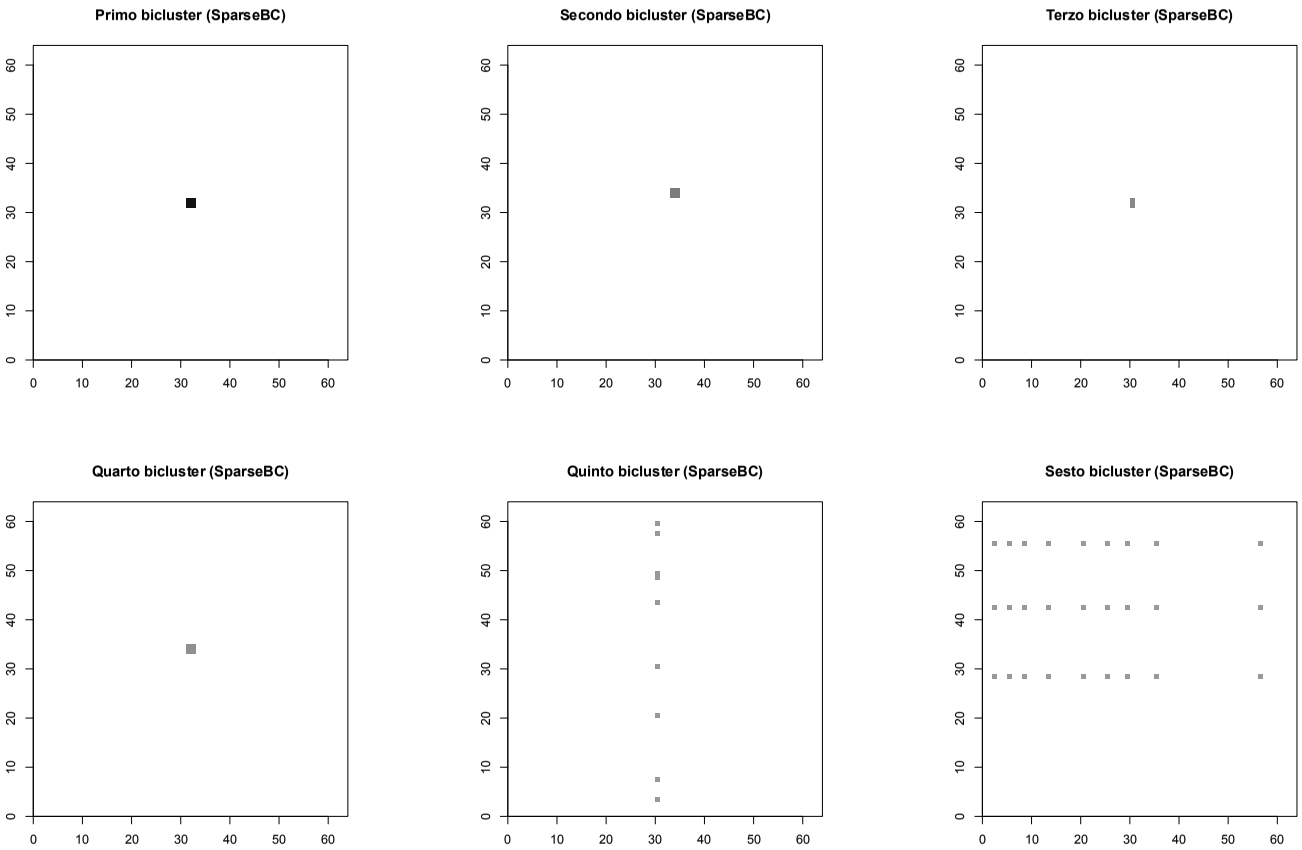

Figura 4.20: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione C2, dal primo al sesto. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

4.3 Riassumendo

In questo capitolo sono stati analizzati i risultati derivanti dall'applicazione dei tre metodi per il *biclustering* teorizzati nel Capitolo 2 a studi di simulazione di diverso grado di difficoltà, al fine di valutarne i limiti e i punti di forza.
Si è riscontrato che:

- Da un punto di vista di ricostruzione dell'immagine simulata, ovvero di capacità di eliminare il rumore di fondo mantenendo il segnale presente nell'immagine, il metodo più performante è SSVD. SparseBC e SSLB includono troppo spesso una quantità elevata di rumore di fondo nella loro immagine ricostruita e in aggiunta SSLB non riesce a ricostruire adeguatamente il segnale proveniente del getto getto di raggi X quando questo è di luminosità debole e delle volte anche quando è di luminosità forte (contesto B4). In generale il metodo più promettente sembra essere dunque SSVD, per la sua capacità di ricostruire l'immagine ma anche e soprattutto perché i *bicluster* da esso stimati si prestano alla tipologia di interpretazione che vogliamo dare nell'ambito astrofisico.

- SSLB e SparseBC non forniscono un'interpretabilità soddisfacente dei *bicluster* che trovano. Tuttavia, in un futuro lavoro questo aspetto potrebbe essere migliorato per esempio utilizzando una diversa strategia di selezione del numero di fattori in SSLB e della tripletta (G, R, λ) in sparseBC. Per quest'ultimo per esempio la nostra strategia basata sul BIC porta sempre alla scelta di un $\lambda = 1$ che corrisponde ad una penalizzazione esigua della verosimiglianza del modello, da cui probabilmente il numero spesso elevato di *bicluster* stimati. Per SSLB invece si dovrebbe dare la possibilità di specificare a priori un numero di fattori da ottenere in modo da poter valutare il modello in una griglia di valori. Questo potrebbe attenuare l'elevata variabilità legata alla stima del numero di *bicluster*.

Capitolo 5

Un esempio con dati reali

In questo capitolo si proveranno ad applicare i modelli descritti nel Capitolo 2 ad un insieme di dati reali presentato in [9]. I dati sono stati ottenuti dagli archivi del *Chandra X-Ray Observatory*.

5.1 I dati

L'immagine, in Figura 5.1, rappresenta una porzione di spazio di circa 22 arcominuti in diagonale che è stata partizionata in 1784×1691 pixel al fine di creare la matrice dei dati. L'oggetto celeste che viene preso in considerazione in questo caso non è un quasar, bensì una stella chiamata pulsar, interna alla nostra galassia. In particolare, in alto a sinistra c'è una rimanenza di una supernova (*supernova remnant*), non di interesse in questo contesto, mentre in basso a destra si possono vedere la pulsar, la nebulosa ad essa associata (*pulsar wind nebula*) e il getto di raggi X (*jet*), come indicato dalle etichette nell'immagine. Anche il dato a disposizione in corrispondenza di ciascun pixel è diverso, in questo caso non si tratta di conteggi di fotoni nella banda elettromagnetica dei raggi X, bensì di conte per unità di tempo di osservazione, di spazio e di energia. Questo tipo di applicazione, differente da quanto visto fino ad ora, nasce dalla volontà di includere nel modello l'informazione proveniente dall'energia dei fotoni rilevati, che non siamo riusciti ad ottenere per insiemi di dati relativi a quasar. Intuitivamente ci si aspetta che fotoni provenienti da quasar portino con se una maggiore energia rispetto a quelli provenienti dai getti di raggi X, che a loro volta dovrebbero essere associati a livelli di energia superiori rispetto ai fotoni provenienti da altre sorgenti che costituiscono il rumore di fondo. Si noti che le misurazioni stanno nell'intervallo $(0, 3.76 \times 10^{-5})$ molto differente da quanto visto fino ad ora e di conseguenza i metodi vanno adattati. Inoltre, in questa immagine è presente un ulteriore elemento di disturbo all'identificazione del getto di raggi X ovvero la nebulosa adiacente alla pulsar, che assieme alla pulsar stessa e al rumore di fondo potrebbe nascondere il segnale proveniente dal getto di raggi X. L'immagine in Figura 5.1 viene ritagliata per evidenziare la zona contenente la pulsar e il suo getto di raggi X, il risultato di questa operazione è visionabile in Figura 5.2, dove a fini puramente rappresentativi abbiamo considerato l'immagine in negativo,

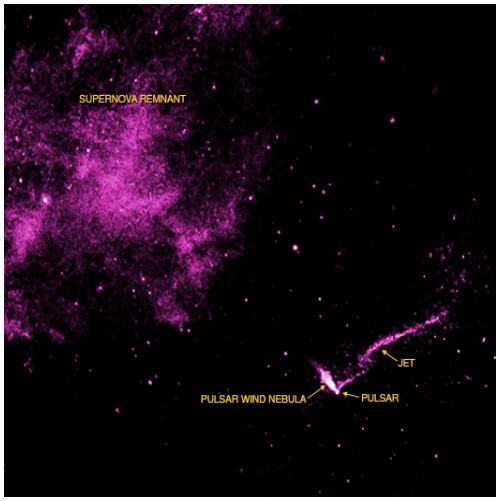

Figura 5.1: *Immagine della pulsar IGR J11014-6103 nella banda elettromagnetica dei raggi X, che rappresenta una porzione di spazio di circa 22 arcominuti, le etichette nell'immagine indicano i principali elementi illustrati: in alto a sinistra la rimanenza di una supernova, in basso a destra la pulsar con la nebulosa ad essa associata e il getto di raggi X.*

ovvero abbiamo trasformato i nostri dati in scala logaritmica e utilizzato un colore nero per pixel in corrispondenza dei quali abbiamo intensità alte e un colore bianco per pixel in corrispondenza dei quali abbiamo intensità basse. L'immagine così rielaborata è suddivisa in 450×400 pixel.

5.2 Applicazione dei metodi

Le principali modifiche che vanno apportate ai modelli sono di seguito elencate.

- Per SSVD ([6]) è necessario diminuire la soglia che determina la convergenza dell'algoritmo, poiché trattando dati di così esigua magnitudine necessitiamo di più accuratezza. In altre parole, una distanza tra due valori di 0.0001, che può sembrare minima, è nel nostro caso maggiore del massimo valore osservato.

- Per SSLB ([8]) la sequenza di λ_0 e $\tilde{\lambda}_0$, che prima non influenzava i risultati di stima, ora ha un peso non irrilevante. In particolare visto che i dati si concentrano in valori vicino a zero, è importante che la parte di *spike* della distribuzione a priori *spike and slab* che regola lo schiacciamento verso zero delle colonne di X e B (si veda paragrafo 2.2.1), sia adeguatamente tarata per tenere in considerazione fin dai primi elementi della sequenza le grandezze che caratterizzano i dati. In particolare, dopo prove empiriche si è riscontrato che delle sequenze adeguate sono $\lambda_0 = (10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9)$ e $\tilde{\lambda}_0 = (10^3, 5 \times 10^3, 5 \times 10^3, 5 \times 10^3, 5 \times 10^3, 5 \times 10^3, 5 \times 10^3)$.

- Per sparseBC ([12]) non sono state necessarie modifiche. Si fa notare, tuttavia, che la nostra strategia per la scelta della tripletta (G, R, λ) che

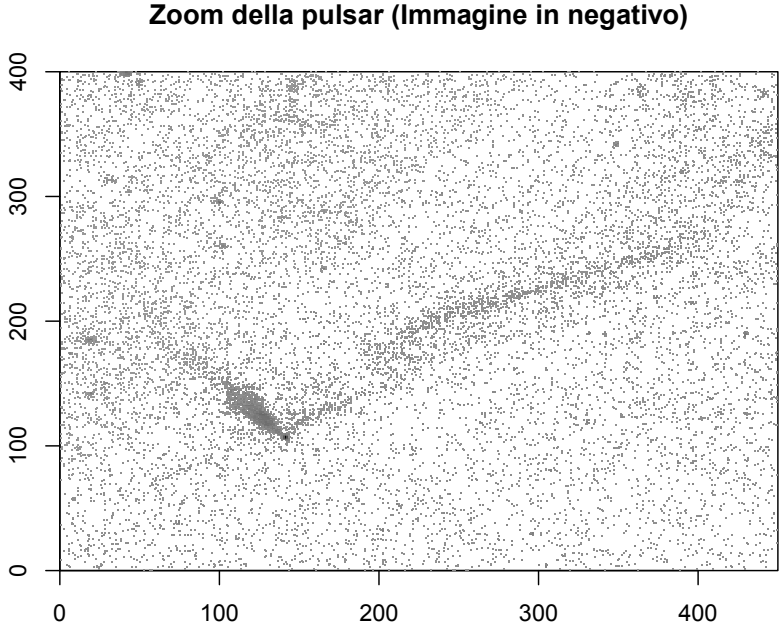

Figura 5.2: *Ingrandimento dell'immagine in Figura 5.1 nella zona di interesse della pulsar. Per evidenziare gli elementi presenti abbiamo ottenuto il grafico a partire dai dati in scala logaritmica e utilizzato il colore nero per intensità alte e il colore bianco per intensità basse.*

corrisponde al miglior adattamento del modello ai dati, seleziona sempre il parametro di lisciamento $\lambda = 0$. Intuitivamente il modello preferisce non schiacciare verso zero le medie di *bicluster* $\{\mu_{gr}\}_{g,r=1}^{G,R}$, di per sé già molto piccole.

5.3 Risultati

Per il modello sparseBC, la nostra strategia di selezione della tripletta (G, R, λ) ha portato a trovare i valori $(19, 17, 0)$ e dunque un totale di $19 \times 17 = 323$ *bicluster* individuati. Non essendo possibile determinare un opportuno arrotondamento, come nel caso di dati di conteggio visto negli studi di simulazione, tutti i *bicluster* identificati vanno considerati. Un numero così elevato di *bicluster* è probabilmente dovuto al fatto che non stiamo effettuando alcun tipo di schiacciamento verso zero, data la scelta di $\lambda = 0$, una interpretazione dei gruppi è dunque impossibile. L'immagine ricostruita dal modello tuttavia (Figura 5.3, in alto a destra), identifica correttamente oltre alla pulsar e alla sua nebulosa, una zona più scura lì dove c'è il getto di raggi X. La ricostruzione non è sufficiente senza la disponibilità di una interpretazione dei *bicluster* che ci permetta di associarli ad un oggetto celeste specifico. Dato l'elevato numero di gruppi e l'impossibilità di una loro interpretazione non vengono riportati i

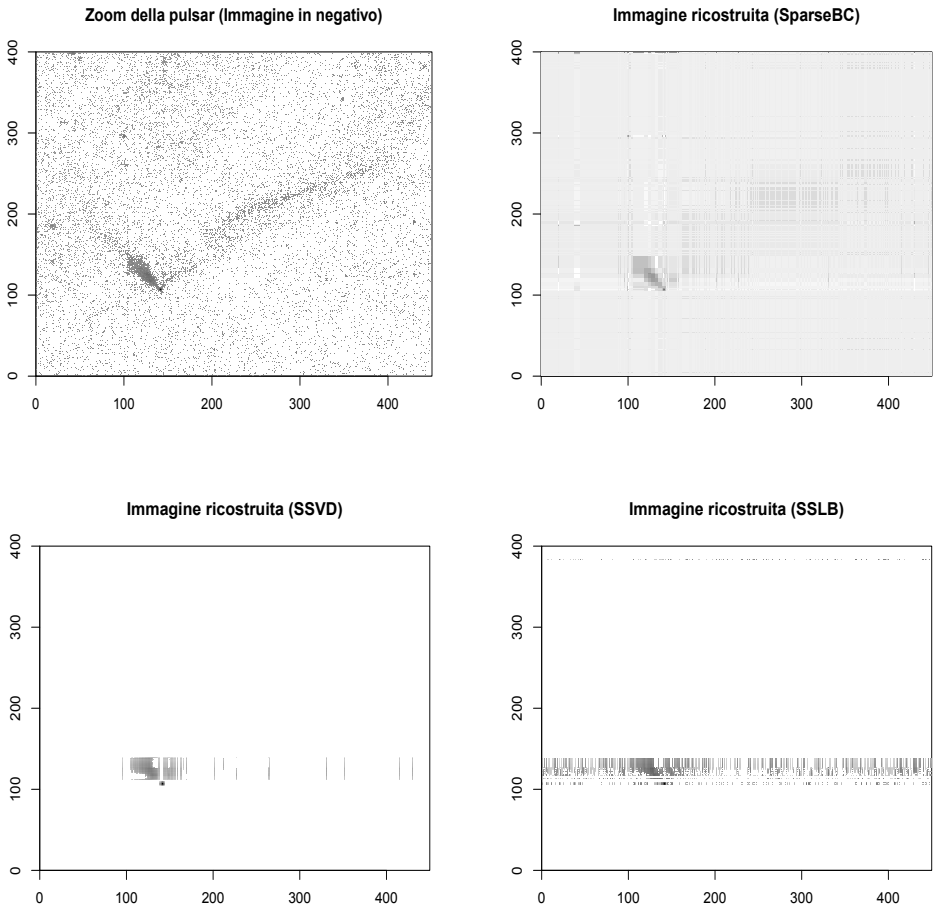

Figura 5.3: *Immagine originale e immagini ricostruite dai tre metodi SparseBC, SSVD e SSLB rispettivamente in alto a sinistra, in alto a destra, in basso a sinistra e in basso a destra, in scala logaritmica, al fine di evidenziarne le caratteristiche. Al colore nero corrispondono intensità alte e al colore bianco intensità basse.*

grafici relativi alla loro composizione.

Per il modello SSVD, come si può vedere in Figura 5.4, seguendo la regola del gomito viene scelto un numero di strati pari a 3. Anche in questo caso è necessario un ingrandimento per andare oltre il primo strato che identifica probabilmente la pulsar. Il modello fallisce nell'identificazione del getto di raggi X, tuttavia elimina gran parte del rumore di fondo (si veda Figura 5.3, in basso a sinistra). Vengono riportate in Figura 5.5 le immagini che rappresentano la composizione degli strati o *bicluster* stimati da SSVD. Come si può notare, il metodo conserva almeno parzialmente la sua caratteristica di interpretabilità dei *bicluster* che individua. Il primo, infatti, è composto da pixel chiaramente in corrispondenza della pulsar, mentre il secondo e il terzo contengono troppo rumore per concludere che identifichino la nebulosa della pulsar.

Per il modello SSLB abbiamo trovato un numero di *bicluster* pari a 7 applicando la nostra strategia di scelta del miglior adattamento tra 100 lanci

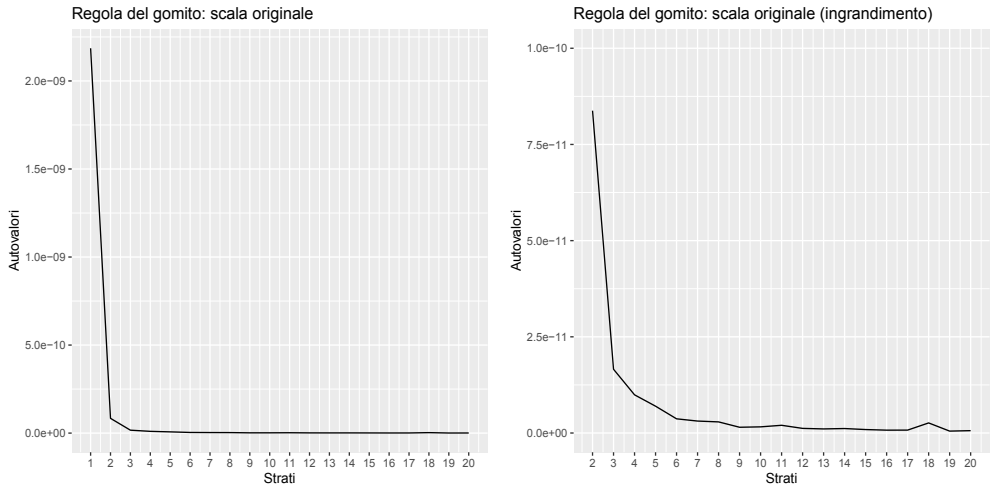

Figura 5.4: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

dell'algoritmo. L'immagine ricostruita (Figura 5.3, in basso a destra), è analoga a quella ottenuta tramite SSVD, ma presenta più rumore di fondo residuo. Ancora una volta, come negli studi di simulazione, SSLB non funziona molto accuratamente nel contesto di ricostruzione di immagini. Non vengono riportate le composizioni dei *bicluster* stimati poiché non sono interpretabili, proprio come si è riscontrato negli studi di simulazione. Ogni fattore stimato include pixel associati al rumore di fondo e dunque non distingue un oggetto celeste.

Si conclude che per questo insieme di dati particolarmente complesso, dato sia l'intervallo dei valori con cui abbiamo lavorato, sia l'ulteriore elemento di disturbo rappresentato dalla nebulosa, i metodi vadano ulteriormente approfonditi. Non solo faticano ad identificare il rumore di fondo, ovvero tutte quelle conte per unità di tempo, di spazio e di energia che non provengono dalle zone occupate da pulsar, nebulosa e getto di raggi X, ma quest'ultimo non viene neanche identificato, se non debolmente da sparseBC. Infine, l'interpretazione dei *bicluster* trovati non è soddisfacente per gli scopi dell'analisi.

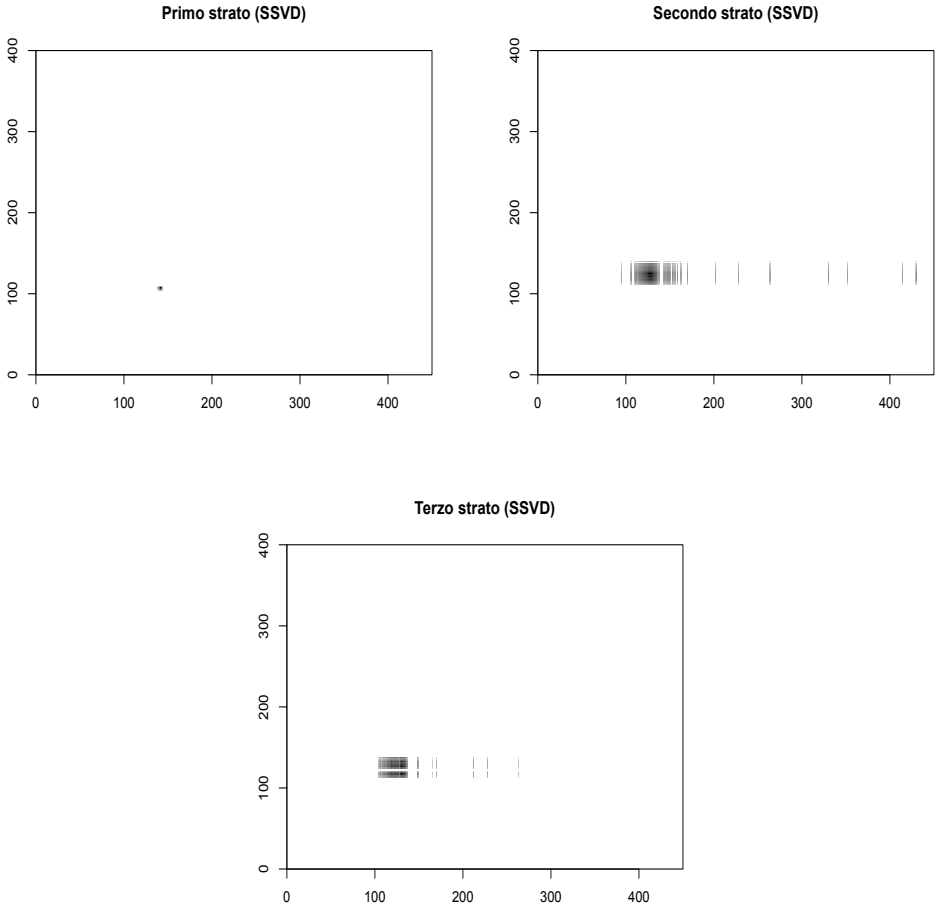

Figura 5.5: *Rappresentazione della composizione degli strati, o bicluster, stimati dal modello SSVD. Un colore bianco in corrispondenza di un pixel indica che il pixel non compone lo strato, più il colore è nero più il pixel è presente nello strato.*

Capitolo 6

Conclusioni

L'identificazione di sorgenti come i quasar e i loro getti di raggi X è di grande importanza per comprendere le forze e i meccanismi che governavano il giovane Universo. In questa tesi abbiamo proposto una classe di metodi alternativa per compiere questa analisi. Abbiamo introdotto nel Capitolo 1 un esempio della strategia di verifica di ipotesi tradizionalmente utilizzata in letteratura. Nel Capitolo 2 abbiamo presentato i metodi per il *biclustering* come un approccio modellistico al problema, sottolineando che il raggruppamento simultaneo di righe e colonne della matrice dei dati coincide con l'identificazione di aree dell'immagine contenenti segnale. Nel Capitolo 3 sono stati delineati i diversi studi di simulazione con i quali abbiamo successivamente saggiato i limiti e le qualità dei metodi in questo contesto, al quale non erano mai stati applicati.
Nel Capitolo 4 abbiamo presentato i risultati più significativi che abbiamo ottenuto, riscontrando che il metodo più promettente è SSVD ([6]), che si basa su una versione per matrici sparse della scomposizione a valori singolari. In particolare, quest'ultimo è l'unico che fornisce un'interpretazione soddisfacente dei *bicluster* stimati, di fondamentale importanza al fine del riconoscimento degli oggetti celesti nell'immagine osservata.
Nel Capitolo 5 si è desiderato applicare i metodi ad un conteso abbastanza differente, soprattutto per la natura dei dati considerati che non erano più conteggi bensì conte per unità di tempo, di spazio e di energia. In questo modo abbiamo attuato un primo tentativo di includere l'informazione dell'energia dei fotoni nel modello, aspetto che riteniamo essere rilevante nel contesto astrofisico. Questa applicazione ha risaltato i limiti dei metodi che, di fronte alla complessità dell'analisi dell'immagine osservata, non sono riusciti ad identificare il getto di raggi X fuoriuscente dalla pulsar. In particolare SSLB ([8]) e SSVD non hanno ricostruito la parte di immagine corrispondente al getto, mentre sparseBC ([12]) l'ha lievemente ricostruita ma senza fornire dei *bicluster* stimati che possano essere interpretati. Si conclude che i metodi, in questo contesto applicativo particolarmente complesso, vadano ulteriormente approfonditi.
Un lavoro futuro potrebbe essere focalizzato proprio sull'aspetto dell'inclusione nell'analisi dell'informazione sull'energia. Si potrebbe ad esempio lavorare con dati di conteggio, per i quali almeno SSVD sembra avere delle potenzialità, ed

utilizzare a posteriori l'energia, per etichettare i *bicluster* trovati. Oggetti celesti differenti emettono fotoni a diversi livelli di energia, si deduce dunque che questo aspetto potrebbe essere determinante per comprendere che tipologia di oggetto celeste i *bicluster* identifichino. Potrebbero essere risolti, con questa strategia, i problemi di interpretazione dei *bicluster* che abbiamo riscontrato in questa tesi. Alternativamente, si potrebbero applicare i metodi per il *biclustering* anche ai dati relativi all'energia e combinare le due analisi per confermare o confutare i risultati ottenuti sui dati di conteggio. In particolare, se le zone dell'immagine identificate dalle due analisi coincidono, è verosimile che esse contengano effettivamente segnale. Infine, combinando le due analisi, avremmo anche la possibilità di determinare il livello di energia associato a ciascun *bicluster* e dunque fornirne una interpretazione.

Appendice A

Ulteriori risultati

A.1 Quasar variabile senza getto (A2)

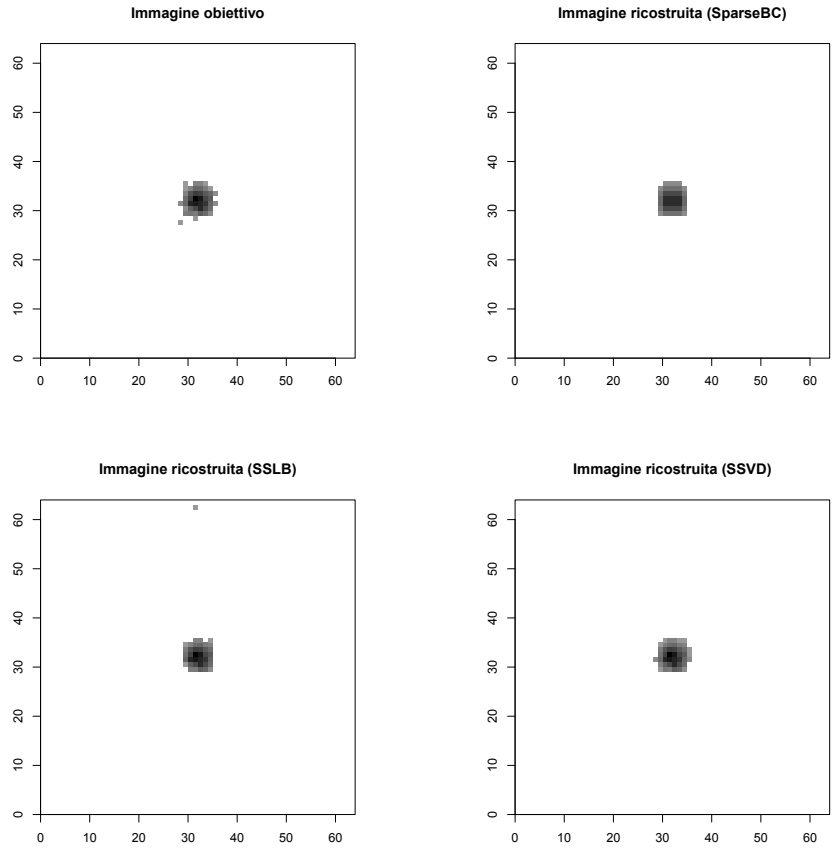

Figura A.1: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione A2, che considera un'immagine composta da un quasar diffuso, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

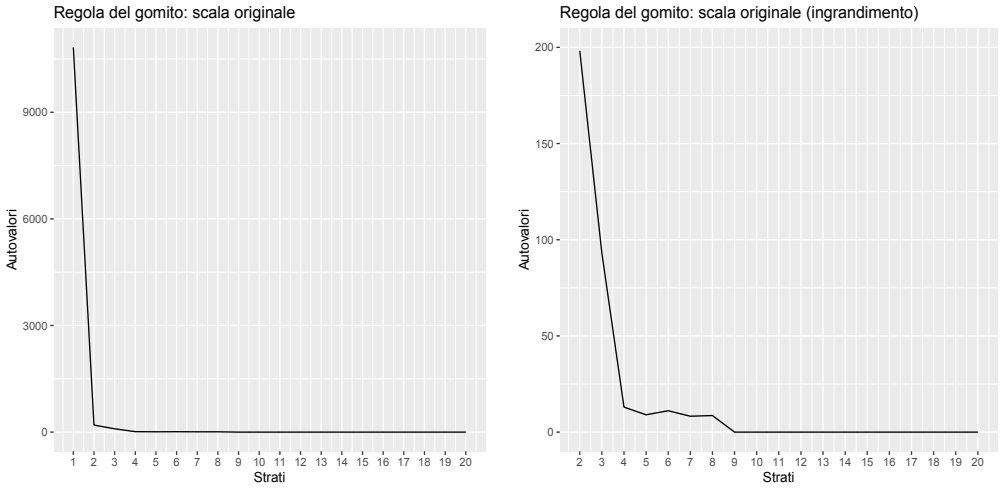

Figura A.2: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione A2. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	1.000	0.867	15.956	7
SSLB	1.000	0.822	1.089	3
SSVD	0.999	0.933	0.578	3

Tabella A.1: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione A2*

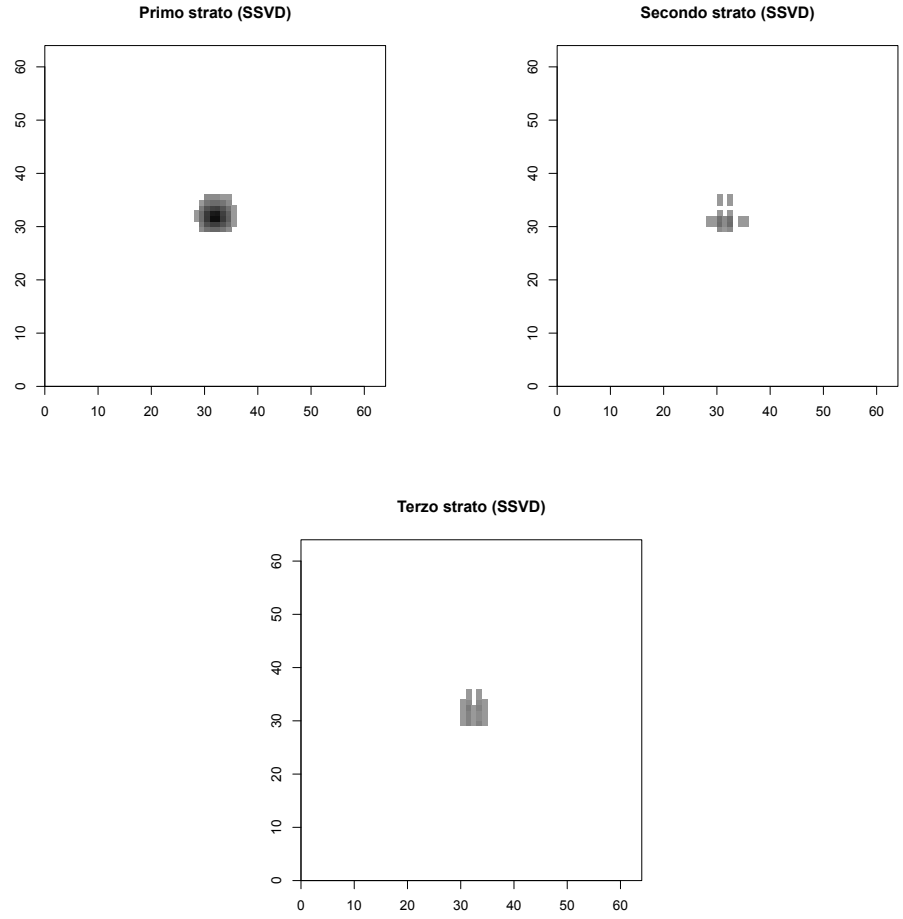

Figura A.3: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione A2, primo strato a in alto a sinistra, secondo in alto a destra e terzo in basso. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

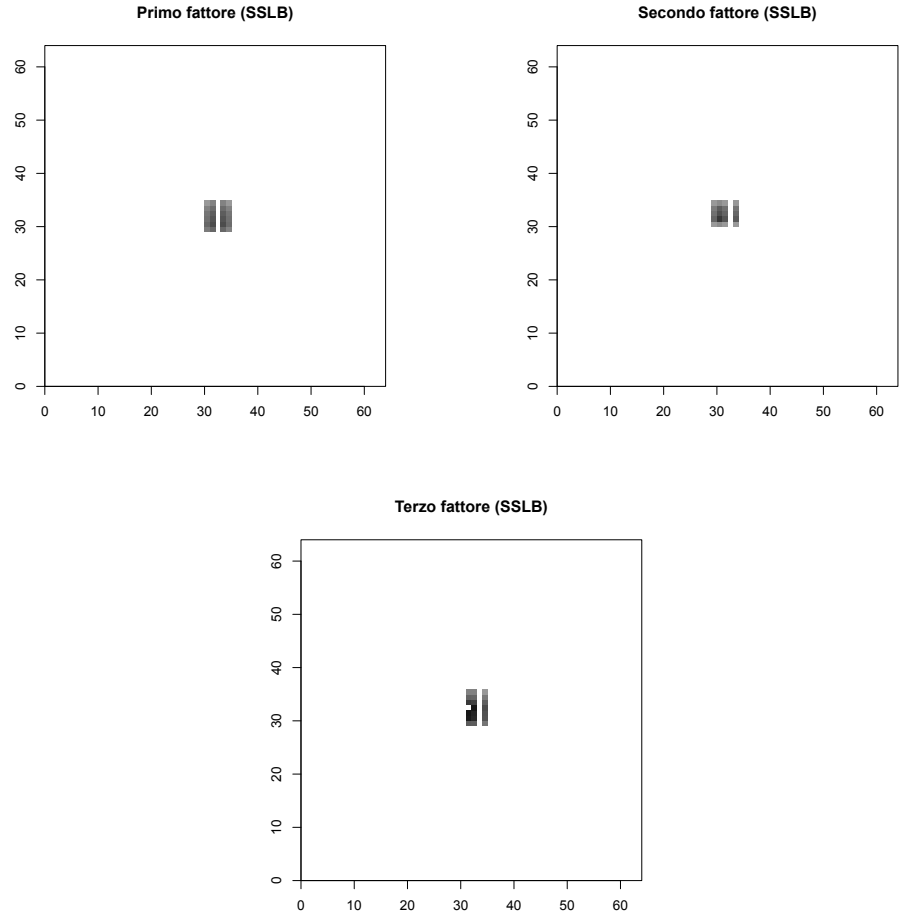

Figura A.4: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione A2, dal primo fattore al terzo, rispettivamente in alto a sinistra, in alto a destra e in basso. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

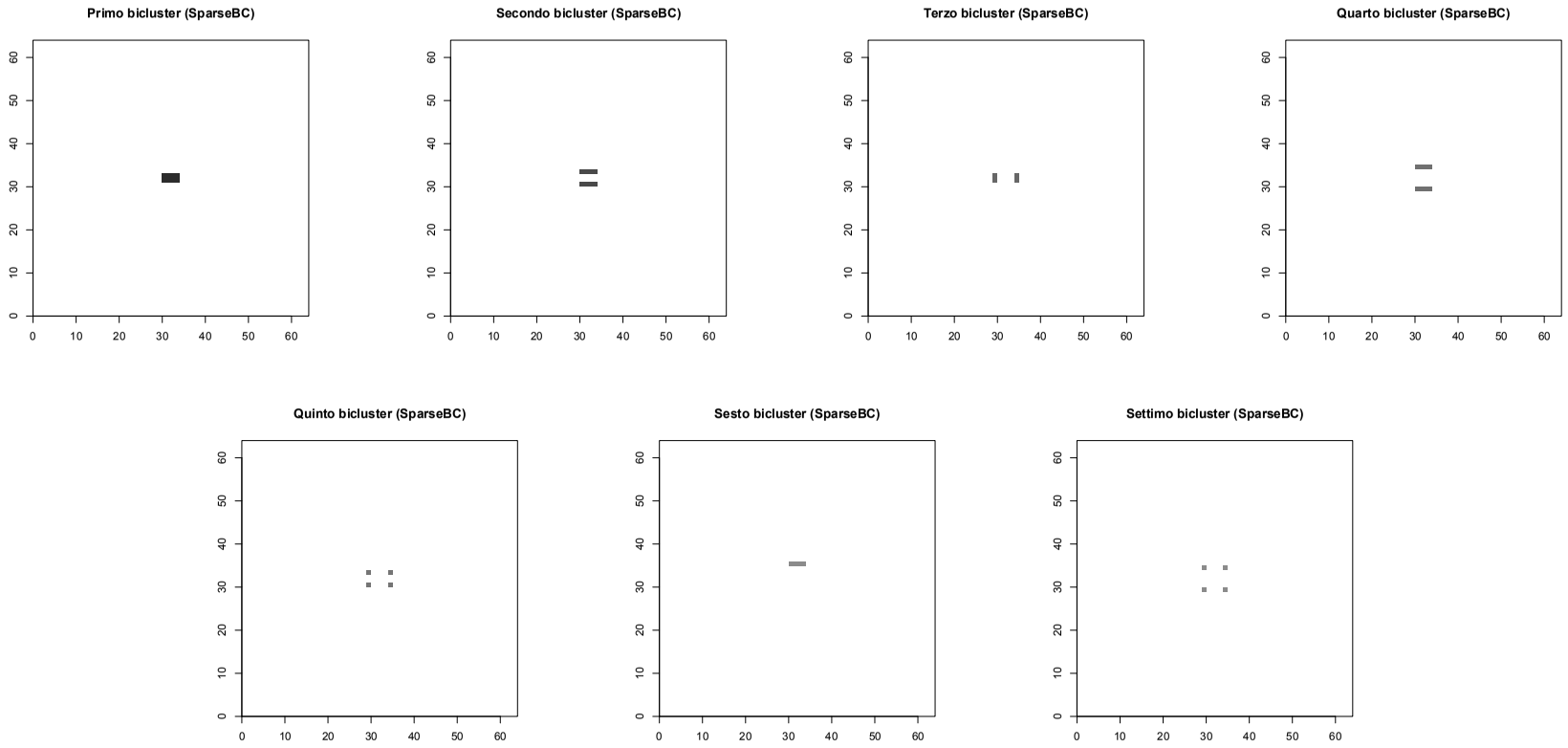

Figura A.5: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione A2, dal primo al settimo. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

A.2 Quasar con getto debole lontano (B1)

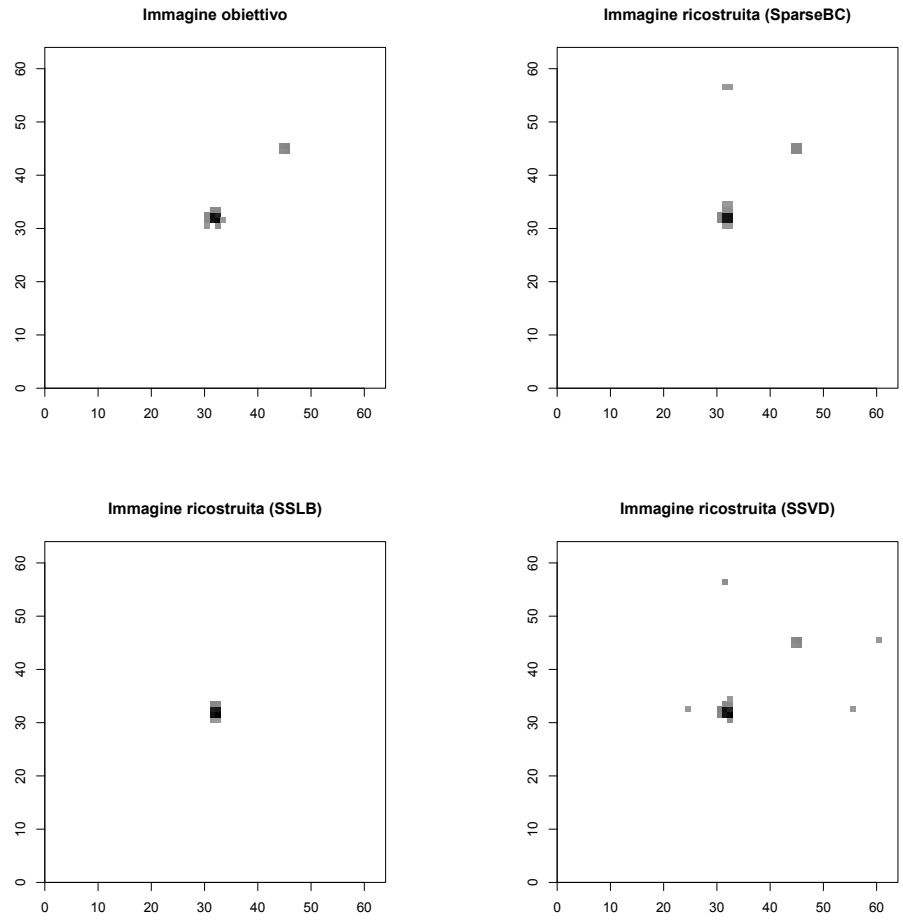

Figura A.6: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione B1, che considera un'immagine composta da un quasar e un getto di raggi X lontano da esso di luminosità debole, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

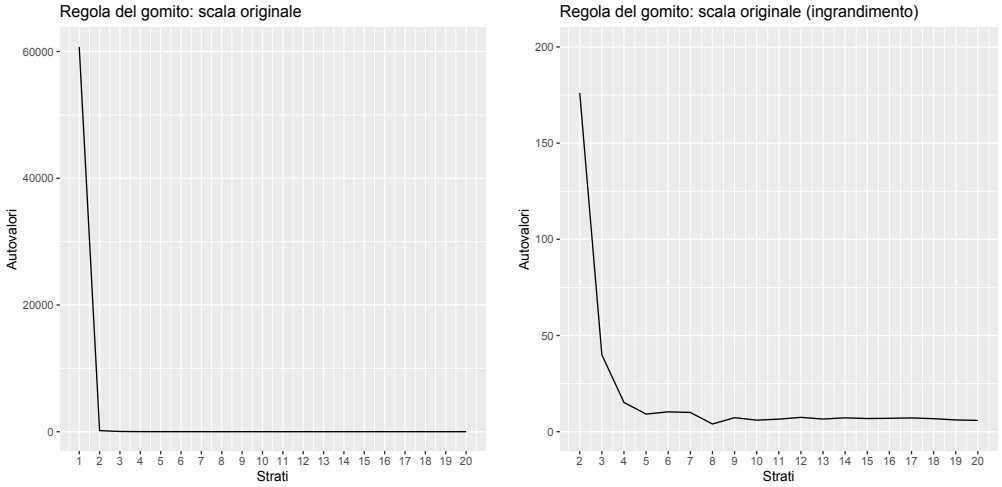

Figura A.7: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione B1. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.999	0.867	22.200	5
SSLB	1.000	0.467	24.800	3
SSVD	0.999	0.867	0.200	3

Tabella A.2: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione B1*

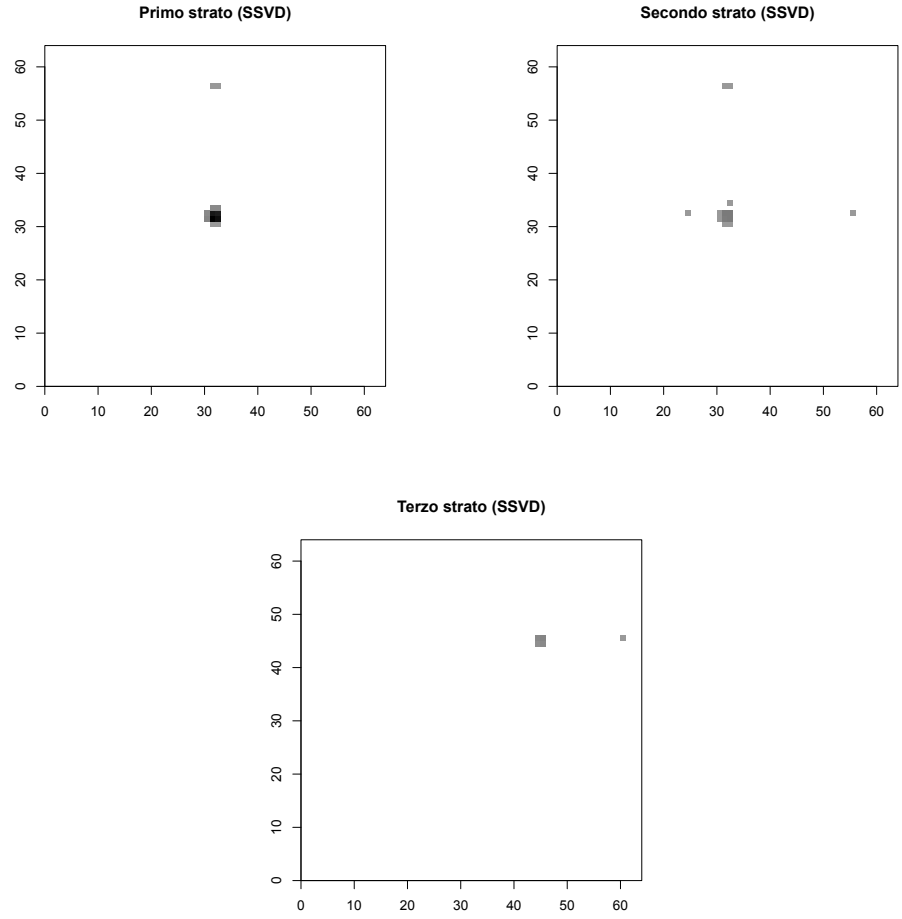

Figura A.8: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione B1, primo strato a in alto a sinistra, secondo in alto a destra e terzo in basso. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

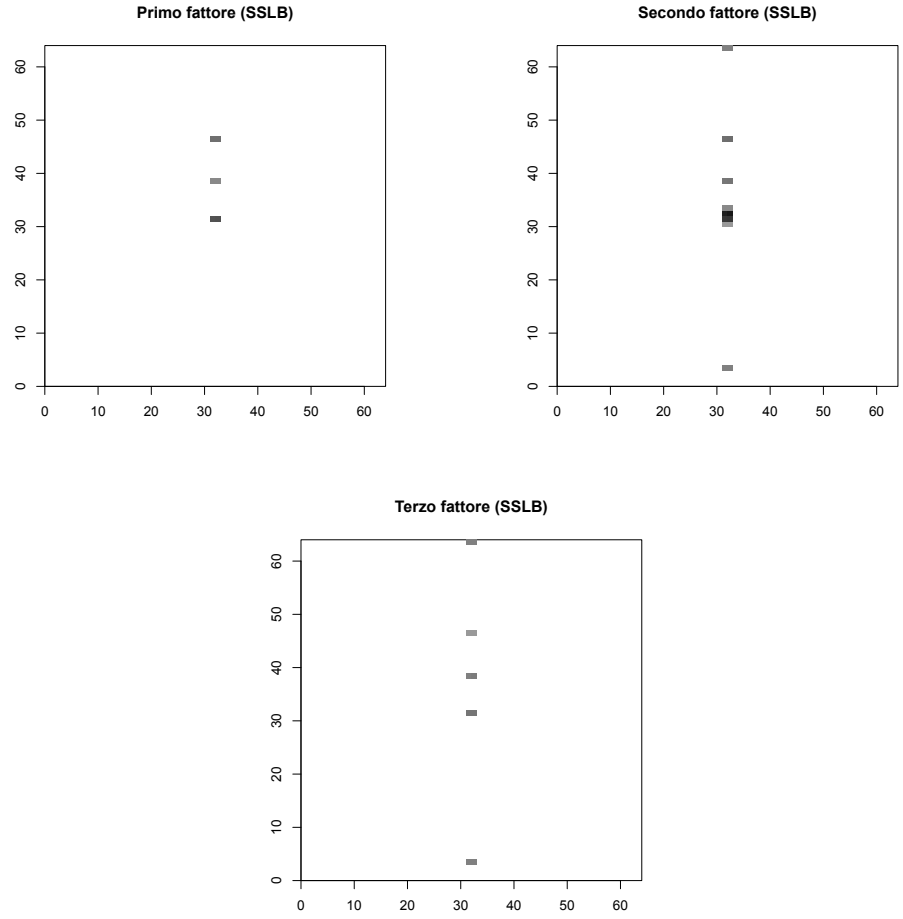

Figura A.9: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione B1, dal primo fattore al terzo. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

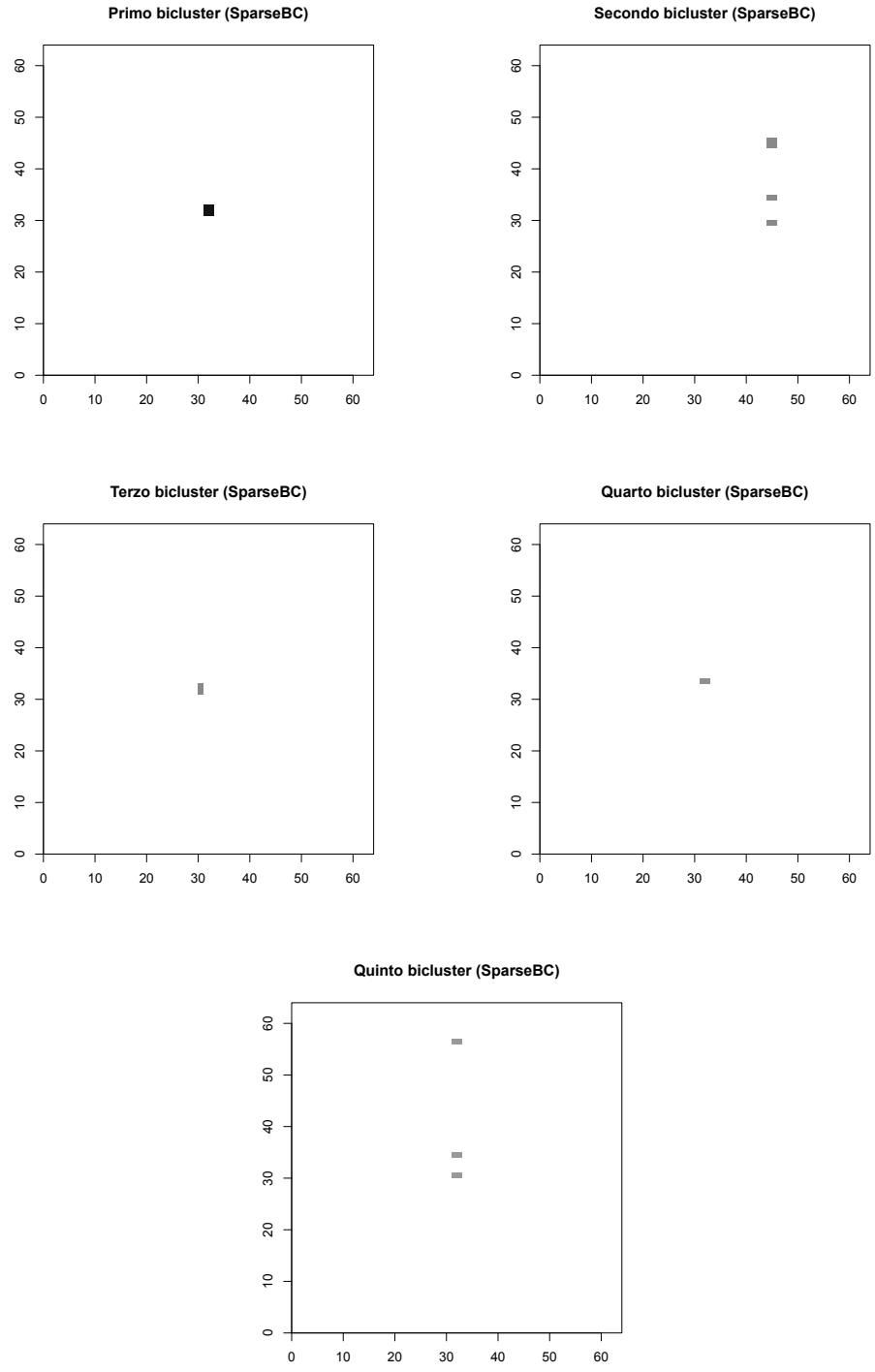

Figura A.10: *Rappresentazione in negativo della composizione dei biclu-ster, stimati da SparseBC per lo studio di simulazione B1, dal primo al quinto. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

A.3 Quasar con getto forte lontano (B2)

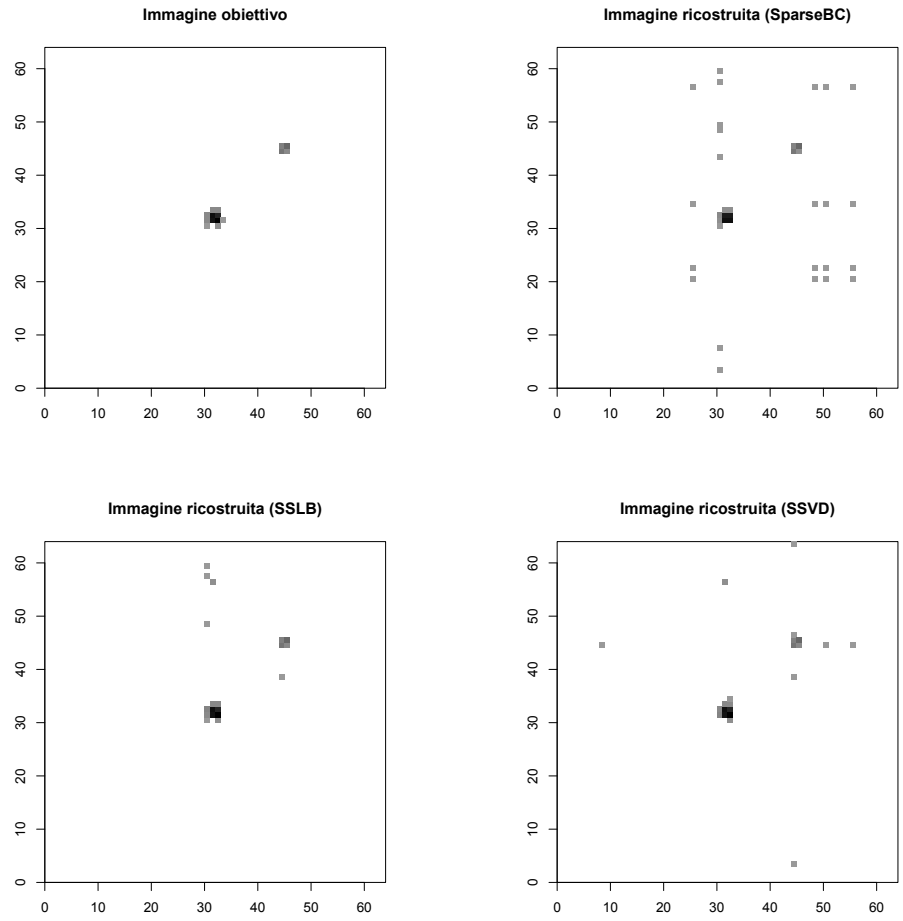

Figura A.11: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione B2, che considera un'immagine composta da un quasar e un getto di raggi X lontano da esso di luminosità forte, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

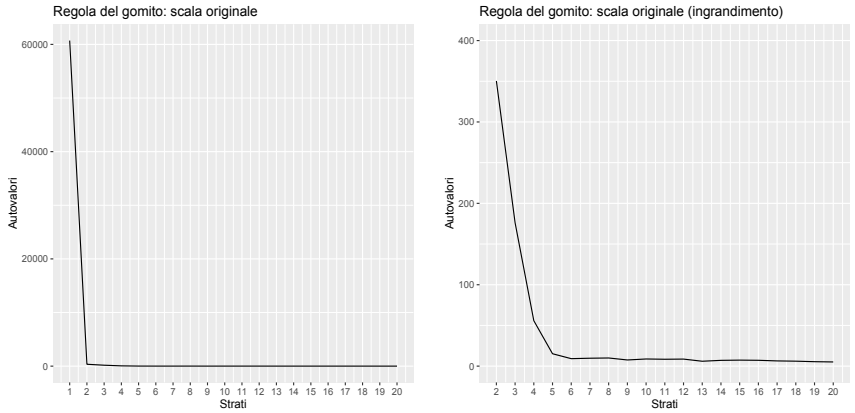

Figura A.12: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione B2. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.994	0.867	14.533	11
SSLB	0.999	0.933	0.133	8
SSVD	0.998	0.867	0.200	4

Tabella A.3: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione B2*

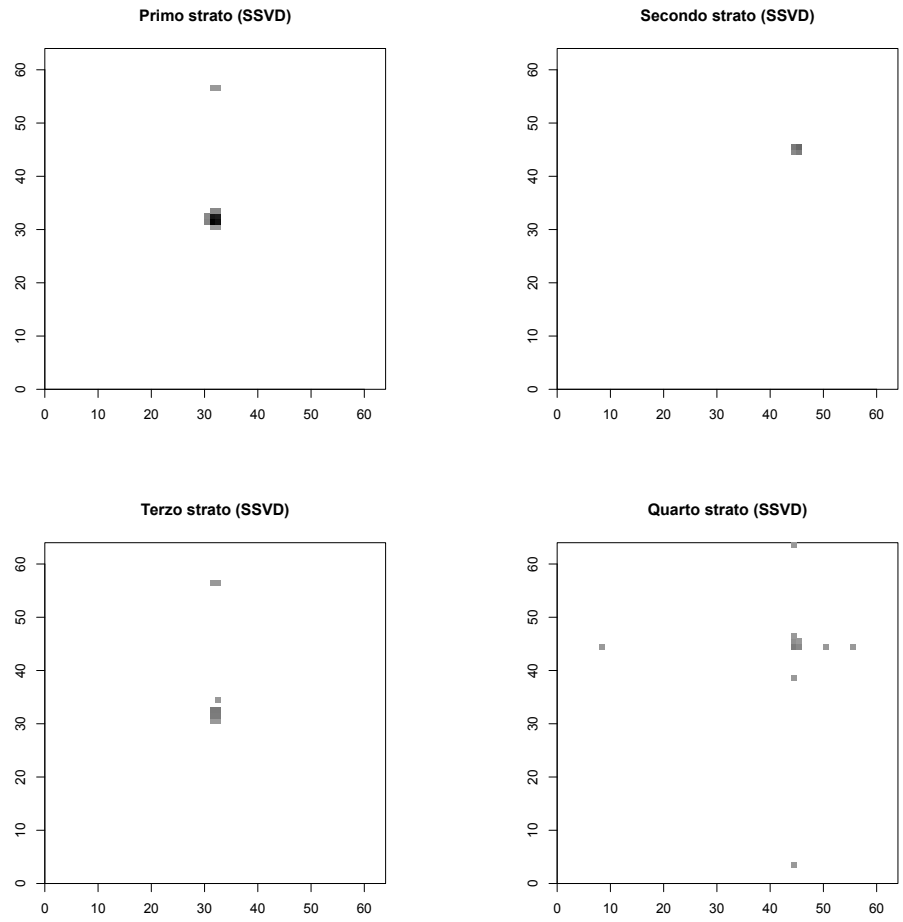

Figura A.13: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione B2, al primo al quarto, in alto a sinistra, in alto a destra, in basso a sinistra e in basso a destra rispettivamente. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

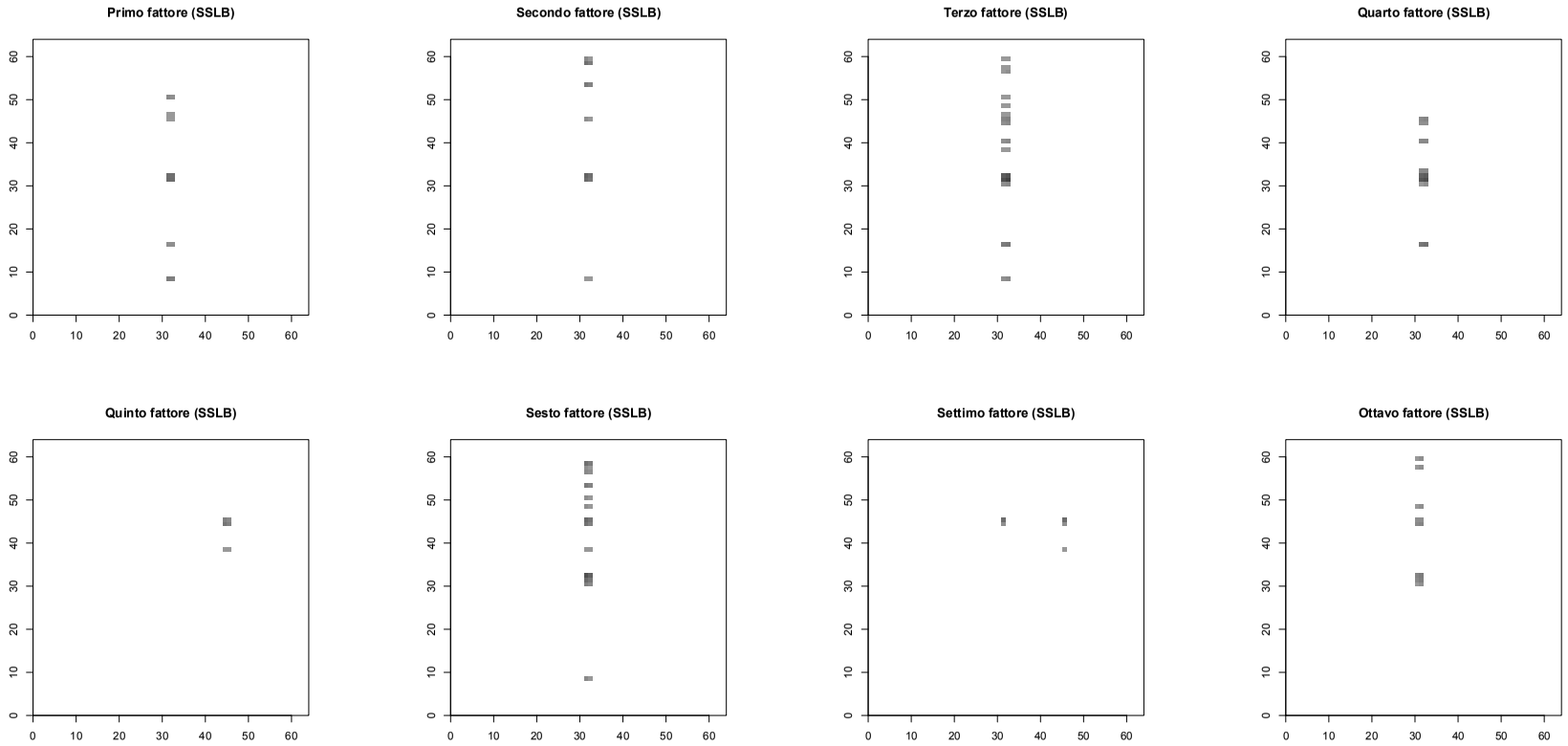

Figura A.14: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione B2, dal primo fattore all'ottavo. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

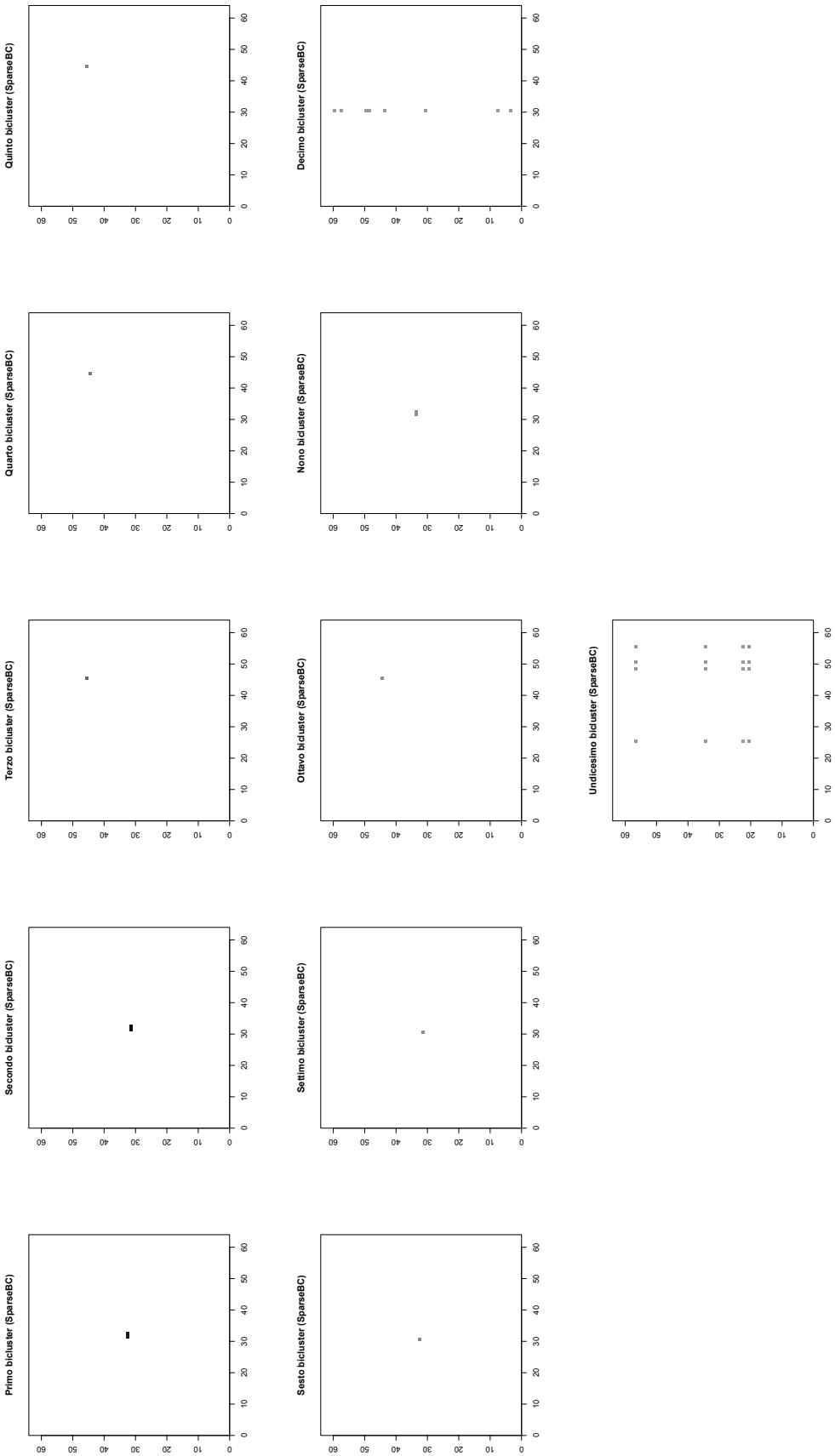

Figura A.15: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione B2, dal primo all'undicesimo. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

A.4 Quasar con due getti deboli lontani (B3)

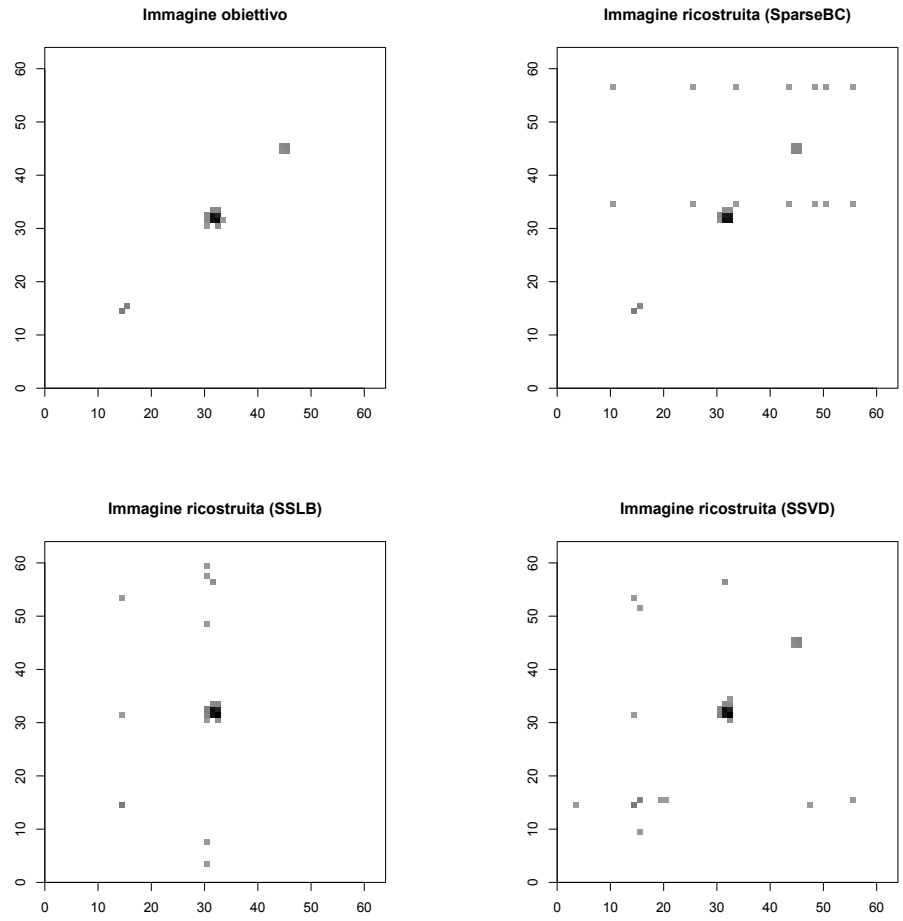

Figura A.16: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione B3, che considera un'immagine composta da un quasar e due getti di raggi X lontani da esso di luminosità debole, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

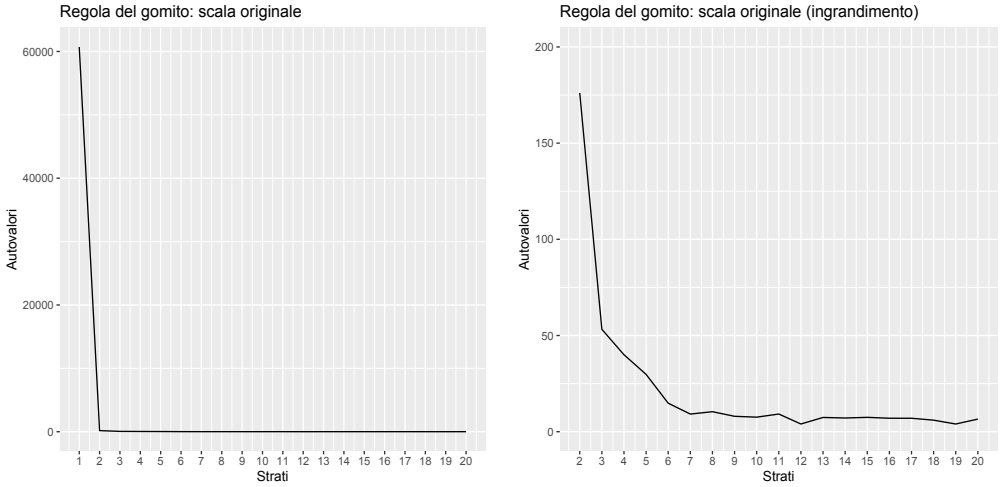

Figura A.17: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione B3. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.997	0.824	12.882	9
SSLB	0.998	0.647	4.412	6
SSVD	0.997	0.882	0.176	5

Tabella A.4: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione B3*

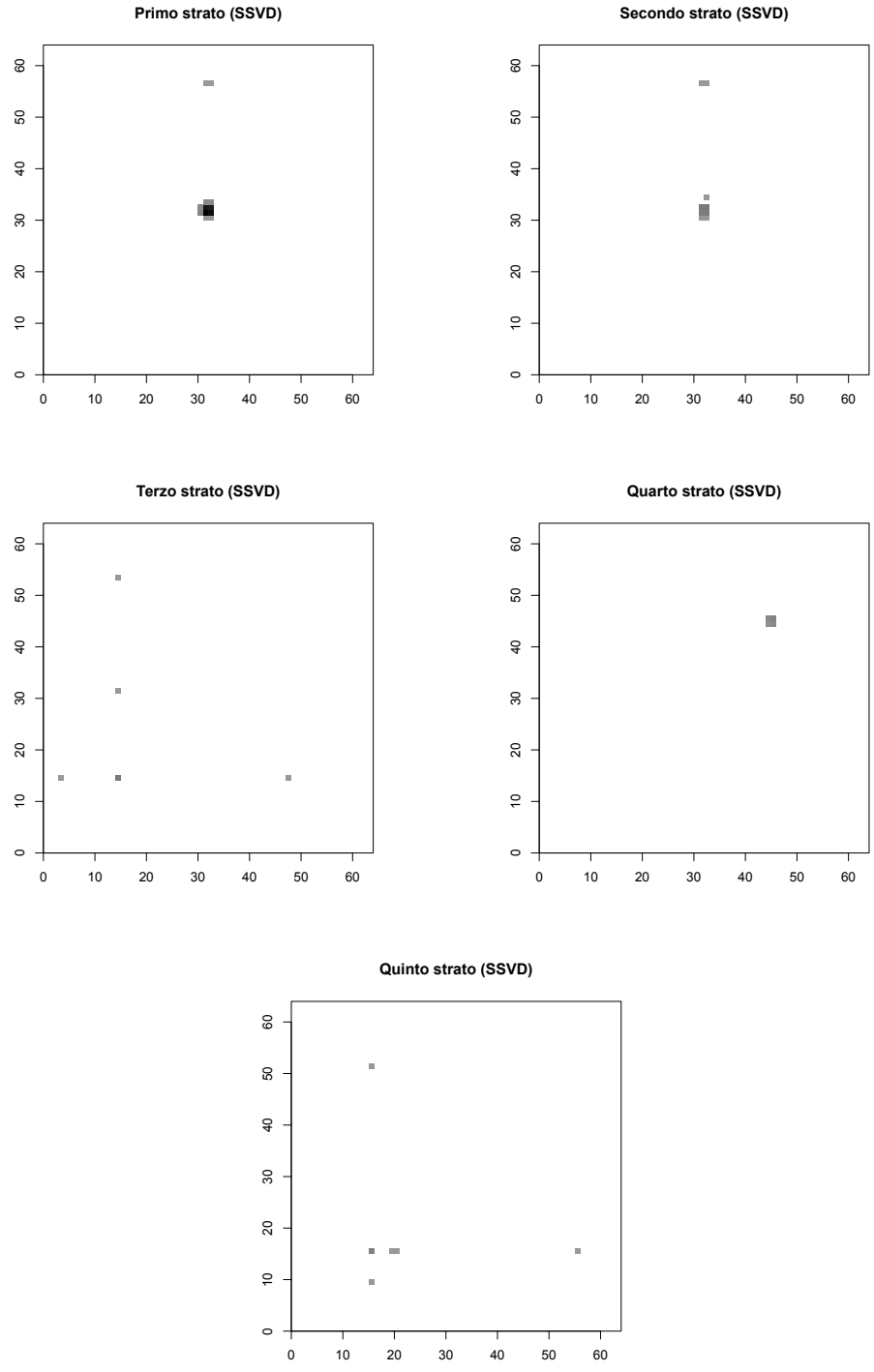

Figura A.18: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione B3, al primo al quinto. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

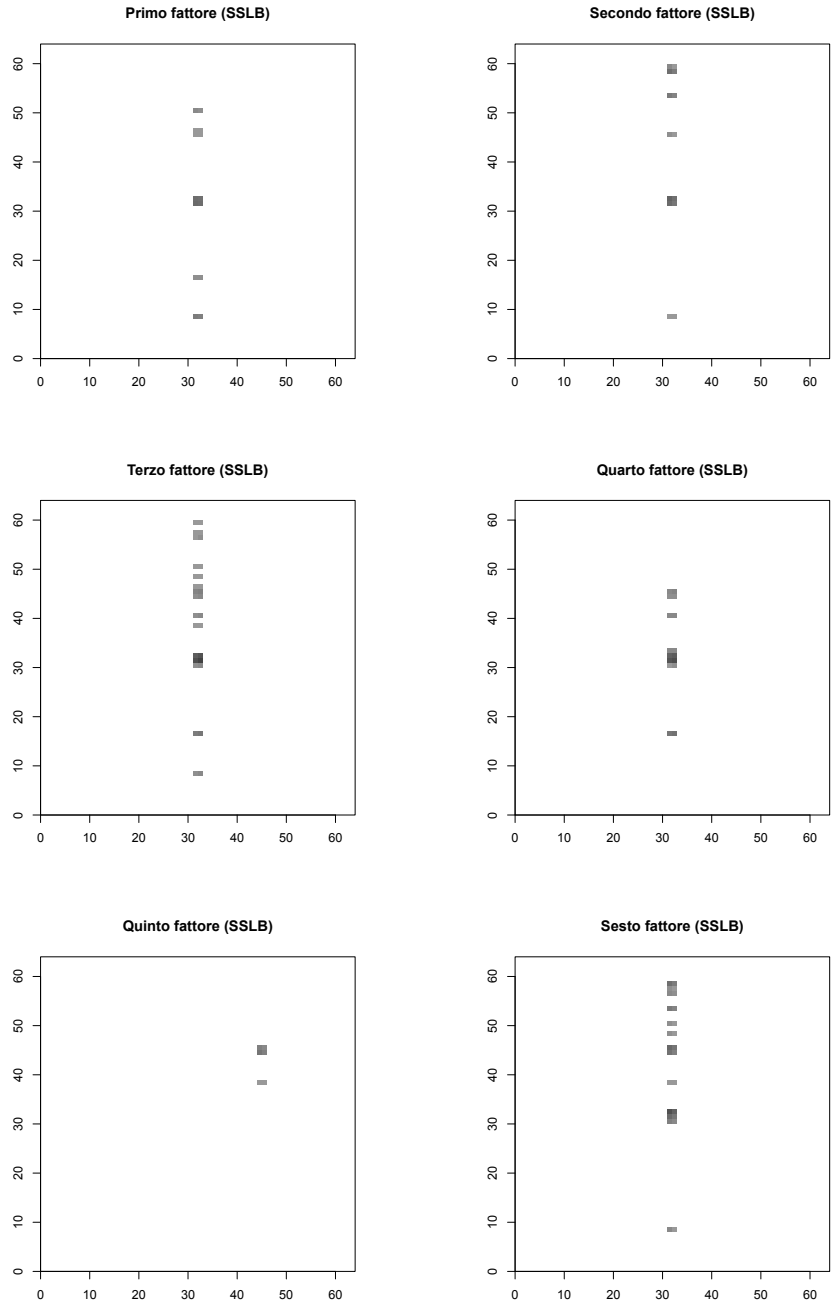

Figura A.19: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione B3, dal primo fattore al quinto. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

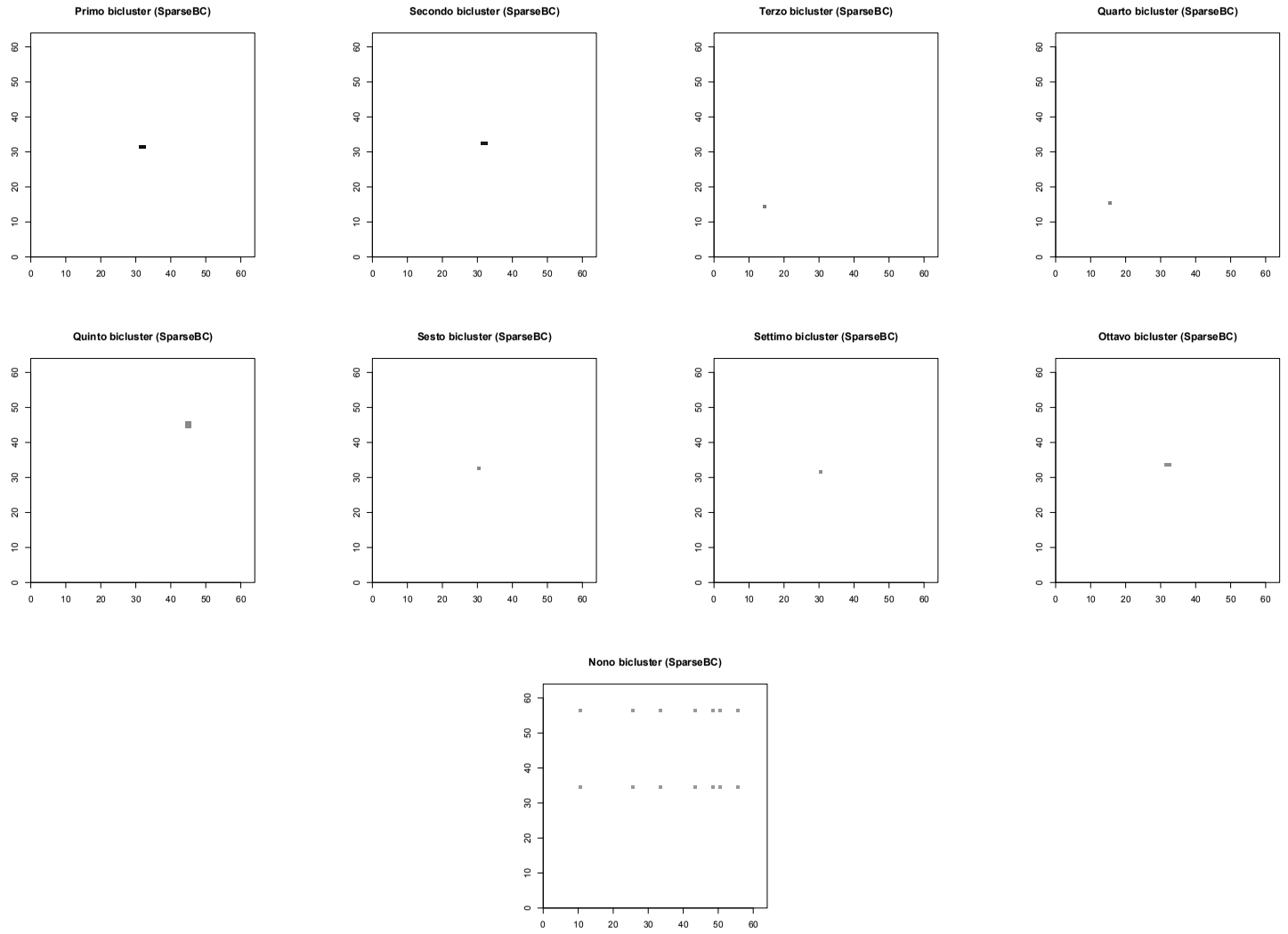

Figura A.20: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione B3, dal primo al nono. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

A.5 Quasar con getto debole contiguo (C1)

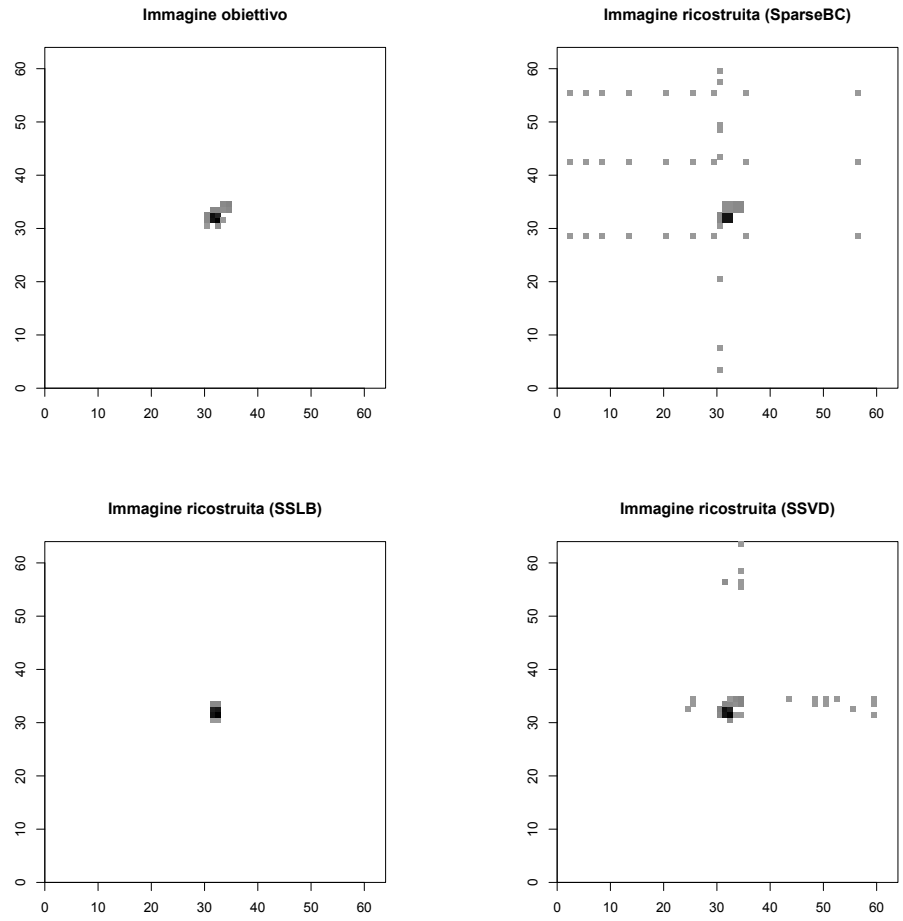

Figura A.21: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione C1 che considera un'immagine simulata composta dal quasar e da un getto di raggi X contiguo ad esso di luminosità debole e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra). Si noti che l'immagine ricostruita da SSLB coincide con la composizione dell'unico fattore trovato.*

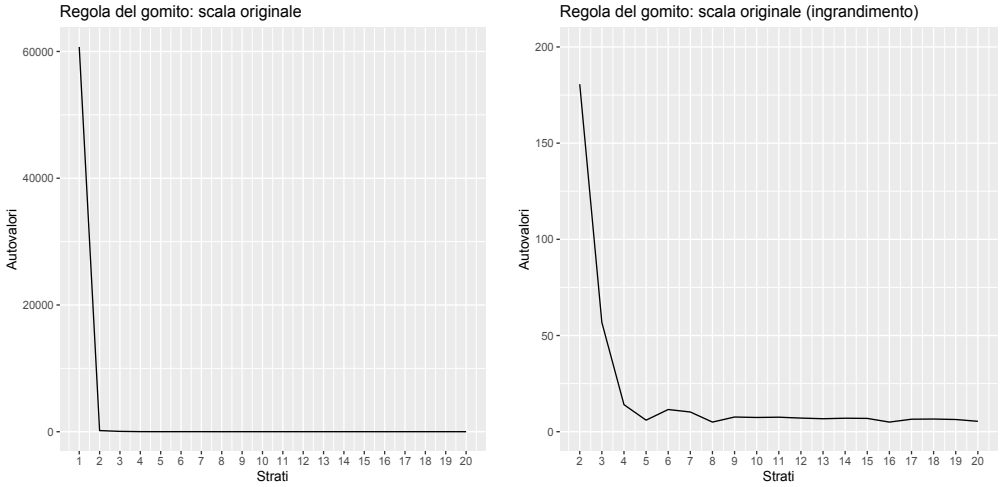

Figura A.22: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione C1. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.991	0.867	22.333	6
SSLB	1.000	0.467	24.8	1
SSVD	0.995	0.933	0.200	3

Tabella A.5: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione C1*

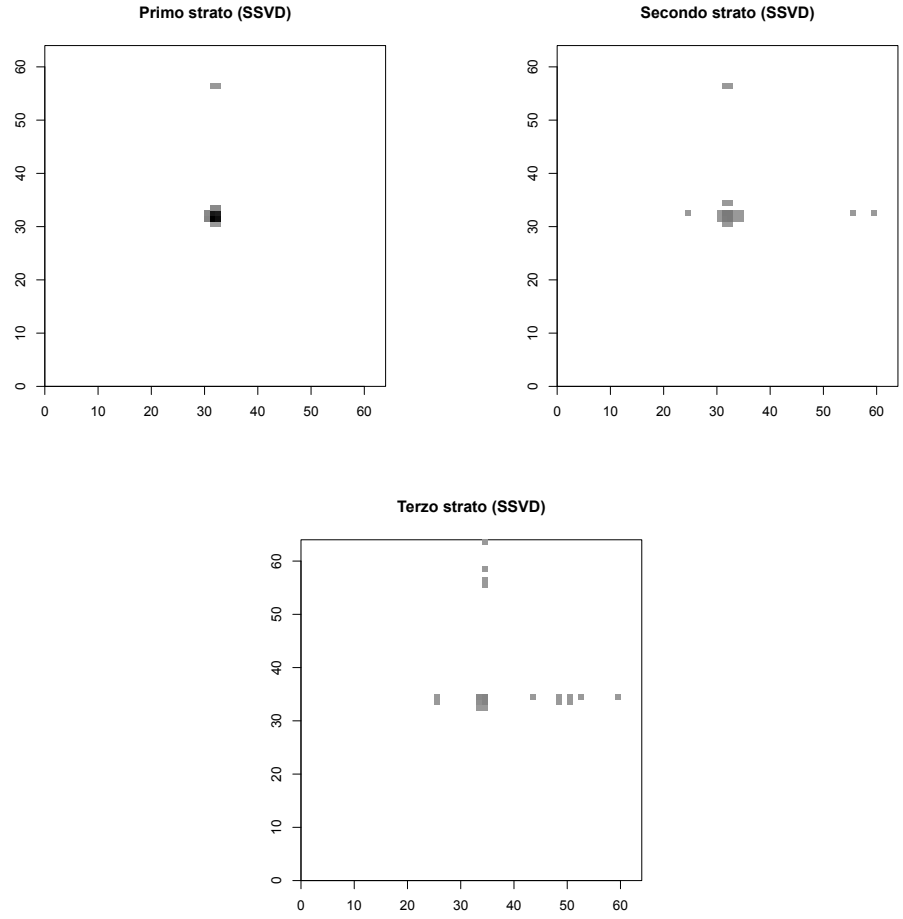

Figura A.23: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione C1, dal primo al terzo, rispettivamente a sinistra, in centro ed a destra. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

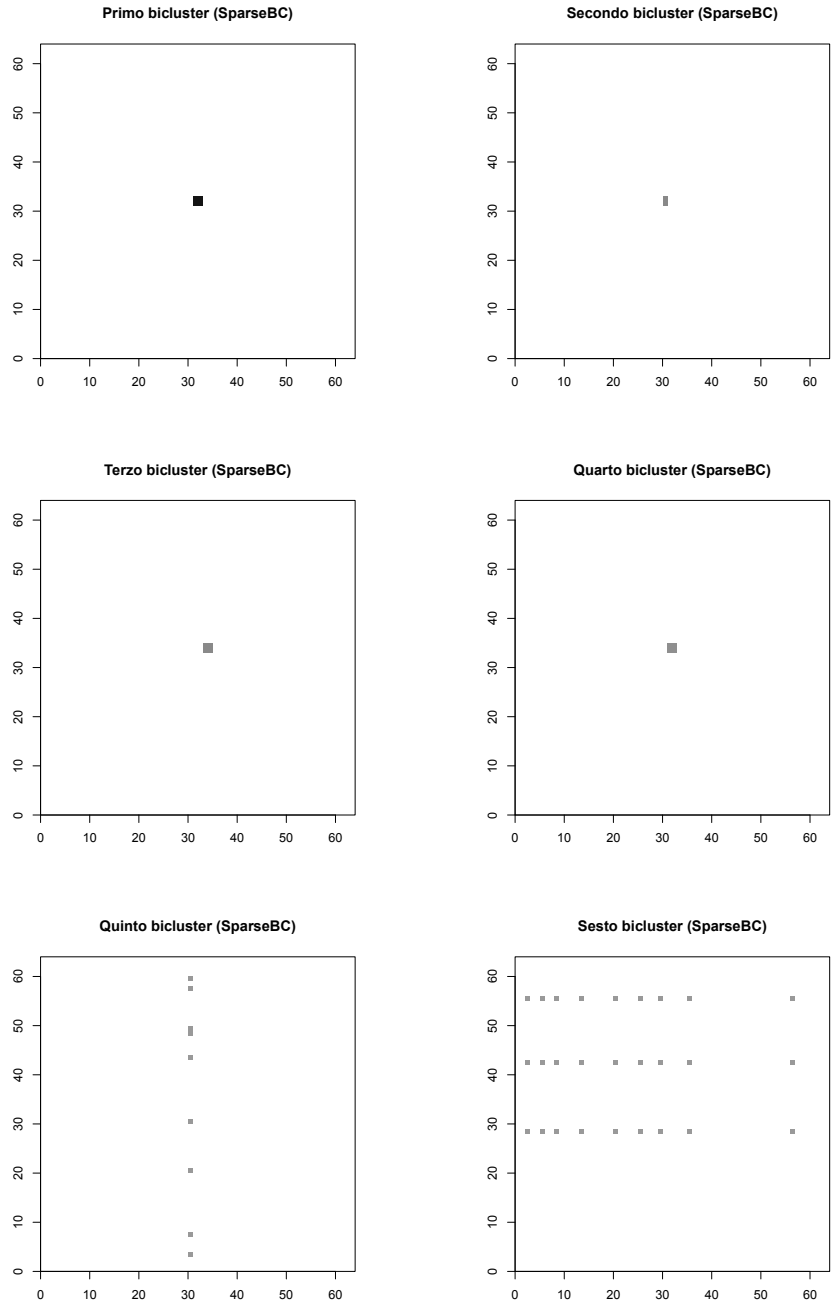

Figura A.24: *Rappresentazione in negativo della composizione dei biclu-ster, stimati da SparseBC per lo studio di simulazione C1, dal primo al sesto. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

A.6 Quasar con getto forte contiguo (C3)

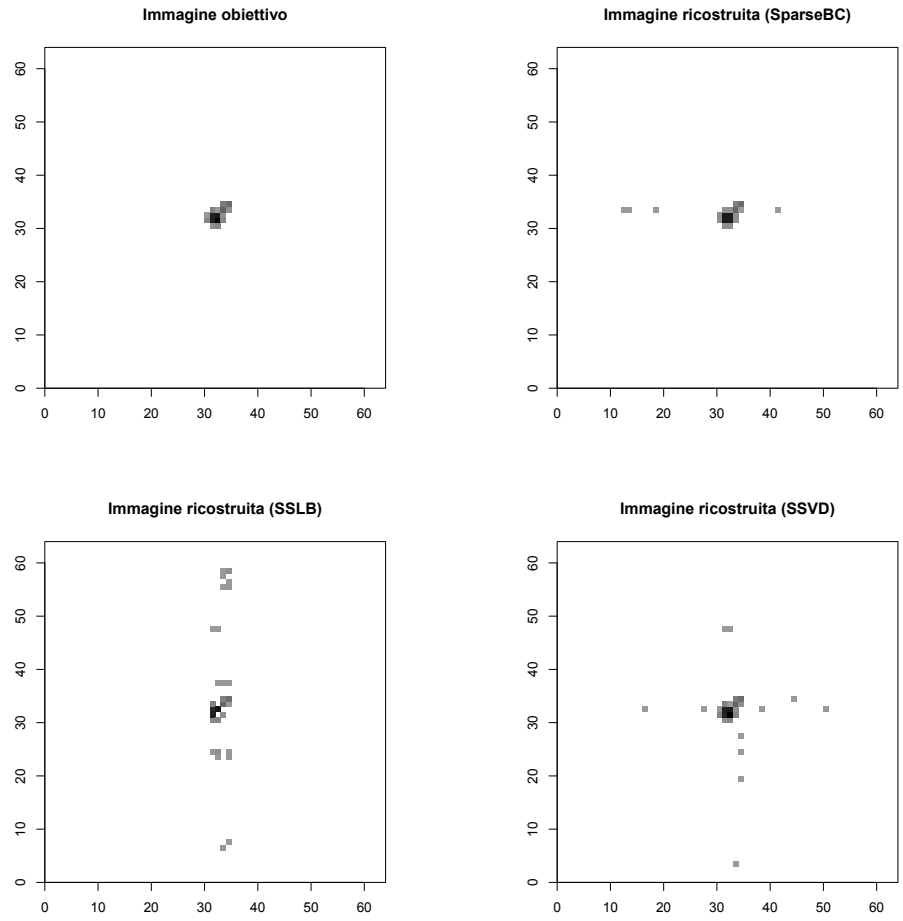

Figura A.25: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione C3 che considera un'immagine simulata composta dal quasar e da un getto di raggi X contiguo ad esso di luminosità forte e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

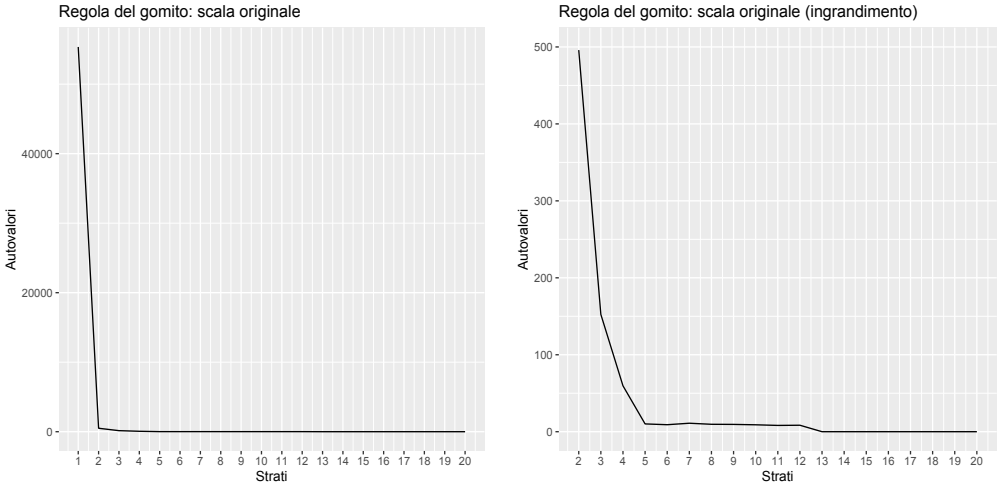

Figura A.26: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione C3. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.999	1.000	22.875	10
SSLB	0.995	0.875	8.188	10
SSVD	0.997	1.000	0.000	4

Tabella A.6: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione C3*

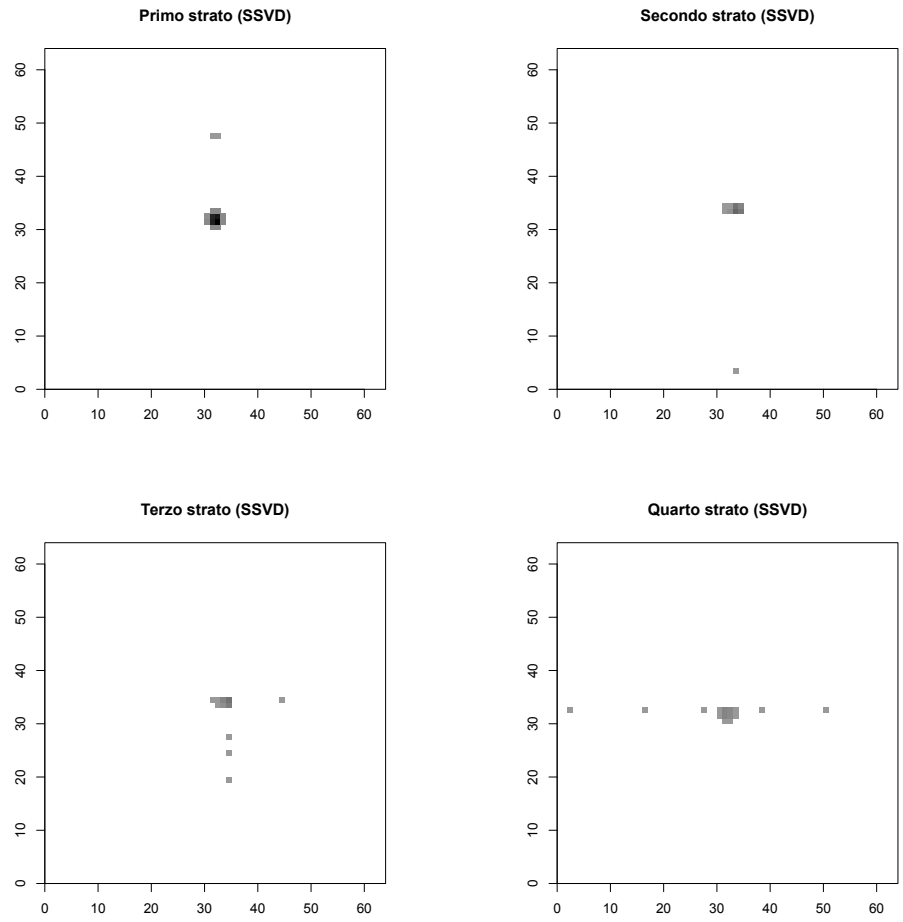

Figura A.27: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione C3, dal primo al quarto, rispettivamente in alto a sinistra, in alto a destra, in basso a sinistra e in basso a destra. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

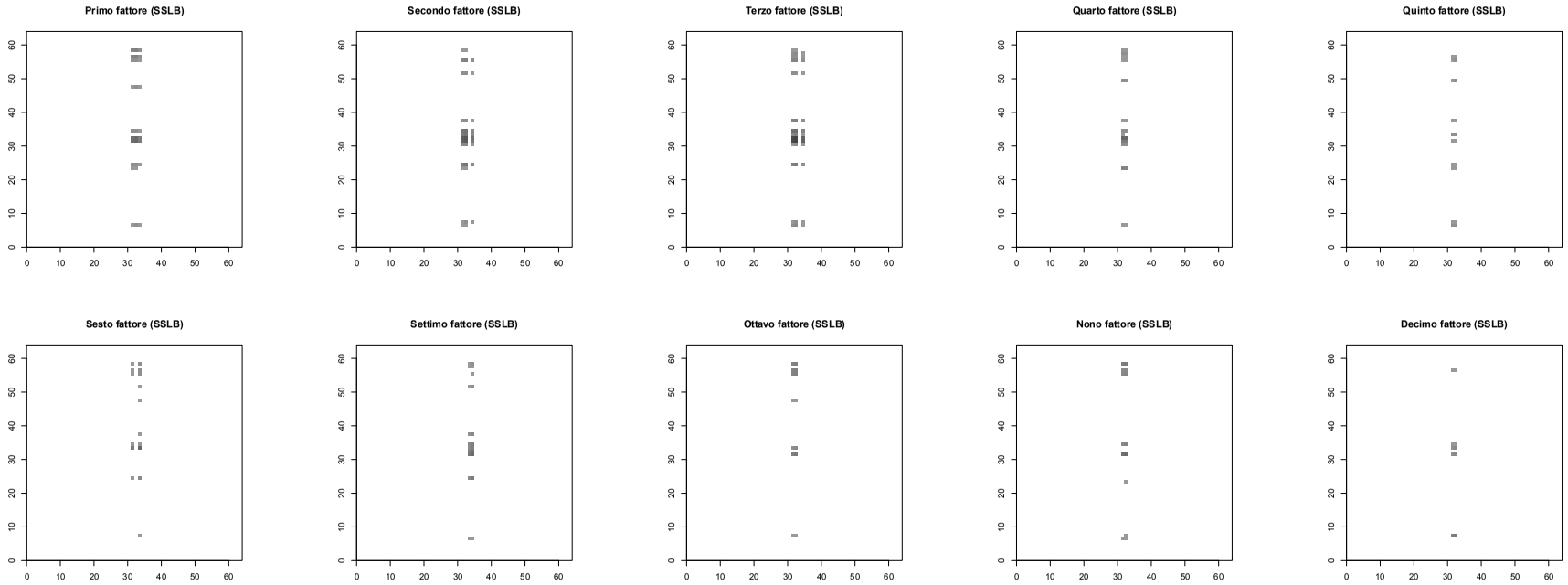

Figura A.28: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione C3, dal primo fattore al decimo. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

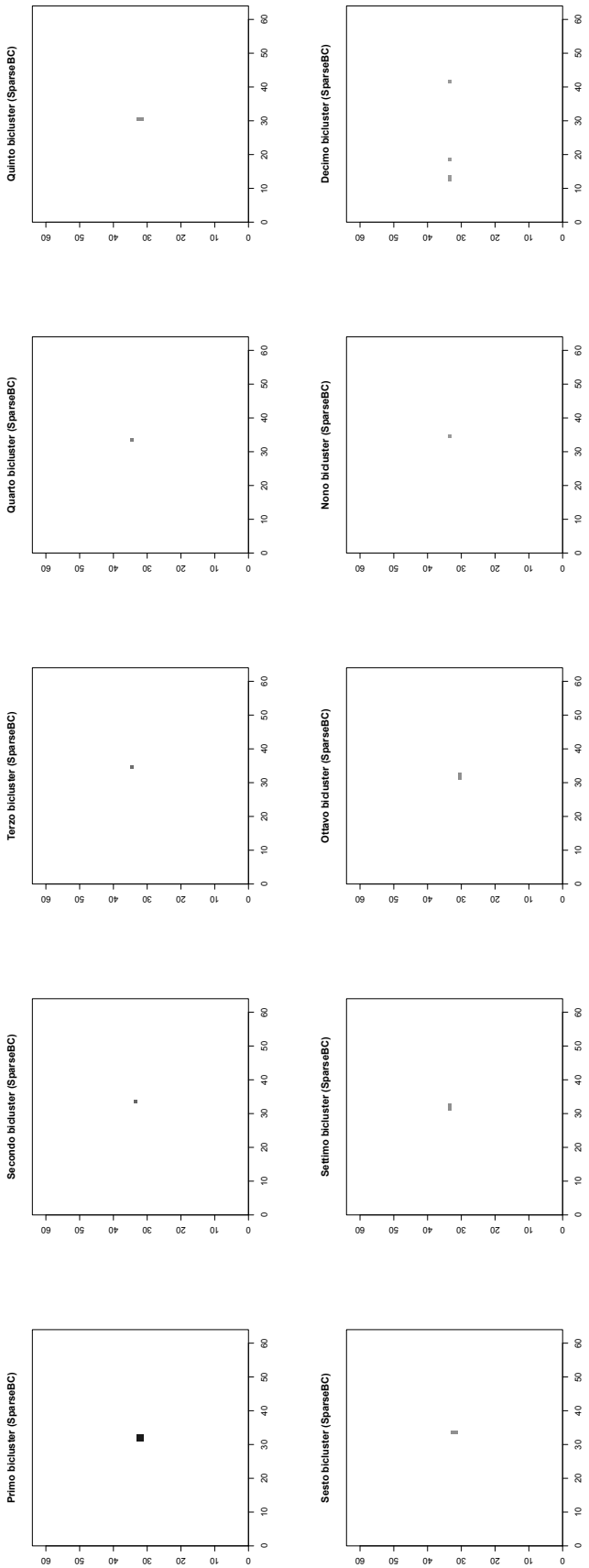

Figura A.29: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione C1, dal primo al sesto. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

A.7 Quasar con getto medio esteso (D2)

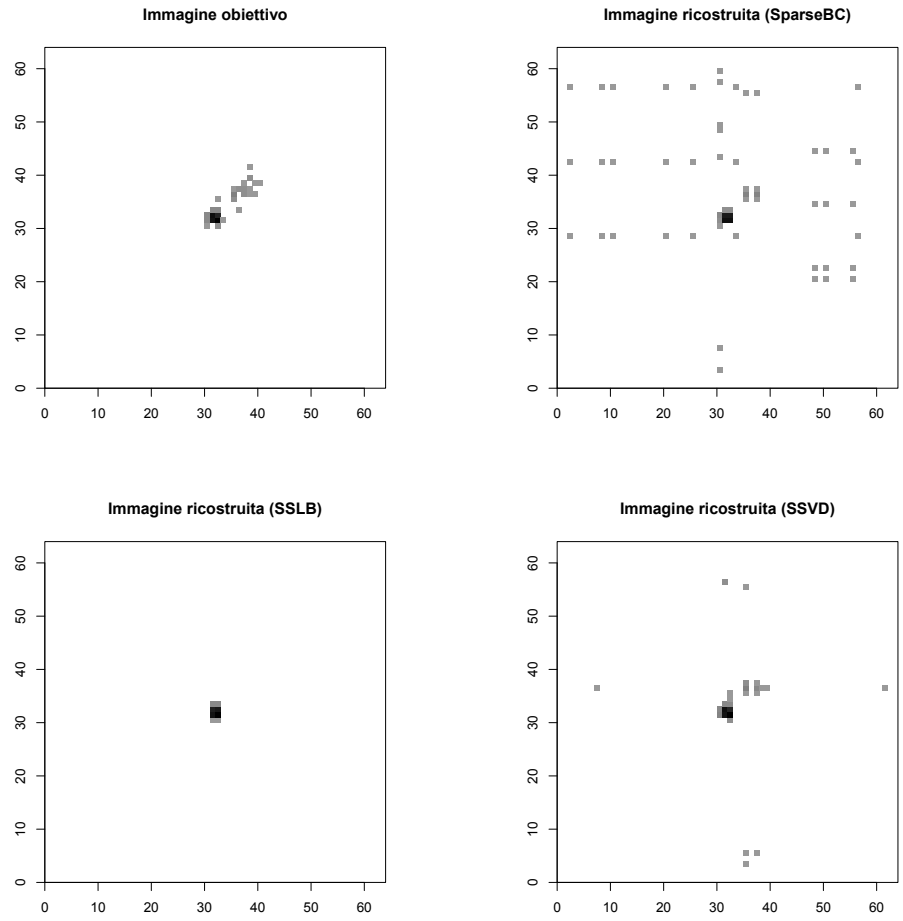

Figura A.30: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione D2, che considera un'immagine composta da un quasar e un getto di raggi X esteso di luminosità media, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

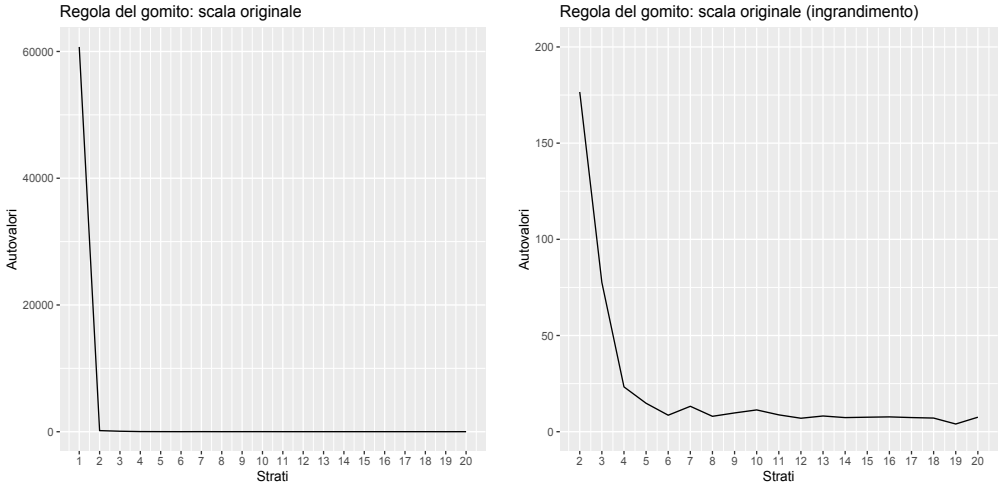

Figura A.31: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione D2. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.989	0.519	8.815	10
SSLB	1.000	0.259	14.407	2
SSVD	0.998	0.630	0.778	3

Tabella A.7: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione D2*

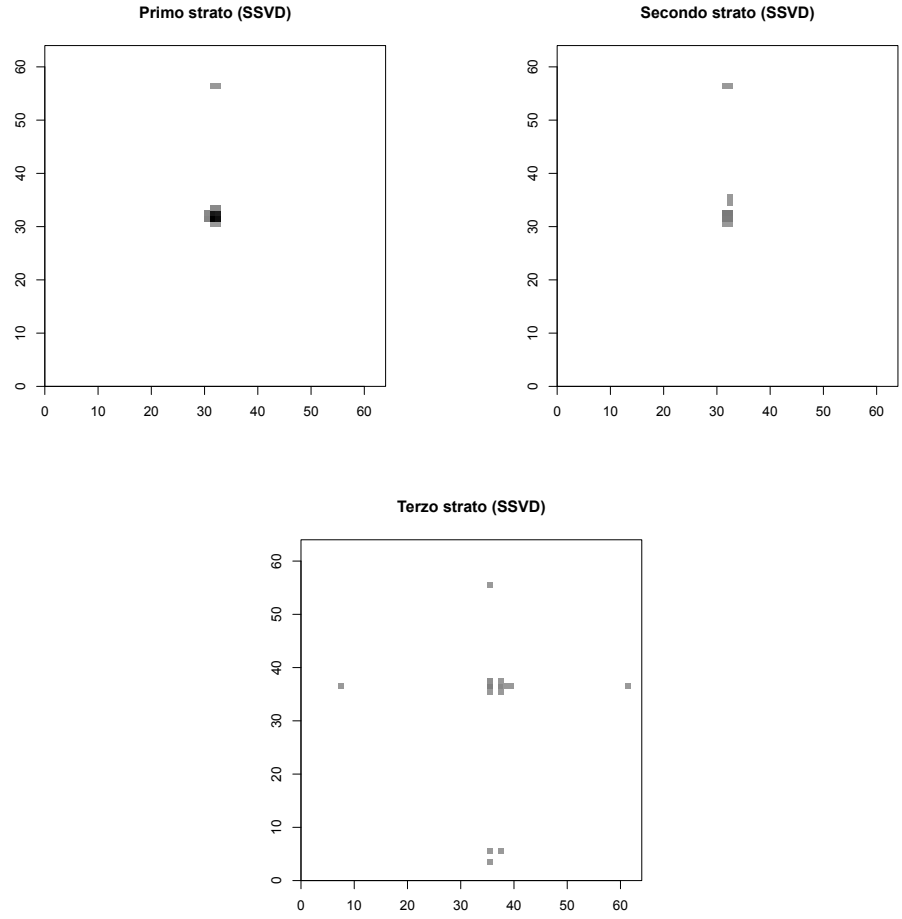

Figura A.32: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione D2, dal primo al terzo, rispettivamente a sinistra, in centro ed a destra. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

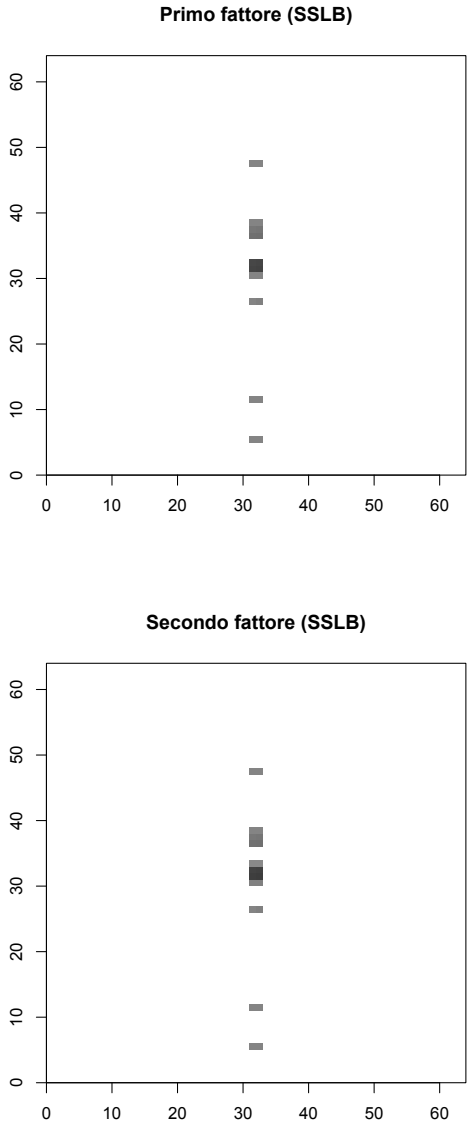

Figura A.33: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione D2, primo in alto e secondo in basso. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

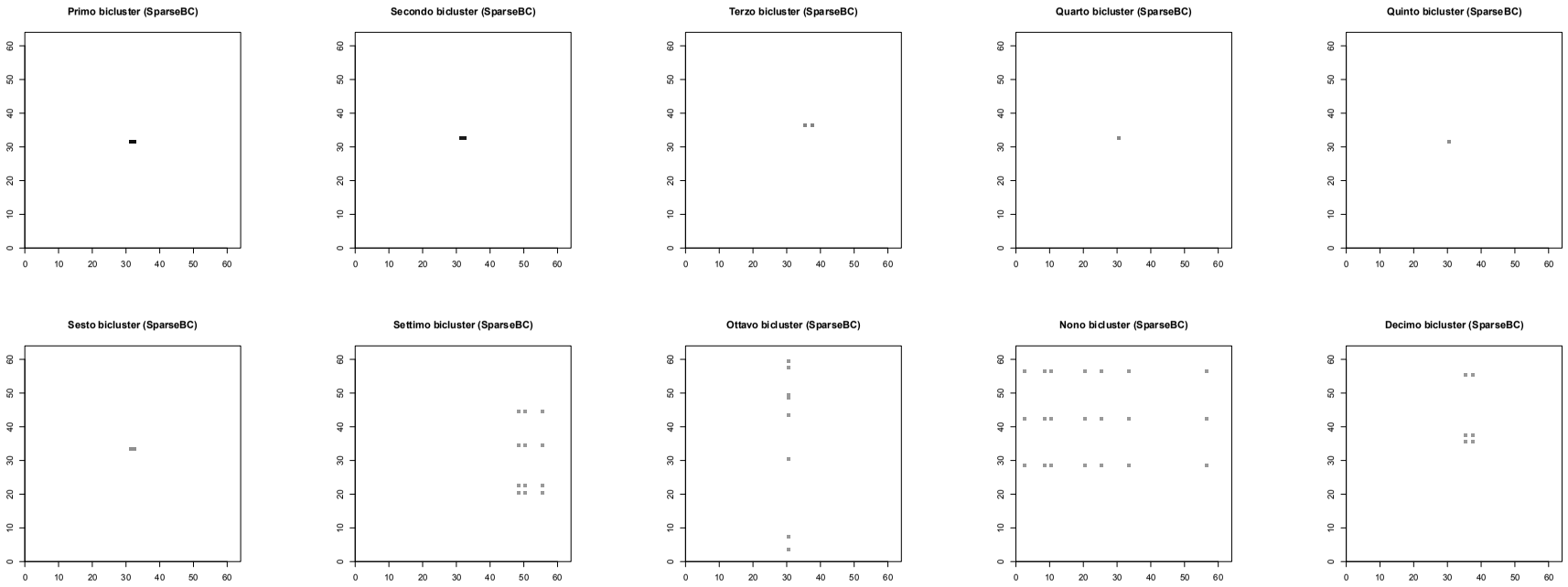

Figura A.34: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione D2, dal primo al decimo. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

A.8 Quasar con getto forte esteso (D3)

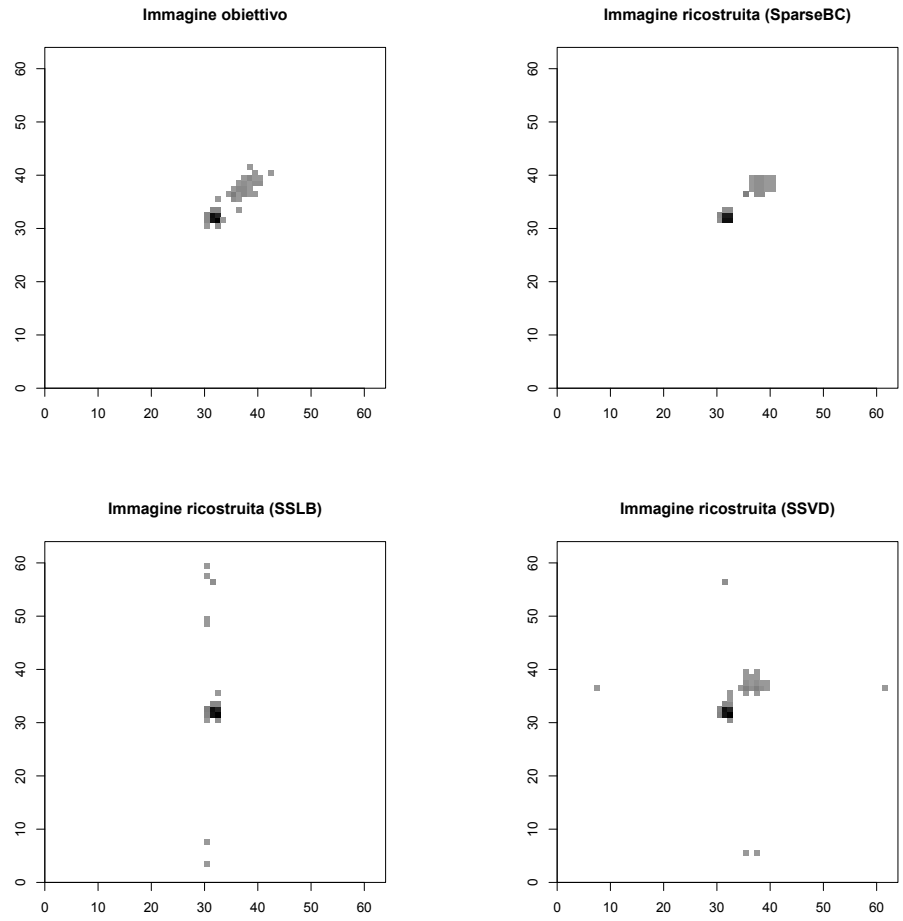

Figura A.35: *Rappresentazione in negativo dell'immagine ideale che vorremmo i modelli ritornassero (in alto a sinistra), per lo studio di simulazione D3, che considera un'immagine composta da un quasar e un getto di raggi X esteso di luminosità media, e rappresentazione in negativo delle immagini ricostruite dai metodi SparseBC (in alto a destra), SSLB (in basso a sinistra) e SSVD (in basso a destra).*

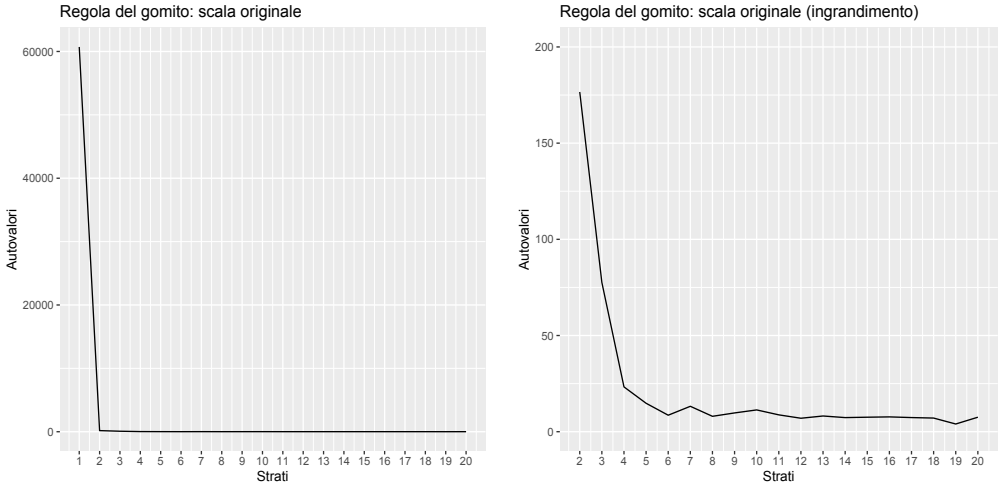

Figura A.36: *Grafico degli autovalori (in ordinata), al variare del numero di strati (ascissa) del modello SSVD, per lo studio di simulazione D3. A sinistra vengono rappresentati gli autovalori relativi ai primi* 20 *strati. A destra viene effettuato un ingrandimento per valutare il cambiamento di pendenza dal secondo strato in poi.*

	ι_0	ι_{not}	MSE	$\hat{K}$
sparseBC	0.999	0.622	6.568	9
SSLB	0.998	0.297	2.973	6
SSVD	0.998	0.649	1.162	3

Tabella A.8: *Percentuale di conteggi correttamente stimati pari a zero (prima colonna), percentuale di conteggi correttamente stimati diversi da zero (seconda colonna), errore quadratico medio tra conteggi non zero dell'immagine obiettivo e corrispondenti conteggi dell'immagine ricostruita (terza colonna) e numero di bicluster stimati (quarta colonna), per i tre metodi sparseBC, SSLB e SSVD, per lo studio di simulazione D3*

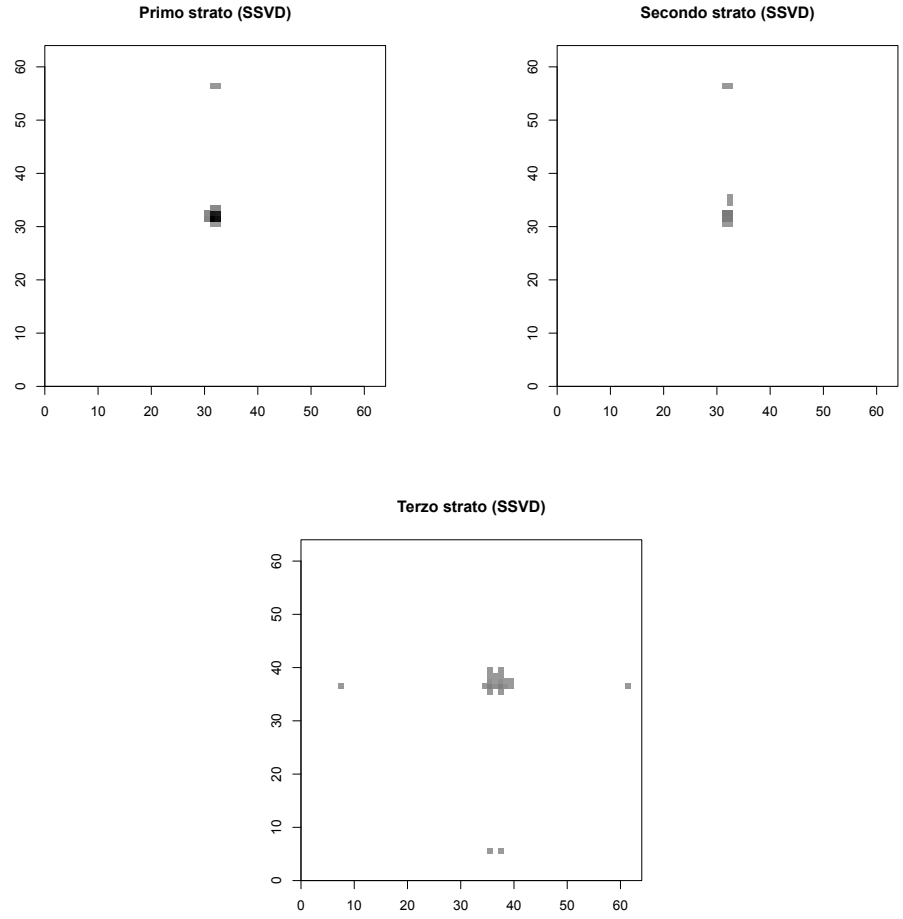

Figura A.37: *Rappresentazione in negativo della composizione degli strati o bicluster stimati da SSVD per lo studio di simulazione D3, dal primo al terzo, rispettivamente a sinistra, in centro ed a destra. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

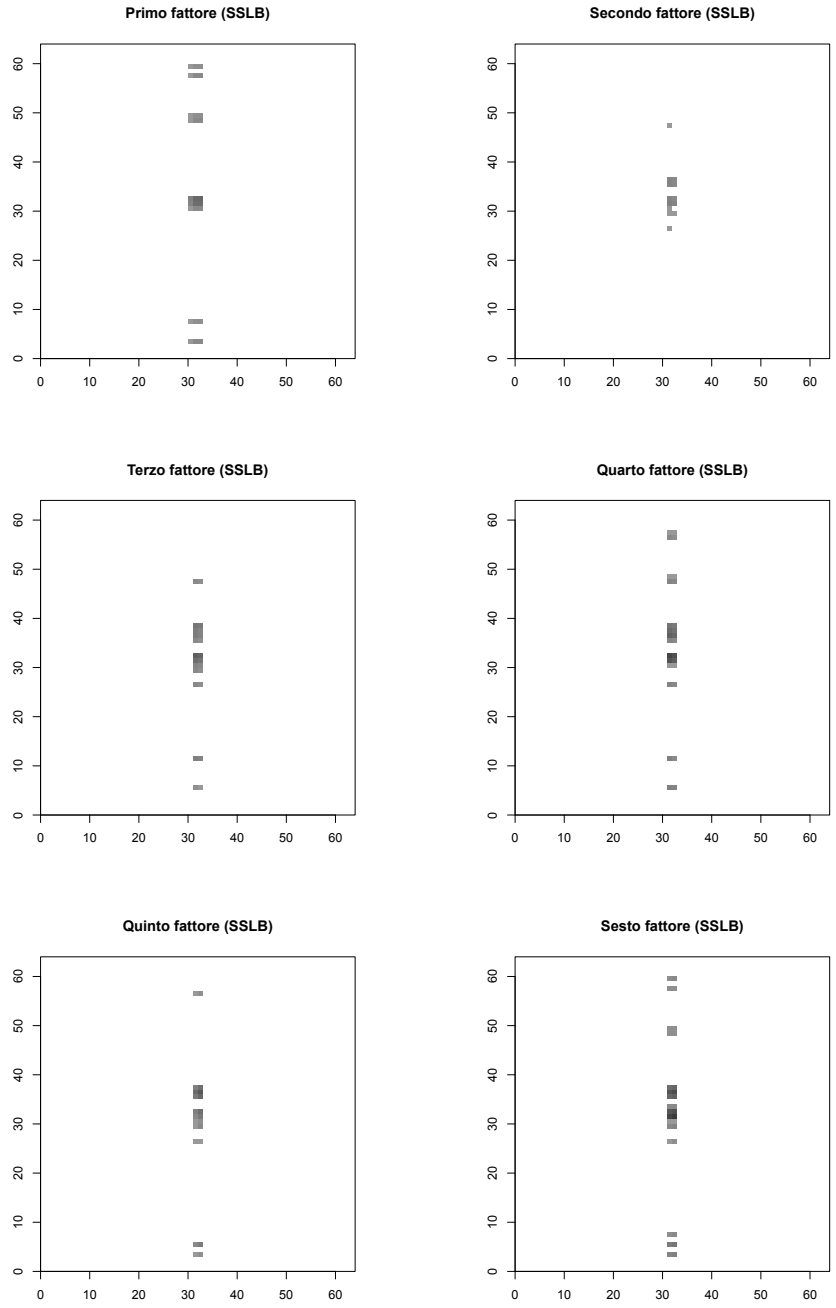

Figura A.38: *Rappresentazione in negativo della composizione dei fattori, o bicluster, stimati da SSLB per lo studio di simulazione D3, dal primo fattore al sesto. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

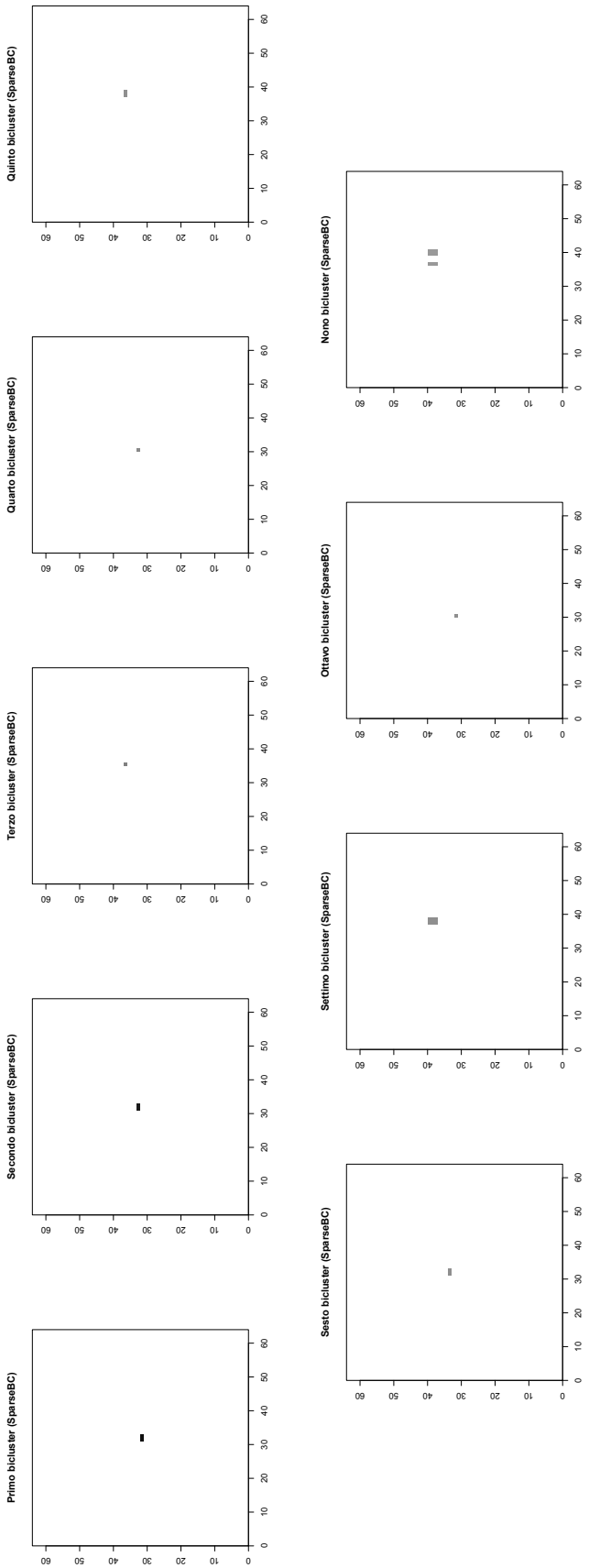

Figura A.39: *Rappresentazione in negativo della composizione dei bicluster, stimati da SparseBC per lo studio di simulazione D3, dal primo al nono. La gerarchia è data dalla media dei conteggi nei bicluster. Il colore bianco in corrispondenza di un pixel indica che quel pixel non compone il bicluster, più il colore si sposta verso il nero, più il logaritmo del conteggio stimato in corrispondenza del pixel è elevato per il bicluster.*

Appendice B

Listato in linguaggio R

```
#---Simulazione dell'immagine di 64*64 pixel

set.seed(1000) # per riproducibilità
library(mnormt)
x <- seq(0.5, 63.5, by = 1) # punti medi ascisse
y <- seq(0.5, 63.5, by = 1) # punti medi ordinate
f <- function(x, y, mu, var){ # funzione che calcola la densità
  dmnorm(c(x, y), mean = mu, varcov = var)
}

# 1. Simulazione conteggi quasar
mu_q <- c(32, 32)
var_q <- matrix(c(0.5^2,0,0, 0.5^2), byrow = T, nrow = 2)
fot_quasar <- 500
fv <- Vectorize(f, c('x', 'y'))
density_value <- outer(x, y, fv, mu = mu_q, var = var_q)
density_value_std <- density_value/sum(density_value)
E_pois <- density_value_std*fot_quasar
counts <- sapply(E_pois, function(x) rpois(1, x))

# matrice di conteggi simulati per il quasar

counts <- matrix(counts, nrow = 64)

# 2. Simulazione conteggi rumore di fondo
fot_rumore <- 200
counts_back <- matrix(rpois(4096, fot_rumore/4096), nrow = 64)
# matrice conteggi simulati rumore di fondo

# 3. Simulazione conteggi rumore di fondo
fot_getto <- 20
mu_g <- c(34,34)
var_g <- matrix(c(0.5^2,0.5^3,0.5^3, 0.5^2), byrow = T, nrow = 2)
density_value_g <- outer(x1, x2, fv, mu = mu_g, var = var_g)
density_value_std_g <- density_value_g/sum(density_value_g)
E_pois_g<- density_value_std_g*fot_getto
```

```
counts_jet <- sapply(E_pois_g, function(x) rpois(1, x),
                     simplify = "array")
# matrice dei conteggi getto
counts_jet <- matrix(counts_jet, nrow = 64)

# 4. Immagine simulata
counts.tot <- counts + counts_back + counts_jet # conteggi totali

# 5. Visualizzazione
image(0:64, 0:64, counts.tot, zlim = c(0, max(counts.tot)),
    main = "Quasar più getto medio vicino",
    ylab = "", xlab = "", col = gray.colors(200, 0, 1, rev = F))

# 6. Conteggi immagine obiettivo
E_counts <- counts + counts_jet

#---Applicazione del modello SSLB

# 1. Funzione per trovare il miglior adattamento
best_fit_SSLB <- function(N = 100, counts.tot, K_init = 50, ...){
  require(SSLB)
  results <- vector("list", N)
  dime <-rep(0, N)
  for(i in 1:N){
    results[[i]] <-  SSLB(counts.tot, K_init = K_init,...)
    dime[i] <-  dim(results[[i]]$X)[2]
  }
  idx <- which(table(dime) == max(table(dime)))
  if(length(idx) > 1){
    Kstar <- min(as.numeric(names(idx)))
  } else {
    Kstar <- as.numeric(names(idx))
  }
  out <- list(Kstar = Kstar, dime = dime, results = results)
  out
}

# 2. Applicazione algoritmo e scelta miglior adattamento
N <- 100
res_sslb <- best_fit_SSLB(N = N, counts.tot = counts.tot,
                          MAX_ITER = 5000)
```

```
# 3. Funzione per calcolare il BIC

bic.SSLB <- function(results, counts.tot, n_obs){
  require(mvtnorm)
  Ys <- list()
  for(i in 1:length(results)){
    Ys[[i]] <- results[[i]]$X%*%t(results[[i]]$B)
  } # matrici stimate per ciascun modello
  sig <- list()
  for(i in 1:length(results)){
    sig[[i]] <- results[[i]]$path$sigmas[,ncol(results[[i]]$path$sigmas)]
  }
  n <- nrow(Ys[[1]])
  bic <- NULL
  su = 0
  for(j in 1:length(results)){
    su = 0
    for(i in 1:n){
      su <- su +dmvnorm(counts.tot[i,],
                        mean = Ys[[j]][i,],
                        sigma = diag(sig[[j]]), log = T)
    }
    curr_bic <- -2*su + (sum(results[[j]]$X != 0) +
                      sum(results[[j]]$B != 0))*log(n_obs)
    bic <- c(bic,curr_bic)
  }
  bic
}


# 4. Scelta miglior adattamento
n_obs <- 64^2
bic <- bic.SSLB(res_sslb$results, counts.tot, n_obs)
best <- which(bic == min(bic))
n.bicluster <- ncol(res_sslb$results[[best]]$X)
Y_hat_sslb <- round(res_sslb$results[[best]]$X%*%
                    t(res_sslb$results[[best]]$B))

# controllo conteggi negativi
neg <- which(Y_hat_sslb < 0)
# se presenti sostituirli con 0
Y_hat_sslb[neg] <- 0

# 5. Visualizzazione immagine ricostruita SSLB (in negativo)
image(0:64, 0:64, log(Y_hat_sslb),
      zlim = c(0, max(log(counts.tot))), ylab = "", xlab = "",
      main = "immagine ricostruita (SSLB)",
      col = gray.colors(200, 0, 0.6, rev = T))
```

```
# per ottenere, per esempio, il primo fattore
fat1 <- round(abs(res_sslb$results[[best]]$X[,1]%*%
              t(res_sslb$results[[best]]$B[,1])))

# 6. Visualizzazione composizione fattori (in negativo)
image(0:64, 0:64, log(fat1),
      zlim = c(0, max(log(counts.tot))), ylab = "", xlab = "",
      main = "Primo fattore (SSLB)",
      col = gray.colors(200, 0, 0.6, rev = T))

#---Applcazione modello SparseBC
# 1. Calcolo verosimiglianza del modello, prova deve essere
#   un oggetto di classe sparseBC
logv_sparseBC <- function(prova, lambda = 1, counts.tot){
  ng <- table(prova$Cs)
  pr <- table(prova$Ds)
  G = length(ng)
  R = length(pr)
  mu <- prova$Mus
  sigma2.num <- sigma2.den <- matrix(0, G, R)
  for(g in 1:G){
    for(r in 1:R){
      sigma2.num[g, r] <-
      sum((counts.tot[prova$Cs == g, prova$Ds == r] - mu[g,r])^2)
      sigma2.den[g, r] <- ng[g] * pr[r]
    }
  }
  sigma2 <- sum(sigma2.num)/sum(sigma2.den)
  logL <- matrix(0, G, R)
  for(g in 1:G){
    for(r in 1:R){
      logL[g, r] <- sum(dnorm(counts.tot[prova$Cs == g, prova$Ds == r],
                    mean = mu[g, r], sd = sqrt(sigma2), log = T))
    }
  }
  logL <- sum(logL) - lambda * sum(abs(mu))
  logL
}
```

```
# 3. Calcolo della verosimiglianza per una griglia di G, R e lambda
choose_g_r_lambda <- function(N,G, R, lambda, counts.tot){
  require(sparseBC)
  ng <- length(G)
  nr <- length(R)
  nl <- length(lambda)
  out <- array( dim = c(N, ng, nr, nl))
  for(i in 1:N){
    for(g in 1:ng){
      for(r in 1:nr){
        for(l in 1:nl){
          ris <- sparseBC(counts.tot, G[g], R[r], lambda = lambda[l])
          out[i,g,r,l] <- logv_sparseBC(ris, lambda = lambda[l], counts.tot)
        }
      }
    }
  }
  out
}

# 4. Scelta del miglior adattamento
set.seed(12333)
N = 10
lambda = 1:4
G = 1:10
R = 1:10
ng = length(G)
nr = length(R)
nl = length(lambda)
sim7_gr <- choose_g_r_lambda(N, G = G, R = R,
           lambda, counts.tot)

# calcolo il BIC corrispondente ad ogni modello
bic <- array(0, dim(sim7_gr))
for(i in 1:10){
  for(g in 1:ng){
    for(r in 1:nr){
      for(l in 1:nl){
        bic[i,g,r,l] <- -2*sim7_gr[i,g,r,l] + (g*r)*log(64^2)
      }
    }
  }
}
idx <- which(bic ==min(bic), arr.ind = T)
set.seed(1234)
final_fit_BC <- sparseBC(counts.tot, k = idx[2], r = idx[3],
                 lambda = idx[4])
Y_hat_BC <- round(final_fit_BC$mus)
```

```
# 5. Visualizzazione immagine ricostruita sparseBC (in negativo)
image(0:64, 0:64, log(Y_hat_BC),
      zlim = c(0, max(log(counts.tot))), ylab = "", xlab = "",
      main = "immagine ricostruita (SparseBC)",
      col = gray.colors(200, 0, 0.6, rev = T))

# 6. Visualizzazione composizione primo bicluster (in negativo)
# indici del primo bicluster
idx1 <- which(round(final_fit_BC$Mus) == max(round(final_fit_BC$Mus)),
               arr.ind = T)
bicluster1<- round(final_fit_BC$mus)
bicluster1[final_fit_BC$Cs != idx1[1],] <- 0
bicluster1[,final_fit_BC$Ds != isx1[2]] <- 0
bicluster1[sparse_sim7$Cs == idx1[1], sparse_sim7$Ds == idx1[2]] <-
          round(sparse_sim7$Mus[idx1[1],idx1[2]])
image(0:64, 0:64, log(bicluster1),
      zlim = c(0, max(log(counts.tot))), ylab = "", xlab = "",
      main = "Primo bicluster (SparseBC)",
      col = gray.colors(200, 0, 0.6, rev = T))

#---Applicazione modello SSVD

# 1. Scelta del numero di strati
c <- counts.tot # salviamo temporaneamente i conteggi simulati in c
library(s4vd)
y_hat <- 0
n_strati <- 20
s <- rep(0, n_strati)
for(k in 1:(n_strati)){
  set.seed(1234)
  results_SSVD <- biclust(counts.tot, method=BCssvd(),
                          K=1,
                          threu = 1,
                          threv = 1,
                          gamu = 2,
                          gamv = 2,
                          niter = 100)
  curr_u <- results_SSVD@info$res[[1]]$u
  curr_v <- results_SSVD@info$res[[1]]$v
  s[k] <- t(curr_u)%*%counts.tot%*%curr_v
  y_hat <- s[k]*(curr_u)%*%t(curr_v)
  counts.tot <- round(counts.tot - y_hat) # matrice dei conteggi residua
  counts.tot[which(counts.tot < 0)] <- 0
}

# Grafico per la regola del gomito
library(ggplot2)
plot.data <- data.frame(cbind(1:(n_strati), s^2))
colnames(plot.data) <- c("strati", "autovalori")
```

```
# grafico classico
ggplot(plot.data, aes(strati, autovalori)) +
  geom_line()+
  labs(title = "regola del gomito: scala originale",
  x = "Strati", y = "Autovalori") +
  scale_x_continuous( n.breaks = 20)
# grafico ingrandito
ggplot(plot.data, aes(strati, autovalori)) +
  geom_line()+
  labs(title = "regola del gomito: scala originale (ingrandimento)",
  x = "Strati", y = "Autovalori")+
  scale_x_continuous(limits = c(2,20), n.breaks = 20)+
  scale_y_continuous(limits = c(0,200))

# 2. Stima del modello con numero di strati scelto precedentemente,
#    in questo caso 2
counts.tot <- c # ripristino i conteggi
set.seed(1234)
results_SSVD_final <- biclust(counts.tot, method=BCssvd(),
                        K=2,
                        threu = 1,
                        threv = 1,
                        gamu = 2,
                        gamv = 2,
                        niter = 100)
s <- rep(0,2)
for(i in 1:2){
  s_final[i] <- t(results_SSVD_final@info$res[[i]]$u)%*%
                 counts.tot%*%results_SSVD_final@info$res[[i]]$v
}
Y_hat_ssvd <- 0
for(i in 1:2){
  Y_hat_ssvd <- Y_hat_ssvd + s_final[i]*
                results_SSVD_final@info$res[[i]]$u%*%
                t(results_SSVD_final@info$res[[i]]$v)
}

# controllo per conteggi negativi
sum(Y_hat_ssvd <0 )
mean(Y_hat_ssvd < 0)
min(Y_hat_ssvd[Y_hat_ssvd <0])
Y_hat_ssvd[Y_hat_ssvd <0] <- 0

# 3. Visualizzazione immagine ricostruita SSVD (in negativo)
image(0:64, 0:64, log(round(Y_hat_ssvd)),
       zlim = c(0, max(log(counts.tot))), , xlab = "", ylab = "",
       main = "immagine ricostruita (SSVD)",
       col = gray.colors(200, 0, 0.6, rev = T))
```

```
# 4. Visualizzazione composizione primo strato (in negativo)
str1 <- abs(round(s[1]*results_SSVD@info$res[[1]]$u%*%
            t(results_SSVD@info$res[[1]]$v)))
image(0:64, 0:64, log(str1),
      zlim = c(0, max(log(counts.tot))), xlab = "", ylab = "",
      main = "Primo strato (SSVD)",
      col = gray.colors(200, 0, 0.6, rev = T))

#---Indici per valutare la ricostruzione dell'immagine
E_counts <- counts.tot + counts_jet # matrice obiettivo

# funzione per ottenere gli indici
perc_zeros_not_mse <- function(Y_hat, ideal_counts){
  idx <- which(ideal_counts == 0)
  out <- list(perc_zeros = mean(Y_hat[idx] == 0),
              perc_not = mean(Y_hat[-idx] != 0),
              mse = sum((Y_hat[-idx] - ideal_counts[-idx])^2)/
              ((dim(Y_hat)[1]*dim(Y_hat)[2]) - length(idx)))
  out
}
ind_sslb <- perc_zeros_not_mse(Y_hat_sslb, E_counts)
ind_ssvd <- perc_zeros_not_mse(Y_hat_ssvd, E_counts)
ind_BC <- perc_zeros_not_mse(Y_hat_BC, E_counts)
```

Bibliografia

[1] Davison (2003). Statistical Models (Cambridge Series in Statistical and Probabilistic Mathematics). *Cambridge University Press*. doi:10.1017/CBO9780511815850.

[2] Director et al. (2022). Contour models for physical boundaries enclosing star-shaped and approximately star-shaped polygons, *Journal of the Royal Statistical Society Series C*, Royal Statistical Society, **71(5)**:1688-1720. doi: 10.1111/rssc.12592.

[3] Esch et al. (2004). An Image Restoration Technique with Error Estimates. *The Astrophysical Journal*, **610**:1213 - 1227. doi: 10.1086/421761.

[4] Friston et al. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, **2**:189-210. doi: 10.1002/hbm.460020402.

[5] Griffiths et al. (2011). The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, **12(32)**:1185-1224. doi: 10.5555/1953048.2021039.

[6] Lee et al. (2010). Biclustering via sparse singular value decomposition. *Biometrics*. **66(4)**:1087-1095. doi: 10.1111/j.1541-0420.2010.01392.x.

[7] McKeough et al. (2016). Detecting Relativistic X-ray Jets in High-Redshift Quasars. *The Astrophysical Journal*, **833**:123-144. doi: 10.3847/1538-4357/833/1/123.

[8] Moran et al. (2021). Spike-and-slab Lasso biclustering. *The Annals of Applied Statistics*, **15(1)**:148-173. doi: 10.1214/20-AOAS1385.

[9] Pavan et al. (2011). IGRJ11014-6103: a newly discovered pulsar wind nebula? *Proceedings of The Extreme and Variable High Energy Sky-PoS(Extremesky 2011)*. doi: 10.22323/1.147.0003.

[10] Ročková et al. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, **113(521)**: 431-444. doi: 10.1080/01621459.2016.1260469.

[11] Stein et al. (2015). Detecting unspecified structure in low-count images. *The Astrophysical Journal*, **813**:66-81. doi: 10.1088/0004-637X/813/1/66.

[12] Tan et al. (2014). Sparse Biclustering of Transposable Data. *Journal of Computational Graphical Statistics*, **23(4)**: 985-1008. doi:10.1080/10618600.2013.852554.

[13] Tibshirani (1996). Regression Shirnkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58(1)**:267-288.

[14] Zou (2006). The adaptive LASSO ad its oracle properties. *Journal of the American Statistical Association*. **101**:1418-1429. doi: 10.1198/016214506000000735.

[15] Zou et al. (2007). On the degrees of freedom of the LASSO. *The Annals of Statistics*. **35(5)**: 2173-2192. doi: 10.1214/009053607000000127.

Sitografia

[1] Moran (2021). *SSLB: Spike and Slab Lasso Biclustering.* R package version 1.0, https://projecteuclid.org/journals/annals-of-applied-statistics/volume-15/issue-1/Spike-and-slab-Lasso-biclustering/10.1214/20-AOAS1385.full.

[2] Sill et al. (2015). *s4vd: Biclustering via Sparse Singular Value Decomposition Incorporating Stability Selection.* R package version 1.1-1, https://CRAN.R-project.org/package=s4vd.

[3] Tan (2019). *sparseBC: Sparse Biclustering of Transposable Data.* R package version 1.2, https://CRAN.R-project.org/package=sparseBC.