POLITECNICO DI TORINO

Laurea magistrale in Ingegneria Matematica

Tesi magistrale

Inferenza bayesiana per l'inverse modeling dello spettro atmosferico di esopianeti

Supervisori

Prof. Flavio GIOBERGIA

Prof. Alkis KOUDOUNAS

Candidato

Alessia TORTONE

2025

Sommario

L'Agenzia Spaziale Europea (ESA) ha organizzato una challenge internazionale ad agosto 2024 con l'obiettivo di individuare transiti planetari. L'analisi si basa sullo studio dei segnali provenienti da stelle vicine al nostro sistema solare. I dati utilizzati per la challenge non provengono da osservazioni reali, ma sono stati generati tramite simulazioni, le quali riproducono il comportamento dei due sensori che saranno presenti sul telescopio spaziale ARIEL, che nel 2029 condurrà il primo studio approfondito su un campione di 1.000 pianeti extrasolari. L'obiettivo principale della challenge è individuare il transito del pianeta, cioè il momento in cui il pianeta passa davanti alla sua stella, oscurandone parzialmente la luce, e stimare le componenti del suo spettro atmosferico. Tale studio permetterà in un secondo momento di poter dedurre la composizione chimica degli esopianeti e di stimarne caratteristiche specifiche, quali temperatura e pressione. Questa tesi propone una soluzione al problema basata su tre fasi principali: (1) preprocessing del segnale, che prevede la rimozione dei pixel difettosi nei dati simulati e la correzione degli errori di lettura generati dai sensori della luce, (2) Inferenza Bayesiana e spettroscopia di trasmissione, che prevede l'applicazione di tecniche di inferenza bayesiana, in particolare la stima del Massimo A Posteriori (MAP), per stimare la profondità di transito per ciascuna lunghezza d'onda di interesse, e di conseguenza dedurre le componenti dello spettro atmosferico del pianeta preso in analisi, (3) training del modello e correzione finale, che prevede l'addestramento del modello sul dataset fornito per apprendere le caratteristiche del segnale, e una successiva correzione lineare ottimizzata attraverso la discesa del gradiente, per migliorare l'accuratezza della stima finale.

Questa metodologia ha permesso di ottenere risultati affidabili nella rilevazione dei transiti e nell'analisi dello spettro atmosferico dei pianeti extrasolari simulati.

Table of Contents

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Nel 2029 l'ESA, Agenzia Spaziale Europea, condurrà in orbita il telescopio spaziale Ariel, il cui compito sarà quello di studiare esopianeti appartenenti a sistemi stellari in prossimità del nostro sistema solare. Ad oggi, sono stati individuati oltre 5.600 esopianeti. Tuttavia, il rilevamento di questi mondi è solo il primo passo: è fondamentale anche caratterizzarne la natura, studiandone la composizione delle atmosfere. In particolare, la missione Ariel dell'ESA condurrà il primo studio completo su 1.000 nuovi esopianeti in un'area della galassia vicina alla nostra.

Il metodo per l'individuazione dei pianeti consiste nell'analisi dello spettro luminoso della stella: quando i pianeti transitano fra noi osservatori e la stella, la luminosità di quest'ultima presenta un piccolo calo nel suo spettro. Individuare questa fase di transito risulta essere proprio il primo obiettivo della missione, in quanto permette una prima identificazione di nuovi pianeti. L'altro importante passaggio è quello di osservare lo spettro luminoso della stella suddiviso per lunghezze d'onda, al fine di comprendere la composizione atmosferica dei pianeti osservati. Durante il transito, una minuscola frazione della luce stellare (tra 50 e 200 fotoni per milione) attraversa l'atmosfera del pianeta, interagendo con gli elementi in essa presenti. Tuttavia, il segnale risultante può essere facilmente corrotto da disturbi strumentali e dal rumore di lettura. La variazione fotometrica generata nel segnale da tale rumore è paragonabile alla variazione prodotta dal transito del pianeta stesso, rendendo complessa la rilevazione di pianeti di piccole dimensioni, come le Super-Terre e i pianeti simili alla Terra.

Per affrontare queste difficoltà, l'ESA ha indetto una challenge internazionale (che ha visto la sua conclusione lo scorso novembre 2024), il cui obiettivo è quello di trovare un modo per stimare la fase di transito del pianeta, lo spettro atmosferico e la relativa incertezza associata. Più nello specifico, sono state generate artificialmente per ogni pianeta immagini 2D sequenziali del piano focale spettrale, scattate nel corso di diverse ore di osservazione dell'esopianeta mentre eclissa la sua stella. Per ottenere le immagini sono stati utilizzati due diversi strumenti di osservazione:

AIRS-CH0 e FGS1. Una parte dei dati dei pianeti simulati viene fornita come dataset di training, accompagnata dai rispettivi valori di ground truth. Altri pianeti, invece, vengono utilizzati per valutare le prestazioni delle diverse soluzioni, formando un dataset di test che non è stato reso disponibile ai singoli partecipanti alla challenge. Abbiamo visto, dunque, come uno dei requisiti principali della competizione sia proprio quello di rimuovere il rumore introdotto nel corso delle osservazioni. Il processo di "detrending" permette di eliminare le distorsioni e ottenere dati puliti per poter fare un'analisi scientifica accurata e una stima precisa degli spettri atmosferici cercati. La corretta estrazione degli spettri atmosferici rappresenta un traguardo essenziale per comprendere la natura e la composizione degli esopianeti, avvicinandoci sempre di più alla possibilità di individuare ambienti favorevoli alla vita al di fuori del nostro sistema solare.

La soluzione proposta per la challenge presentata in questo progetto si articola in diverse fasi:

1. Preprocessing: Viene inizialmente applicata una fase di preprocessing delle immagini dello spettro stellare per ripulire il segnale, correggendo i pixel morti e bruciati. Per rimuovere il rumore di fondo del segnale si utilizzano i file di calibrazione degli strumenti. Anche gli errori legati alla lettura degli elettroni da parte dei sensori vengono corretti tramite i file di calibrazione forniti nel dataset di partenza. Infine vengono applicate altre tecniche di correzione del segnale quali la conversione analog to digital (ADC) e il campionamento Correlated Double Sampling (CDS).

2. Finestra di transito: Viene individuata la fase di transito del pianeta davanti alla stella mediante l'analisi della derivata prima del segnale medio ottenuto lungo l'asse delle componenti dello spettro atmosferico. In particolare si individuano i due momenti nel quale il segnale raggiunge le pendenze maggiori in modulo.

3. Modello di inferenza bayesiana: Viene creato un modello di inferenza bayesiana posto all'individuazione dell'andamento spettrale della luminosità della stella, indipendente dal passaggio stesso del pianeta, e della profondità di transito del segnale per ogni lunghezza d'onda. L'idea utilizzata in questa fase è quella dell'analisi della spettroscopia di trasmissione: durante il transito del pianeta la luce della stella attraversa l'atmosfera del pianeta subendo un assorbimento selettivo a seconda della composizione chimica dell'atmosfera stessa del pianeta. I gas assorbono la luce a specifiche lunghezze d'onda ed essendo presenti nell'atmosfera in quantità differenti anche la luce assorbita per ognuna di esse risulterà differente. La profondità di transito per una

specifica lunghezza d'onda, dunque, coincide esattamente con l'assorbimento atmosferico per quella determinata lunghezza d'onda.

Il modello prevede la costruzione di una prior informativa per ogni componente dello spettro, che utilizza conoscenze a priori apprese durante l'analisi del segnale medio. Viene inoltre definita una funzione di likelihood che restituisce i residui del segnale reale e quello corretto tramite l'andamento stellare stimato, tenendo conto del passaggio del pianeta e della profondità di transito per ciascuna lunghezza d'onda dello spettro atmosferico del pianeta. Infine tramite la massimizzazione della posterior, metodo chiamato in letteratura Massimo a Posteriori (MAP), si stima la profondità di transito per ciascuna lunghezza d'onda. In particolare il problema di minimo viene risolto tramite l'applicazione del Nelder-Mead method. Per quanto riguarda l'incertezza associata, si utilizzano stime che assumono la distribuzione dei parametri della posterior pressochè gaussiane in un intorno del MAP.

4. Correzione bias e offset: Grazie al dataset di train, le predizioni finali vengono riscalate tramite una matrice di correzione che individua bias frequenti e corrette grazie a un vettore che salva invece quantità fisse da aggiungere/togliere alle predizioni. Tale matrice dei bias e il vettore di offset si ottengono tramite tecniche di discesa del gradiente poste a minimizzare i residui generati dalle predizioni e lo spettro target.

Nel capitolo 2 viene fornita una panoramica più dettagliata della challenge e dei dataset forniti per la risoluzione del problema. Si riportano, inoltre, i principali articoli che hanno permesso lo studio della spettroscopia di trasmissione, dei modelli di inferenza bayesiana e dei modelli di minimizzazione, quali il Nelder-Mead method. Vengono inoltre descritti gli articoli che trattano l'utilizzo di tecniche di inferenza bayesiana per l'inverse modelling di parametri stellari e di esopianeti. Le implementazioni testate per generare un preprocessing robusto e i principali modelli utilizzati per ottenere la stima dello spettro atmosferico vengono descritte nel capitolo 3. Nel capitolo 4 si mostrano i risultati ottenuti e più dettagliatamente i contributi legati a ciascuna nuova introduzione nel modello, per trarne le conclusioni finali nel capitolo 5.

Capitolo 2

Background and Related works

Analiziamo ora più nel dettaglio la tecnica utilizzata per la stima dello spettro atmosferico degli esopianeti, chiamata Spettroscopia di trasmissione (2.1). In seguito verranno presentati gli algoritmi e la teoria matematica che sta alla base del modello creato, quali in Nelder Mead method (2.2) e la Teoria Bayesiana (2.3) e come essa sia stata utilizzata in letteratura per l'analisi di segnali stellari e di esopianeti (2.4). Infine viene fornita una panoramica più dettagliata della challenge, dei suoi dataset e di come venga calcolato lo score in fase di valutazione dei progetti (2.5).

2.1 Spettroscopia di trasmissione

La spettroscopia di trasmissione [1] è una tecnica utilizzata per analizzare l'atmosfera degli esopianeti. Il segnale luminoso della stella viene osservato prima, durante e dopo il transito del pianeta, e l'analisi dei suoi cambiamenti nel tempo permette di ricavare importanti informazioni in merito al pianeta stesso. La luce stellare viene filtrata dalle atmosfere planetarie a specifiche lunghezze d'onda e lo studio di come varia il segnale per le varie lunghezze d'onda permette di comprendere la chimica e la fisica di queste atmosfere [2]. I passaggi principali della tecnica per ricavare la composizione atmosferica dal segnale della stella sono:

1. Osservazione del transito e raccolta dei dati: Durante il transito, la luce della stella attraversa l'atmosfera del pianeta. Le molecole presenti nell'atmosfera del pianeta assorbono la luce della stella in corrispondenza di specifiche lunghezze d'onda, creando un calo della luminosità nel segnale luminoso proprio in corrispondenza del transito del pianeta stesso. Poichè si vuole risalire alla

composizione atmosferica, dunque, non è sufficiente il segnale luminoso su tutto lo spettro della luce. Risulta necessario utilizzare spettrografi ad alta risoluzione per registrare lo spettro della luce per diverse lunghezze d'onda, coprendo un ampio intervallo spettrale.

2. Sottrazione dello spectrum di riferimento: È importante rimuovere il contributo del segnale della stella, che può avere un impatto decisivo durante le analisi e isolare il solo effetto del transito. Le variazioni nel segnale stellare generate dal transito del pianeta alle volte risultano minime rispetto al segnale intrinseco della stella. Bisogna quindi cercare di individuare quali sono le fluttuazioni dovute al passaggio dell'esopianeta e quali sono invece oscillazioni dovute alla stella stessa. Questo viene fatto confrontando lo spettro rilevato durante il transito con quello della stella quando il pianeta non la sta eclissando, creando uno "spectrum di riferimento".

3. Calcolo della profondità di transito: viene individuata la profondità di transito del pianeta, cioè la variazione di intensità della luce stellare durante il transito. Tale quantità corrisponde a quanta luce è stata assorbita dalle molecole nell'atmosfera del pianeta.

4. Recupero del segnale atmosferico: L'assorbimento della luce non è uguale lungo tutto lo spettro luminoso, ma dipende dalla quantità in atmosfera di molecole che assorbono la luce a lunghezze d'onda differenti. Si utilizzano dunque modelli atmosferici per ricavare esattamente le lunghezze d'onda in cui avviene l'assorbimento della luce e in che proporzioni esso avviene. Da tali analisi si ricava così lo spettro atmosferico del pianeta, mostrando le lunghezze d'onda in corrispondenza delle quali si ha un assorbimento maggiore/minore della luce e di conseguenza in quali proporzioni sono presenti le varie molecole che assorbono la luce a quelle specifiche lunghezze d'onda.

5. Identificazione delle molecole: L'analisi dello spettro atmosferico consente di identificare le molecole presenti, come H_2O, CO_2, CH_4 e altri composti. Confrontando i dati osservati con modelli teorici, si possono ricavare informazioni relative alle condizioni atmosferiche del pianeta, quali temperatura, pressione e composizione chimica.

Nel caso della challenge Ariel 2025 [3], l'obiettivo è quello di giungere al quarto step dell'intero processo con i dati forniti dal primo step. La richiesta infatti è quella di stimare lo spettro atmosferico del pianeta suddiviso per lunghezze d'onda [4, 5, 6], mentre l'identificazione delle molecole che compongono effettivamente l'atmosfera (quinto step) non era parte delle richieste.

Molti articoli mostrano come questa tecnica sia molto efficace per la stima della composizione atmosferica dei pianeti. [7], [8] analizzano la caratterizzazione delle atmosfere degli esopianeti attraverso la spettroscopia di trasmissione di cui viene fornita la presentazione generale trattata precedentemente, approfondendone le limitazioni e le prospettive future. [9] presenta uno studio sulla spettroscopia di transito dell'esopianeta HAT-P-11b, per capire se in corrispondenza di determinate lunghezze d'onda i dati osservati fornissero informazioni reali sulla composizione atmosferica del pianeta o se i segnali fossero contaminati dalla presenza di macchie solari sulla superficie della stella e che quindi alcuni cali di luminosità fossero dovuti alla presenza di quest'ultime. Nell'articolo [10] la spettroscopia di trasmissione viene utilizzata per lo studio dell'atmosfera del pianeta Venere, per poi essere utilizzata come prototipo per l'analisi delle atmosfere di esopianeti rocciosi simili alla Terra ma altamente irradiati, mentre [11] analizza la spettroscopia di trasmissione del pianeta WASP-7 b, per stimare in particolare la presenza di sodio, Na I, attraverso lo spettrografo UVES.

2.2 Nelder Mead method

Nelder e Mead nel loro articolo [12] hanno descritto un algoritmo per la minimizzazione di una funzione a n variabili, chiamato Metodo del simplesso, ora meglio conosciuto come Metodo Nelder-Mead. Si tratta di un metodo di ricerca euristica che non fa uso delle derivate, come invece ne fanno uso i più classici algoritmi di minimizzazione quali Gradient descent o metodo di Newton, e può dunque convergere verso punti non stazionari.

L'algoritmo si basa sul concetto di simplesso:

Definizione. Un simplesso S in $\mathbf{R}^n$ è l'inviluppo convesso di $n+1$ punti $x_i \in \mathbf{R}^n$ per $i = 1, ..., n+1$:

$$S = \{y \in \mathbf{R}^n : y = \sum_{i=1}^{n+1} \lambda_i x_i,\ \lambda_i \geq 0,\ \sum_{i=1}^{n+1} \lambda_i = 1\}$$

e su 4 diversi parametri:

- ρ: riflessione, $\rho > 0$
- χ: espansione, $\chi > 1$
- γ: contrazione, $0 < \gamma < 1$
- σ: restrizione, $0 < \sigma < 1$

L'idea alla base dell'intero processo è quella di sostituire uno degli $n+1$ punti, dove la funzione f assume il valore peggiore, con un nuovo punto per migliorare il simplesso. Vediamo come si svolge una singola iterazione k dell'algoritmo.

Sia S_k il simplesso non singolare al passo k e siano $x_1^{(k)}, ..., x_{n+1}^{(k)}$ i punti che generano S_k. Si utilizza la notazione tale per cui $f(x_n^{(k)}) = f_n^{(k)}$.

1. Fase di riordino:

   Si calcola f nei punti del simplesso e si riordinano i vertici in modo tale che

   $$f_1^{(k)} \leq f_2^{(k)} \leq ... \leq f_{n+1}^{(k)}$$

2. Fase di riflessione:

   Sia $\bar{x}$ il baricentro degli n punti migliori, cioè $\bar{x}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(k)}$. Si calcoli ora la riflessione $x_R^{(k)}$ di $x_{n+1}^{(k)}$ rispetto al baricentro $\bar{x}^{(k)}$:

   $$x_R^{(k)} = \bar{x}^{(k)} + \rho(\bar{x}^{(k)} - x_{n+1}^{(k)})$$

   Indichiamo da qui in poi $f(x_R^{(k)})$ come $f_R^{(k)}$. Se $f_1^{(k)} \leq f_R^{(k)} < f_n^{(k)}$, allora si accetta $x_R^{(k)}$ come nuovo punto di S_{k+1} e si passa all'iterazione successiva.

3. Fase di espansione:

   Se $f_R^{(k)} < f_1^{(k)}$ allora si calcola l'espansione $x_E^{(k)}$ come:

   $$x_E^{(k)} = \bar{x}^{(k)} + \chi(x_R^{(k)} - \bar{x}^{(k)})$$

   Indichiamo da qui in poi $f(x_E^{(k)})$ come $f_E^{(k)}$. Se $f_E^{(k)} < f_R^{(k)}$, allora si accetta $x_E^{(k)}$ come nuovo punto di S_{k+1} e si passa all'iterazione successiva.

4. Fase di contrazione:

   Se $f_R^{(k)} \geq f_n^{(k)}$ si calcola una contrazione tra $\bar{x}^{(k)}$ e il migliore tra $x_R^{(k)}$ e $x_{n+1}^{(k)}$, che indichiamo con $x_B^{(k)}$. Il nuovo punto è dato da:

   $$x_C^{(k)} = \bar{x}^{(k)} - \gamma(\bar{x}^{(k)} - x_B^{(k)})$$

   Se $f_C^{(k)} < f_{n+1}^{(k)}$, allora si accetta $x_C^{(k)}$ come nuovo punto di S_{k+1} e si passa all'iterazione successiva.

5. Fase di restrizione:

   Si restringe il simplesso attorno al punto migliore $x_1^{(k)}$, ottenendo così un set di nuovi punti per il simplesso S_{k+1}:

   $$x_i^{(k+1)} = x_1^{(k)} + \sigma(x_i^{(k)} - x_1^{(k)}), \quad \forall\, i = 2, ..., n+1$$

2.3 Modelli Bayesiani e Maximum at Posterior

Come possiamo leggere in [13] la statistica bayesiana e i modelli che ne derivano si basano sul fatto che la probabilità assegnata ad un evento sia una misura della conoscenza o della convinzione personale che si attribuisce ad esso, piuttosto che una frequenza osservabile, come succede invece nella statistica frequentista. Il metodo bayesiano ha origine dal teorema di Bayes che afferma che:

Date due variabili aleatorie (anche vettoriali) X e Y, allora

$$f(x|y) = \frac{f(y,x)}{f(y)} = \frac{f(y|x)f(x)}{f(y)}$$

dove $f(y) = \int f(y,x)dx$ se X è continua e $f(x) = \sum f(y,x)$ se è discreta.

Il teorema di Bayes permette di passare dalla condizionata di y|x a quella di x|y. Si può anche darne un'interpretazione di tipo iterativo. Si assegna una probabilità iniziale detta prior a x, $f(x)$, che rispecchia il grado di affidabilità iniziale che si assegna all'evento descritto da tale distribuzione. In seguito osservo una nuova variabile y, che dipende da x, tramite $f(y|x)$. L'informazione che ho su x, dopo aver osservato y, cambia in $f(x|y)$, detta posterior. La nuova distribuzione aggiornata dalle nuove evidenze può essere utilizzata come nuova prior di partenza per l'iterazione successiva. Questo approccio fornisce un modo per calcolare la plausibilità di un'ipotesi, esprimendola come un valore tra 0 e 1.

Indichiamo ora con θ un vettore di parametri di interesse, di cui non conosciamo la distribuzione e y un set di valori osservabili. Siamo interessati alla quantità $f(\theta|y)$, cioè alla distribuzione di probabilità del parametro θ osservato y. Anche in questo caso è possibile applicare Bayes:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

dove $f(y|\theta)$ è la congiunta delle osservazioni, che si può interpretare come la verosimiglianza, $f(\theta)$ è la distribuzione a priori e $f(y)$ è la costante di normalizzazione, poco importante in quanto non dipende dal parametro θ. Una distribuzione è interamente descritta dal suo kernel, che in questo caso per la variabile aleatoria $f(\theta|y)$ risulta essere $f(y|\theta)f(\theta)$. In un approccio di tipo frequentista, essendo il parametro da stimare un singolo valore e non una variabile aleatoria, l'obiettivo è quello di massimizzare la verosimiglianza. In questo caso invece la parte che assume maggiore rilevanza è proprio il kernel della distribuzione, per cui da qui in poi la costante di normalizzazione non verrà ulteriormente presa in considerazione e si tratterà direttamente con:

$$f(\theta|y) \propto f(y|\theta)f(\theta) \tag{2.1}$$

Tendenzialmente è difficile individuare $f(\theta|y)$ poichè non appartiene a una famiglia di distribuzioni ben precisa. Quando invece $f(\theta|y)$ corrisponde a una distribuzione nota, si dice che la prior e la verosimiglianza sono coniugate. I casi in cui questo accade sono molto limitati, per cui sono stati trovati altri metodi per campionare dalla posteriori, quali Gibbs sampling o Metropolis-Hasting di cui viene fornita una presentazione generale in [14], mentre in [15] viene approfondito l'algoritmo di Gibbs per la stima dei parametri. Nel caso preso in analisi però un campionamento diretto dalla posterior non risulta possibile dato l'alto numero di parametri. Nell'articolo [16] si fornisce una spiegazione più dettagliata del MAP, Maximum At Posterior, che prevede la massimizzazione della posterior. Essa presenta un grande potenziale in quanto non richiede la stima dell'intera distribuzione della posterior o un campionamento da essa, ma i parametri vengono stimati direttamente tramite classici algoritmi di ottimizzazione. Tale metodo è strettamente legato a quello utilizzato nella statistica frequentista della massimizzazione della verosimiglianza. Possiamo interpretare la tecnica MAP come una regolarizzazione della massimizzazione della verosimiglianza, data la natura aleatoria dei parametri e dunque la presenza di una prior che la descrive. Vediamo più nel dettaglio la differenza tra i due approcci.

Supponiamo di voler stimare il parametro θ date le osservazioni y. Sia f la distribuzione campionaria di y, tale per cui $f(y|\theta)$ sia la distribuzione di y generata dal parametro θ.

- Approccio frequentista.

  Sappiamo che $f(y|\theta) = L(\theta|y)$, dove L rappresenta la verosimiglianza di θ osservati i dati y. La tecnica della massimizzazione della verosimiglianza prevede che il parametro θ sia stimato tramite la seguente:

  $$\hat{\theta}(y) = argmax_{\,\theta}\, L(\theta|y) = argmax_{\,\theta}\, f(y|\theta)$$

- Approccio bayesiano.

  Assumiamo che θ sia variabile aleatoria e dunque ammettiamo una prior $f(\theta)$. In questo caso la tecnica del MAP, combinata al risultato raggiunto in 2.1, ci permettono di ottenere:

  $$\hat{\theta}_{MAP}(y) = argmax_{\,\theta}\, f(\theta|y) = argmax_{\,\theta}\, f(y|\theta)\, f(\theta)$$

Vediamo dunque come nella formulazione bayesiana sia presente il termine di regolarizzazione dato dalla prior.

La tecnica sopra presentata è molto potente nonostante la sua semplicità. Vediamo, infatti, come essa sia utilizzata in vari contesti quali la risoluzione

di ODE in [17] o la stima di parametri per problemi non parametrici con prior definite in spazi di Besov presentata in [18].

Come viene sottilineato in [19], risulta molto utile descrivere la funzione di prior come distribuzione log-concava, in quanto essa permette, grazie a una formulazione convessa del problema di minimizzazione, di utilizzare algoritmi di ottimizzazione computazionalmente stabili, che scalano in modo efficiente con le grandi dimensioni. Vediamo come in tale progetto si utilizzi infatti lo Stochastic gradient descent.

Anche in altri casi si è visto come metodi Montecarlo o similari fossero computazionalmente insostenibili per i costi computazionali associati. In alternativa al metodo del MAP, ad esempio in [20] viene utilizzata una decomposizione della verosimiglianza in poligoni ortogonali rispetto alla prior. Questo viene fatto proprio per evitare l'uso di MCMC e ottenere un stima semi-analitica della posterior, data dal prodotto della prior e dei polinomi ortogonali della verosimiglianza.

Nel progetto preso in analisi viene richiesta anche una stima dell'incertezza associata alla stima dei parametri. Il MAP permette di stimare direttamente il valore migliore per i parametri, ma non fornisce alcuna informazione sulla sua distribuzione. Soltanto un campionamento potrebbe dare informazioni precise, ad esempio, sulla varianza della distribuzione. Per riuscire a stimare la varianza della posterior viene utilizzata l'approssimazione presentata nel libro [13], per cui si assume che in un intorno del MAP la distribuzione sia pressoché gaussiana. Fatta tale assunzione si può vedere come il logaritmo della posterior si possa esprimere come una funzione quadratica, in particolare tramite un'espansione di Taylor si ha:

$$log f(\theta|y) = log\, f(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} log\, f(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + ... \tag{2.2}$$

dove $\hat{\theta}$ è il valore ottimizzato attraverso il MAP. Il termine dell'espansione al primo ordine è uguale a 0, in quanto la log-posterior ha derivata 0 nel suo punto di ottimo. Osservando l'equazione 2.2 si può vedere come il primo termine risulti essere costante rispetto al parametro θ, mentre il secondo termine, data l'approssimazione iniziale, sia proporzionale al logaritmo di una normale:

$$f(\theta|y) \approx N(\hat{\theta},\, [I(\hat{\theta})]^{-1}) \tag{2.3}$$

dove $I(\theta)$ è dato da:

$$I(\theta) = -\frac{d^2}{d\theta^2} log\, f(\theta|y)$$

Grazie a tale assunzione possiamo stimare la varianza della posterior come la varianza data dalla 2.3

2.4 Modelli bayesiani e analisi di segnali stellari

In letteratura si può osservare come l'adozione di tecniche di inferenza bayesiana, utilizzate per la modellazione e la stima dei parametri, abbia mostrato un grande potenziale nel migliorare l'accuratezza e la quantificazione dell'incertezza di modelli stellari, transiti di esopianeti e le loro composizioni atmosferiche. Essi risultano particolarmente utili per la stima di parametri fisici, quali temperatura, pressione, composizione e copertura nuvolosa. Un esempio in cui è possibile osservare tale approccio è [21], in cui viene presentato il codice SteParSyn che impiega campionamenti Markov Chain Monte Carlo (MCMC) per stimare i parametri atmosferici stellari (temperatura, gravità superficiale e metallicità) da dati spettrali delle stelle prese in analisi. Questo metodo è particolarmente utile perchè permette di stimare le distribuzioni di tali parametri e non solo stime puntuali, utili soprattuto per quanto riguarda le stime dell'incertezza associata.

Nell'articolo [22] i modelli di inferenza bayesiana vengono invece utilizzati per l'individuazione dei transiti di pianeti extrasolari. In particolare il metodo si basa su tecniche bayesiane utilizzate per identificare periodicamente le caratteristiche dei transiti planetari nelle curve di luce generate dalla stella e di ricostruirne la forma. L'algoritmo è in grado di determinare la periodicità con buona precisione (dell'ordine di un'ora) e di discriminare tra eventi di transito planetario e artefatti, tramite un processo completamente automatizzato.

Altri approcci bayesiani sono stati utilizzati nel campo delle atmosfere stellari e degli esopianeti, come ad esempio i modelli di recupero atmosferico, approccio che utilizza l'inverse modelling e algoritmi di inferenza bayesiana per stimare i parametri atmosferici e le loro incertezze. Uno dei contributi più importanti in materia è dato da TauREx, [23], che mira a stimare la composizione chimica e le condizioni fisiche delle atmosfere degli esopianeti tramite lo studio e la modellazione degli spettri. In particolare vengono utilizzati due algoritmi indipendenti per l'analisi Bayesiana: Nested Sampling e Markov Chain Monte Carlo (MCMC), con selezione iterativa del modello più adatto. Sono state rilasciate altre due versioni di TauREx, che migliorano ulteriormente il codice originale. In particolare nella terza versione, descritta in [24], è stata migliorata l'interfaccia con l'utente, tramite la creazione della libreria python corrispondente. L'ultima versione presenta inoltre tempi di calcolo ancora più veloci rispetto alle versioni precedenti e rispetto ad altri codici per il recupero atmosferico per gli esopianeti, quali Nemesis o Hidra, trattati rispettivamente in [25] e [26].

Nell'articolo [27] si può leggere, come uno dei metodi principali per misurare le proprietà atmosferiche sia il recupero atmosferico, come sopra citato, ma i modelli

stessi possono adattarsi ai dati osservativi, ignorando le incertezze sulla validità del modello stesso. L'autore qui propone una combinazione di modelli, tra cui tecniche che combinano le distribuzioni posteriori dei parametri pesandole in base all'evidenza bayesiana di ciascun modello, in modo tale da stimare un'incertezza che non sia sottostimata dall'overfitting del modello sui dati osservati.

2.5 Ariel data Challenge e dataset di training

Come viene presentato nel sito ufficiale della challenge [3], il dataset fornito per la soluzione del problema è composto dai seguenti documenti, per ciascuno dei pianeti presi in analisi:

1. I metadati.
2. I file di calibrazione degli strumenti.
3. I file dei segnali

Per quanto riguarda i metadati, vengono resi disponibili i seguenti file:

- Gain e offset per conversione ADC: nei seguenti file vengono salvati i parametri di conversione ADC, Analog to Digital, per ciascun pianeta. In genere il segnale ricevuto dal rilevatore viene elaborato per convertire la tensione del pixel captata in un numero intero. Nei file di metadati vengono salvati il guadagno e l'offset utilizzati per effettuare questa conversione. Viene inoltre indicata la stella di riferimento del pianeta. Per il dataset di train, in particolare, i pianeti provengono da due sistemi stellari differenti, mentre per il dataset di test viene segnalato che i pianeti possono essere simulati anche da stelle diverse da quelle presenti nel dataset, ma non viene fornita alcuna altra informazione precisa sul numero effettivo di stelle considerate e su quali esse siano.
- train labels: per il dataset di train viene fornito lo spettro atmosferico atteso.
- axis info: Informazioni sugli assi per entrambi gli strumenti. Le immagini di ciascun pianeta vengono captate a intervalli di tempo regolari, la cui durata è registrata in questi file e varia a seconda dello strumento utilizzato.
- wavelength: Vengono riporate le lunghezze d'onda associate a ciascuna componente dello spettro preso in considerazione. In particolare si hanno 283 diverse lunghezze d'onda e dunque 283 componenti da stimare per ciascun spettro atmosferico.

I file di calibrazione registrano le caratteristiche elettroniche del sensore e vengono utilizzati per il miglioramento e pulizia delle immagini. Nello specifico si hanno:

- dark parquet: catturano il rumore termico e il livello di bias del sensore, poichè sono ottenuti dall'analisi delle esposizioni a sensore oscurato. Questi vengono utilizzati per sottrarre la corrente oscura dalle immagini.
- dead parquet: Identifica i pixel morti o caldi sul sensore. I pixel caldi producono costantemente alti livelli di segnale indipendentemente dalla luce in ingresso, mentre i pixel morti, al contrario, non reagiscono alla luce.
- flat parquet: Sono fotogrammi creati esponendo una superficie illuminata in modo uniforme. Vengono utilizzati per correggere le variazioni nella sensibilità pixel-per-pixel e le irregolarità nel sistema ottico.
- linear correction parquet: contengono informazioni sulla correzione della lettura lineare del sensore. I pixel non reagiscono in maniera lineare all'accumulo di elettroni sul sensore. Più nello specifico, a mano a mano che ci si avvicina al punto di saturazione, il pixel registra in maniera sempre più scorretta il numero effettivo di elettroni, fino al punto da non poter più raccogliere elettroni aggiuntivi e dunque fornire una risposta alla luce completamente piatta. Per una stima accurata del segnale, la risposta deve essere calibrata tramite un polinomio di grado n. Questo polinomio permette di convertire il numero di elettroni raccolti/misurati dal pixel nel numero di elettroni che il sensore avrebbe generato con una risposta lineare. In questi file in particolare vengono salvati i coefficienti del polinomio che serve per invertire questa tendenza dei sensori.
- read parquet: Anche durante la lettura stessa del sensore vengono registrati degli errori. Essi sono portati dal rumore elettronico, presente anche quando nessuna luce arriva al rivelatore. Tale rumore viene registrato in questi fotogrammi.

Per generare i file veri e propri dei pianeti sono stati utilizzati due diversi strumenti ottici, FGS1 e AIRS-CH0, specializzati in analisi di bande spettrali diverse. FGS1 è specializzato nell'allineamento e messa a fuoco del satellite. Inoltre fornisce dati maggiormente precisi nello spettro del visibile. AIRS-CH0 è invece uno spettrometro infrarosso. La rilevazione del segnale da parte dei due strumenti si articola in due momenti: prima una sotto-esposizione del segnale registra la carica, in seguito, dopo un tempo di esposizione prefissato, viene nuovamente misurata la carica accumulata. Questo processo viene continuamente resettato per l'intero tempo di osservazione della stella durante il transito del pianeta. I fotogrammi generati dagli strumenti presentano 135.000 fotogrammi per FGS1 e 11.250 fotogrammi per AIRS-CH0. Vediamo più nello specifico i file forniti:

- Segnale AIRS-CH0: salva i dati del segnale provenienti dallo strumento AIRS-CH0. Ogni file contiene 11.250 righe di immagini catturate a intervalli regolari.

Ogni immagine 32 x 356 è stata appiattita in 11392 colonne. Nella figura 2.1 possiamo vedere l'immagine in 2 dimensioni (spazio e lunghezza d'onda) relativa ad uno dei pianeti del dataset di train nell'istante di tempo iniziale, in cui il pianeta ancora non oscura la sua stella.

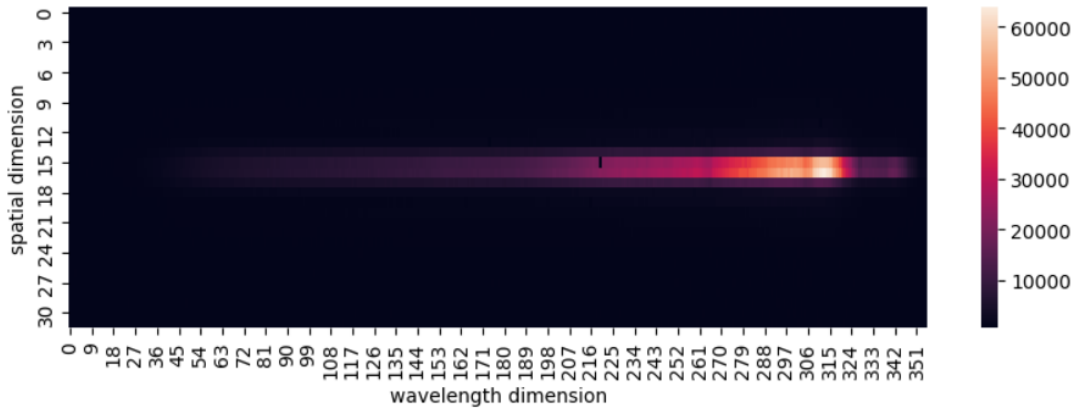

Figura 2.1: Immagine 2D del segnale relativo al pianeta 4249337798, tramite sensore AIRS-CH0, nell'istante iniziale

- Segnale FGS1: salva i dati del segnale provenienti dallo strumento FGS1. Ogni file contiene 135.000 righe di immagini a passi temporali di 0,1 secondi. Ogni immagine 32x32 è stata appiattita in 1024 colonne.

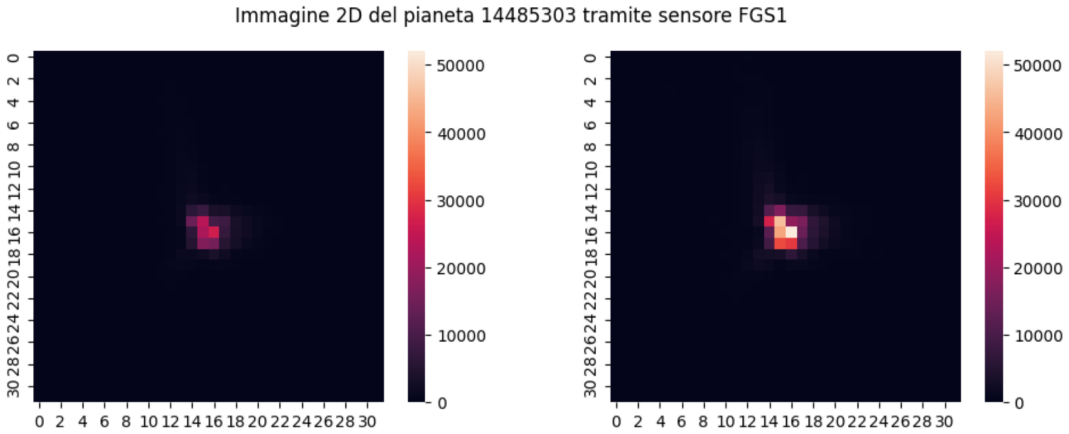

Figura 2.2: Immagini 2D del segnale relativo al pianeta 14485303, tramite sensore FGS1, prima e dopo la fase di accumulo di carica.

Nella figura 2.2 possiamo vedere l'immagine in 2 dimensioni relativa ad uno dei pianeti del dataset di train nell'istante di tempo iniziale. Per ogni istante

di tempo, come anticipato precedentemente, vengono salvate due immagini, una prima e una dopo l'accumulo di carica sul sensore. Vediamo infatti come la seconda immagine sia più luminosa della prima.

Per quanto riguarda il punteggio della challenge, esso viene calcolato tramite la funzione di Log verosimiglianza gaussiana (GLL), la cui formulazione è data dall'equazione 2.4:

$$GLL = -\frac{1}{2}\left(\log(2\pi) + \log(\sigma_{\text{user}}^2) + \frac{(y - \mu_{\text{user}})^2}{\sigma_{\text{user}}^2}\right) \tag{2.4}$$

Possiamo osservare una rappresentazione grafica della GLL in funzione dell'incertezza σ_{user} nella figura 2.3:

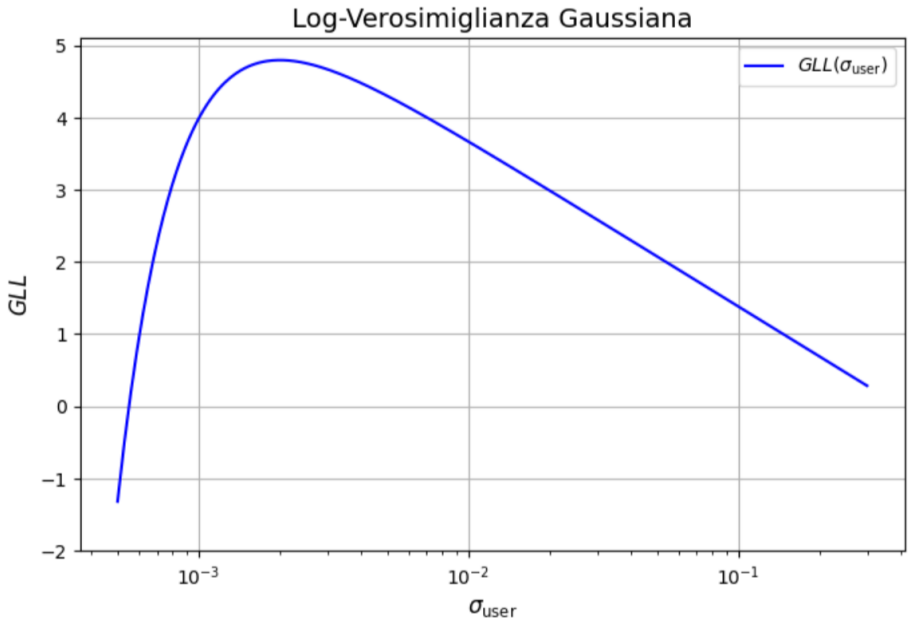

Figura 2.3: Grafico GLL, al variare di σ_{user}, fissati i valori di μ_{user} e y

dove μ_{user} e σ_{user} sono rispettivamente il valore stimato dal partecipante dello spettro atmosferico per una specifica lunghezza d'onda e l'incertezza associata. In seguito, i valori di GLL ottenuti per ogni coppia vengono sommati su tutte le lunghezze d'onda e su tutto il dataset di test per ottenere un valore finale L. Lo score finale è dato da:

$$\text{score} = \frac{L - L_{\text{ref}}}{L_{\text{ideal}} - L_{\text{ref}}} \tag{2.5}$$

dove L_{ideal} rappresenta il caso in cui lo spettro stimato sia esattamente quello reale, mentre L_{ref} è definito utilizzando media e varianza del dataset di train come previsione base per tutti i campioni.

Il dataset su cui viene calcolato lo score della challenge invece non viene fornito ai singoli partecipanti. Durante la sottomissione si può comunque far riferimento e leggere anche per il test i file di calibrazione e i metadati, a parte ovviamente il file dei target. Come già anticipato precedentemente, inoltre, il dataset di test valuta il modello su pianeti appartenenti a sistemi stellari diversi da quelli del dataset di train.

Sia per il dataset di train sia per quello di test non vengono date informazioni molto precise sull'incertezza associata alle componenti dello spettro. Dalla funzione 2.4 possiamo notare come vengano penalizzati maggiormente valori con incertezza associata più piccola di quella corretta, rispetto invece a valori che permettono maggiore incertezza associata. Anche dal grafico 2.3 si vede chiaramente come, fissati i valori di μ_{user} e y si abbia una pendenza maggiore per valori più piccoli del valore ottimo di σ_{user} che coincide con il punto di massimo della GLL, mentre si ha un calo meno importante alla dua destra, cioè per stime meno restrittive di σ_{user}. Questo ha anche un'interpretazione logica ben precisa, in quanto se il valore del parametro stimato μ_{user} non coincide con quello corretto, esso deve essere meno penalizzato se ad esso è associata un'incertezza maggiore.

Capitolo 3

Metodologia

Nel seguente capitolo vengono approfonditi i passaggi principali del modello, di cui vediamo una rappresentazione grafica dell'architettura nella figura 3.1.

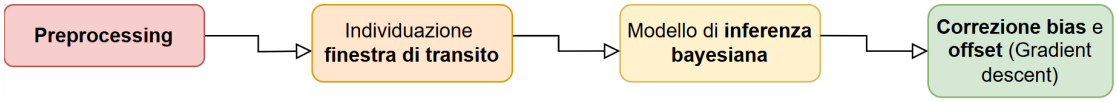

Figura 3.1: Architettura del modello utilizzato.

Gli step principali che costituiscono il modello sono i seguenti:

1. Preprocessing: Vengono inizialmente corretti i pixel difettosi, in particolare quelli morti e bruciati. Viene in seguito rimosso il rumore elettronico e corretti gli errori di lettura non lineare degli elettroni da parte dello strumento. Altre tecniche di correzione del segnale vengono applicate per pulire il segnale, quali la conversione analog to digital (ADC) e il campionamento CDS, Correlated Double Sampling.

2. Individuazione finestra di transito: Si analizza la derivata prima del segnale medio ottenuto lungo l'asse delle componenti dello spettro atmosferico. In particolare si individuano i due momenti nel quale il segnale raggiunge le pendenze maggiori in modulo che andranno poi a coincidere con gli istanti di ingresso ed egresso dalla fase di transito.

3. Modello di inferenza bayesiana: Viene creato un modello di inferenza bayesiana posto all'individuazione dell'andamento spettrale della luminosità della stella, indipendente dal passaggio stesso del pianeta, e della profondità di transito del segnale per ogni lunghezza d'onda. L'idea alla base è quella della spettroscopia di trasmissione, presentata nella sezione 2.1. Il modello ha dunque l'obiettivo di individuare la profondità di transito per una specifica lunghezza d'onda,

quantità che coincide esattamente con l'assorbimento atmosferico per quella determinata componente dello spettro. Viene inizialmente trovata la profondità di transito del segnale medio, stima iniziale uguale per tutte le lunghezze d'onda. In questa fase viene inoltre individuato il polinomio che meglio descrive l'andamento del segnale luminoso della stella, indipendente dal passaggio del pianeta. Il modello in seguito prevede la costruzione di una prior informativa data da una normale troncata per ogni componente dello spettro, la cui media è data dalla profondità di transito trovata nella fase precedente di analisi del segnale medio. La likelihood del modello restituisce i residui del segnale reale e quello corretto tramite l'andamento stellare stimato, tenendo conto del passaggio del pianeta e della profondità di transito per ciascuna lunghezza d'onda dello spettro atmosferico. Viene in seguito applicata la tecnica MAP (Mazimum at Posterior), presentata nella sezione 2.3 che stima il massimo della posterior, la quale è data dalla somma della likelihood e la log pdf della prior. Esso ci fornisce la stima della profondità di transito per ciascuna lunghezza d'onda. In particolare il problema di minimo viene risolto tramite l'applicazione del Nelder-Mead method. La posterior viene assunta gaussiana in un intorno del MAP, assunzione che ci permette di utilizzare la 2.3 per la stima dell'incertezza associata.

4. Correzione bias e offset (Gradient descent): Viene infine applicata una fase di training del modello per l'individuazione di una correzione lineare delle predizioni tramite una matrice diagonale di correzione di bias frequenti e un vettore di offset. Tale matrice dei bias e il vettore di offset si ottengono tramite tecniche di discesa del gradiente poste a minimizzare i residui generati dalle predizioni e lo spettro target.

Vediamo ora più nel dettaglio ciascuna di queste fasi.

3.1 Preprocessing

I dati provenienti dal rilevatore sono affetti da rumore, distorsioni strumentali e pixel non funzionanti che devono essere opportunamente gestiti. Il preprocessing del segnale presenta dunque diverse fasi per ripulirlo e garantire accuratezza.

Nella figura 3.2 possiamo vedere un esempio di segnale proveniente dal sensore AIRS-CH0 prima della fase di preprocessing. In particolare, vediamo il segnale relativo al pianeta 785834 nell'istante iniziale.

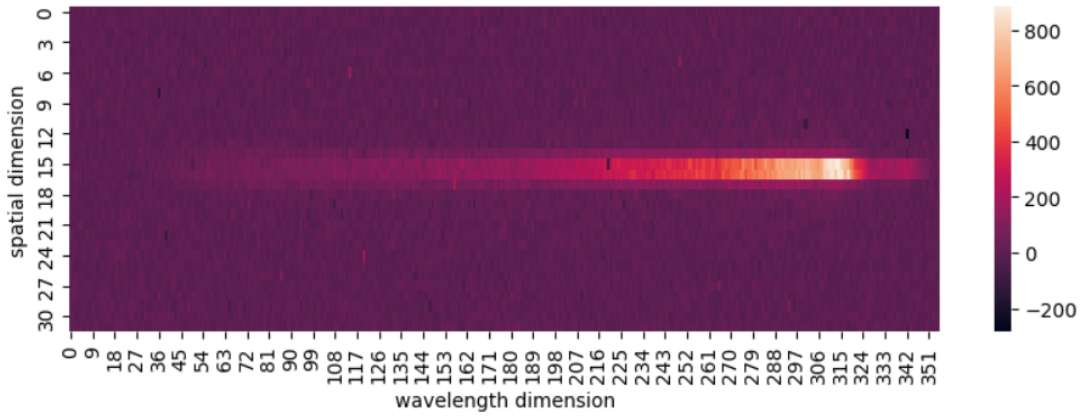

Figura 3.2: Segnale prima del proprocessing (pianeta 785834 istante iniziale)

Come primo step risulta necessaria l'inversione della conversione Analogico-Digitale (ADC). La conversione effettuata dal rilevatore deve essere invertita, utilizzando il guadagno e l'offset, salvati nei file di calibrazione dei pianeti, garantendo così che i dati analizzati siano proporzionali al segnale analogico originale.

Gli errori legati alla lettura da parte del sensore, in particolare generati da rumore elettronico, vengono subito eliminati. Questo passaggio viene effettutato sottraendo dal segnale i dati provenienti dal file di calibrazione dei fotogrammi read.

In un secondo momento si cerca di correggere i pixel delle immagini che presentano difetti: i pixel "caldi" vengono mascherati mediante metodi che individuano valori anomali rispetto al rumore di fondo. Questi, insieme ai pixel morti, di cui era fornita una mappa direttamente dai file di calibrazione, vengono impostati in una prima fase come valori nulli.

In seguito, il contributo del rumore intrinseco del sensore, fornito nei file di calibrazione come mappa dei dark frame, viene sottratto dal segnale. La sottrazione tiene conto del tempo di esposizione variabile tra le esposizioni.

Viene anche applicata una correzione lineare al segnale, utilizzando nuovamente i file di calibrazione per ogni pianeta. In particolare i file di Linear correction forniscono i coefficienti del polinomio inverso da utilizzare per mitigare l'effetto di non linearità, dovuta alla perdita capacitiva di lettura dei pixel al trascorrere del tempo. Si applica dunque la seguente:

$$S_{Lin}(t, x, \lambda) = P_{(x,\lambda)}(S(t, x, \lambda))$$

dove $S(t, x, \lambda)$ è il segnale a tempo t, componente spaziale x e componente spettrale λ, $P_{(x,\lambda)}$ è il polinomio inverso fissati x e λ, mentre S_{Lin} è il segnale corretto.

Nella figura 3.3 possiamo osservare il segnale in seguito all'inversione ADC, alla correzione dei read frame, del rumore oscuro salvato dei dark frame e della correzione lineare. Il segnale è stato inoltre ridimensionato, poiché sono stati eliminati i frame della parte superiore e inferiore dell'immagine. Sono chiaramente visibili i frame difettosi che ancora risultano impostati al valore NaN, identificabili nell'immagine come i frame completamente bianchi.

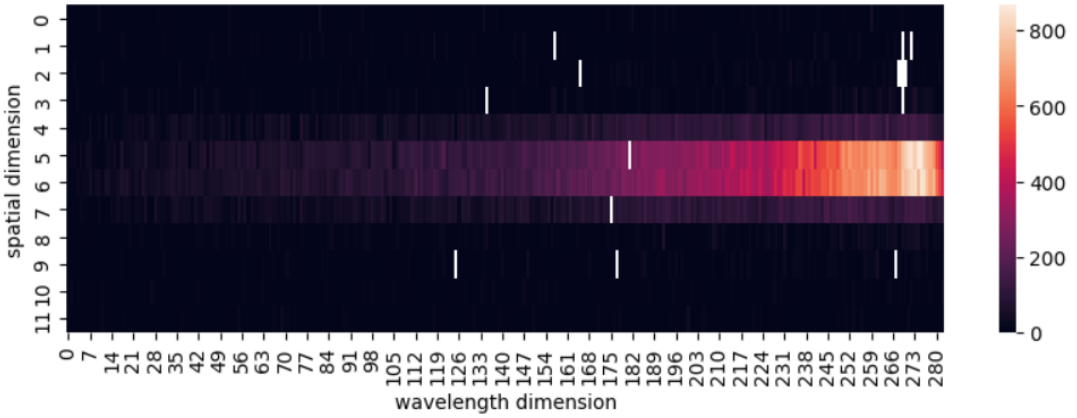

Figura 3.3: Segnale dopo la correzione dei read frame, dark frame, linear correction (pianeta 785834 istante iniziale)

Per rimuovere il rumore a bassa frequenza e migliorare la qualità del segnale, viene applicato il Correlated Double Sampling (CDS). Il rilevatore viene letto due volte, una volta all'inizio dell'esposizione e una volta alla fine. Questo metodo sottrae la prima lettura alla seconda, riducendo il rumore introdotto da variazioni termiche o elettroniche, in particolare se S è il segnale, t l'istante temporale, x la componente spaziale, λ la componente spettrale, mentre i e $i+1$ corrispondono agli indici che salvano la misurazione iniziale e finale del segnale per l'istante t:

$$S_{CDS}(t, x, \lambda) = S(i+1, x, \lambda) - S(i, x, \lambda)$$

Vediamo i miglioramenti apportati da questo processo nell'immagine 3.4.

I pixel difettosi, visibili nelle immagini precedenti come frame completamente bianchi, vengono ora sostituiti interpolando linearmente il segnale per ciascuna lunghezza d'onda. Per ogni istante temporale, dunque, nel caso in cui il frame dell'immagine (x, λ) sia nullo, esso viene sostituito mediante:

$$S_t(x, \lambda) = \frac{S_t(x, \lambda - 1) + S_t(x, \lambda + 1)}{2}$$

dove S_t è il segnale nell'istante di tempo t, x la componente spaziale e λ la componente spettrale. I pixel, invece, ai margini dell'immagine sono stati sostituiti mediante il primo frame valido nella stessa regione.

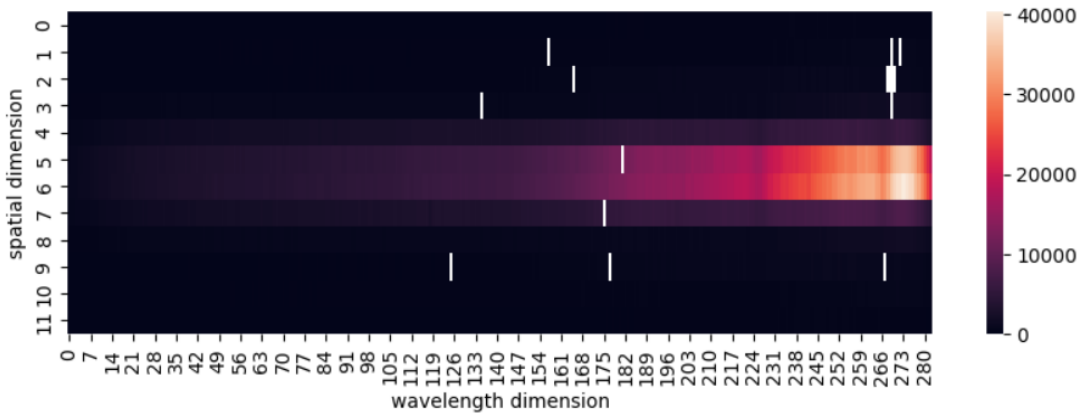

Figura 3.4: Segnale dopo la correzione CDS (pianeta 785834 istante iniziale)

Possiamo osservare graficamente questa fase di eliminazione dei frame nulli nell'immagine 3.5

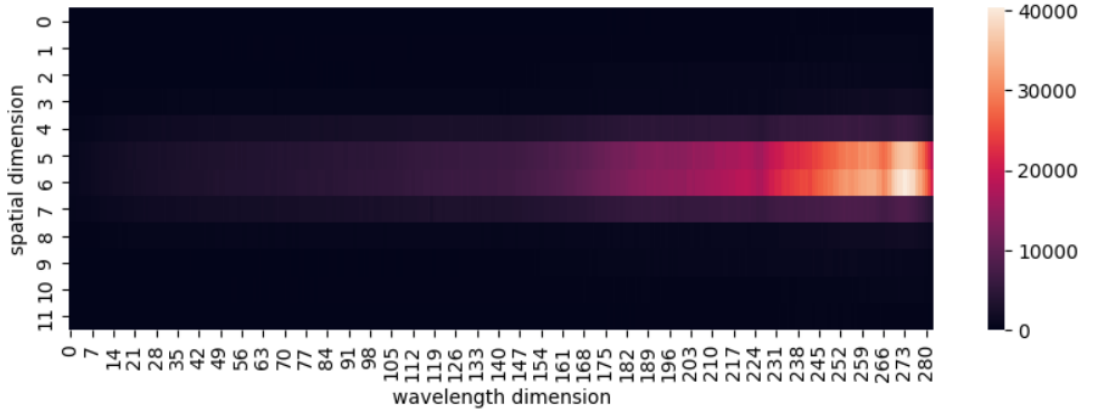

Figura 3.5: Segnale dopo la correzione dei pixel difettosi (pianeta 785834 istante iniziale)

Viene poi eliminata la dimensione spaziale dalle immagini, facendo la media lungo tale asse, ottenendo in questo modo l'evoluzione nel tempo dell'intensità del segnale per ciascuna lunghezza d'onda. Nell'immagine 3.6 viene fornita un'interpretazione grafica.

Nello step finale di preprocessing del segnale, viene applicata una tecnica di binning per ridurre la dimensionalità dei dati e migliorare il rapporto segnale-rumore. In particolare, il segnale viene suddiviso in blocchi di lunghezza pari al fattore di binning prestabilito e successivamente, per ciascun blocco, ne viene calcolata la media lungo l'asse temporale, ottenendo così una versione del segnale

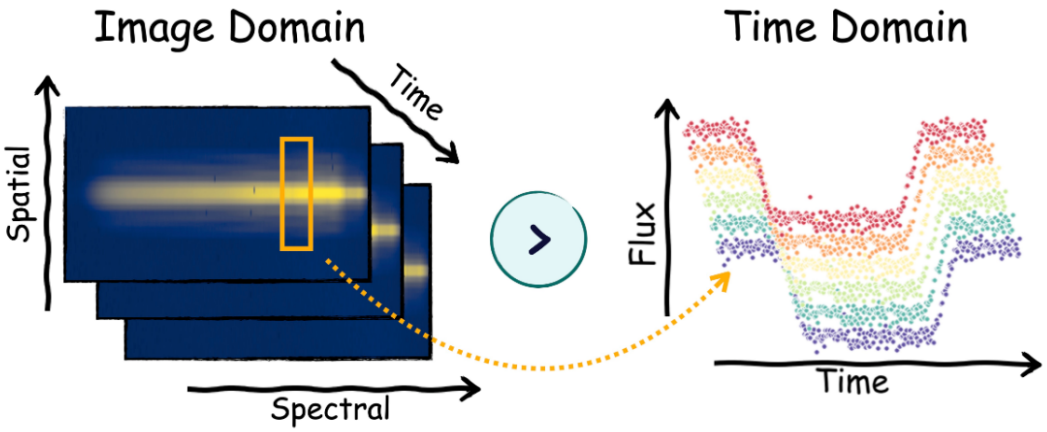

Figura 3.6: Illustrazione dei cambi di domino del segnale. Fonte: [28]

con una risoluzione temporale ridotta, ma un'accuratezza migliorata.

Da analisi di significatività del segnale nel modello, che sarà analizzato nella sezione 3.3, si è potuto notare come i dati provenienti dal sensore FGS1 fossero soggetti a troppo rumore e fornissero risultati fuorvianti senza un'effettiva ulteriore rimozione di quest'ultimo, oltre a ciò che è stato presentato in questa sezione. Per questo motivo si è deciso di non considerare tali dati. D'ora in avanti solo i dati provenienti dal sensore AIRS-CH0 saranno considerati.

3.2 Individuazione della finestra di transito

Si è potuto osservare come i segnali suddivisi per lunghezza d'onda fossero molto rumorosi nonostante il preprocessing, rendendo difficile l'individuazione degli istanti precisi in cui il pianeta eclissa la sua stella, come possiamo vedere nella figura 3.7. In tale immagine possiamo osservare come il segnale registrato per alcune lunghezze d'onda prese in analisi risulti molto rumoroso e presenti delle grandi oscillazioni, nonostante il binning temporale che ne ha ridotto la variabilità.

Per questo motivo per l'individuazione della fase di transito si è lavorato sul segnale medio lungo l'asse delle lunghezze d'onda, di cui vediamo un esempio nella figura 3.8, in modo tale che una semplice analisi della derivata prima del segnale potesse dare informazioni molto precise sull'inizio e fine della fase di transito.

Viene individuato il punto di minimo del segnale m e in seguito viene calcolata la derivata prima nella prima parte del segnale S_1, prima del punto di minimo, e nella

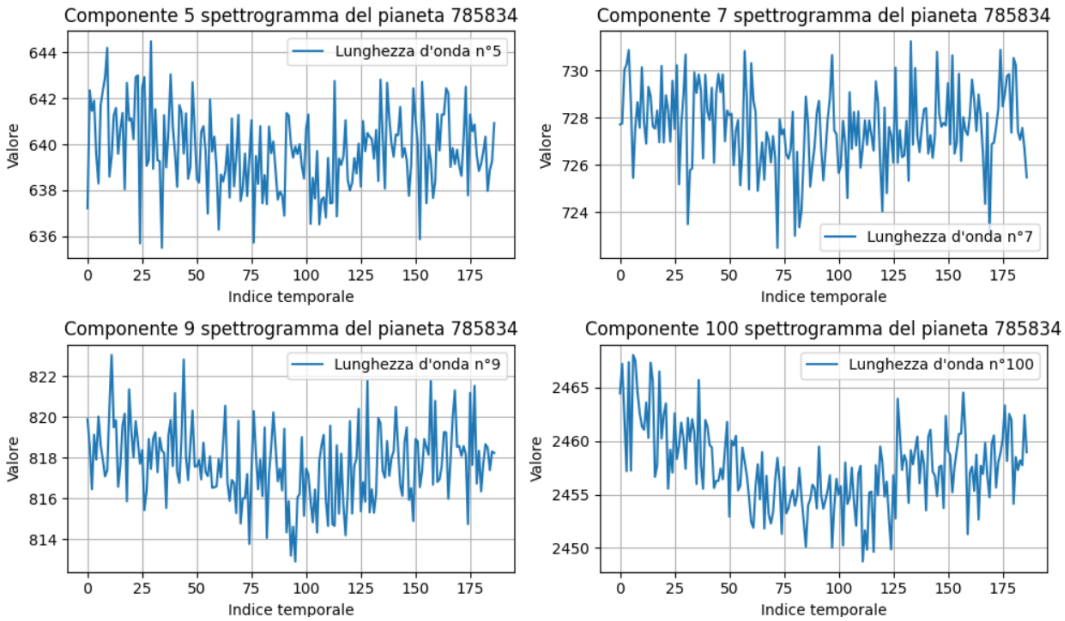

Figura 3.7: Segnale del pianeta 785834 per quattro diverse componenti dello spettrogramma

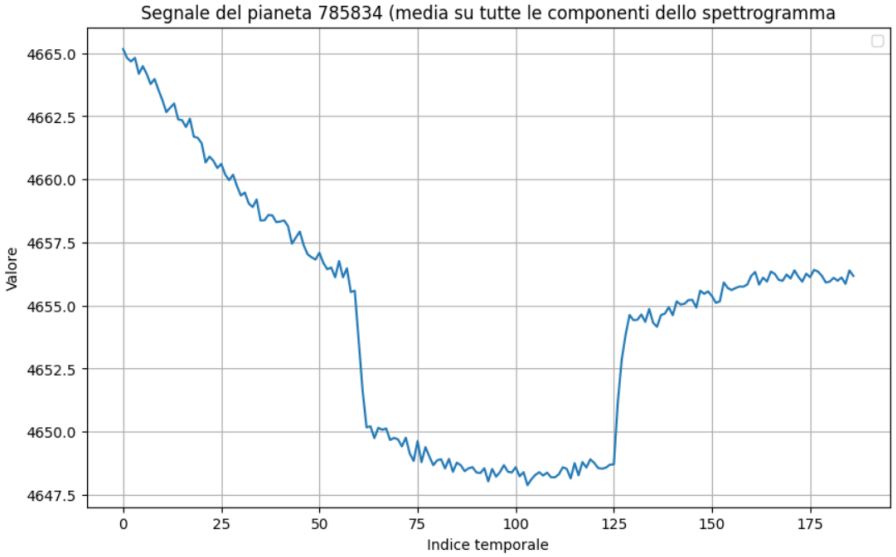

Figura 3.8: Segnale medio del pianeta 785834 (media su tutte le componenti dello spettrogramma)

seconda metà del segnale S_2, derivate che indichiamo rispettivamente con D_1 e D_2. Esse vengono poi normalizzate rispetto al valore massimo assunto dalla derivata.

In questo modo risulta maggiormente evidente il punto in cui il segnale raggiunge le pendenze più elevate in modulo, indipendentemente dalla scala. Vediamo più nel dettaglio i passaggi sopra descritti:

$$D_1(k) = \frac{\frac{dS_1}{dk}}{\max\left(\frac{dS_1}{dk}\right)}$$

$$D_2(k) = \frac{\frac{dS_2}{dk}}{\max\left(\frac{dS_2}{dk}\right)}$$

Il punto di pendenza più negativa viene fatto coincidere con l'inizio della fase di transito, ϕ_1, mentre quello con la pendenza più elevata con la fine della fase di transito, ϕ_2. Dunque si ha:

$$\phi_1 = \arg \min_{0 \leq k < M} D_1(k)$$

$$\phi_2 = \arg \max_{M \leq k < N} D_2(k)$$

3.3 Modello di inferenza Bayesiana

Sono stati testati vari modelli per la stima dei parametri, quali, ad esempio, modelli di apprendimento supervisionato o reti neurali, per approdare infine a modelli di inferenza bayesiana, risultati essere i più promettenti. La spettroscopia di trasmissione, presentata nella sezione 2.1, mostra come sia possibile stimare lo spettro atmosferico del pianeta direttamente analizzando il segnale della stella eclissata da suo pianeta e individuando la profondità di transito generata nella fase oscura per ogni componente dello spettro, senza passare da un addestramento del modello vero e proprio, come invece è richiesto nell'applicazione di tecniche di machine learning. Il modello presentato, dunque, può essere direttamente applicato sui pianeti di test.

Il modello presenta la seguente struttura:

1. Stima della media della profondità di transito applicando metodi per la minimizzazione dei residui sul segnale medio.

2. Stima dell'andamento dello spettro della stella ottenuto dall'elaborazione del segnale medio della fase precedente, indipendente dal passaggio del pianeta.

3. Applicazione di un modello bayesiano con risoluzione del MAP per ogni lunghezza d'onda, adattando l'andamento stellare precedentemente individuato per ognuna di essa.

4. Stima dell'incertezza, supponendo il segnale pressochè gaussiano in un intorno del MAP.

5. Composizione del valore medio e del valore stimato per ogni lunghezza d'onda sia per la profondità di transito sia per l'incertezza associata.

Vediamo più nel dettaglio ognuna di queste fasi. Per la realizzazione della prima parte 1 per ogni pianeta si estrae una media delle componenti relative alle varie lunghezze d'onda per uniformare il segnale nel tempo. Quello che si ottiene è dunque lo spettro luminoso della stella congiunto per tutte le lunghezze d'onda durante il transito del pianeta nei diversi istanti di tempo. Indichiamo con $S(t, \lambda)$ il segnale nell'istante t per la lunghezza d'onda λ e con N il numero di lunghezze d'onda considerate (nel nostro caso 283). Si ha dunque che il segnale medio cercato è dato dalla seguente:

$$S_{\text{medio}}(t) = \frac{1}{N} \sum_{i=1}^{N} S(t, i)$$

In questa fase l'obiettivo è quello di trovare la profondità di transito media, che coincide con la variazione percentuale necessaria per correggere il segnale durante il transito, affinché questo descriva il calo di luminosità del segnale medio. Il parametro viene ottimizzato utilizzando una funzione obiettivo, la quale esegue i seguenti passaggi:

- Trova il segnale che descrive l'andamento della luminosità della stella, eliminando il calo di luminosità dovuto al passaggio del pianeta, grazie al parametro della profondità di transito. Le porzioni del segnale prima e dopo il transito rimangono invariate, mentre durante il transito il segnale originale viene corretto in maniera progressiva: ai bordi del transito, il fattore è applicato in frazioni del parametro decrescenti, per garantire una transizione fluida, nella fase centrale il segnale viene moltiplicato interamente per il fattore di correzione. Questi passaggi risultano necessari in quanto nella fase di individuazione del transito si trovano due istanti temporali di inizio e fine del transito, ma essi non descrivono totalmente il processo. In realtà infatti il pianeta non eclissa la sua stella in un singolo istante ma il processo necessita di alcuni momenti, in cui si ha un calo della luminosità graduale. Nella fase in cui il pianeta si trova interamente davanti alla sua stella il segnale ha pressochè una tendenza piatta, per poi ritornare ai valori iniziali di luminosità in modo graduale, applicando lo stesso ragionamento dell'inizio della fase di oscuramento. Si viene dunque a creare una zona "cuscinetto" attorno agli istanti di inizio e fine del transito individuati nella fase precedente, in cui si vuole che il cambio di luminosità sia graduale.

Indichiamo con i l'istante di ingresso nella fase di transito, e l'istante di egresso dalla fase di transito individuati nella fase precedente, s il fattore di profondità di transito da stimare e con δ la porzione di segnale assunta come zona di transizione attorno agli istanti i e e. Il segnale stellare S_{stella} dopo la rimozione della fase di transito tramite la tecnica sopra descritta è dato dalla seguente:

$$S_{\text{stella}}(t) = \begin{cases} S_{medio}(t) & t < i - \delta \quad \text{oppure} \quad t > e + \delta \\ S_{medio}(t) \cdot \left(1 + \frac{1}{3}s\right) & i - \delta \leq t \leq i \quad \text{oppure} \quad e \leq t \leq e + \delta \\ S_{medio}(t) \cdot \left(1 + \frac{2}{3}s\right) & i < t < i + \delta \quad \text{oppure} \quad e - \delta < t < e \\ S_{medio}(t) \cdot (1 + s) & i + \delta \leq t \leq e - \delta \end{cases}$$

- Individua la migliore modellazione polinomiale del segnale testando gradi diversi del polinomio che descrive la curva. In particolare si testano polinomi che variano dal grado 1 al grado 4, in modo da individuare tra questi quello che meglio descrive la curva.

$$P_{\text{stella}}(t) = \arg\min_{P_n \in \mathcal{P}} \sum_t \left(S_{\text{stella}}(t) - P_n(t)\right)^2, \quad n \in \{1, 2, 3, 4\}$$

- Calcola la discrepanza tra il migliore polinomio individuato e il segnale originale della stella, corretto del transito del pianeta, e ne restituisce l'errore quadratico medio.

Facendo riferimento alla notazione sopra utilizzata l'obiettivo principale di questa fase è quello di individuare s fattore di correzione della profondità di transito del segnale medio, ottenuto tramite l'ottimizzazione della funzione obiettivo. Tale problema di minimo viene risolto con un algoritmo di tipo Nelder-Mead, presentato nel capitolo 2.

Abbiamo visto come in questa fase viene anche individuato il polinomio P_{stella}che meglio descrive l'andamento del segnale, eliminato il calo di luminosità dovuto al transito del pianeta. Tale polinomio verrà chiamato successivamente drift stellare. Vediamo nella figura 3.9 graficamente i passaggi per ottenere il polinomio di drift stellare per uno dei pianeti presi in analisi.

La parte centrale del processo, introdotta precedentemente in 3, vede la vera e propria applicazione dell'inferenza bayesiana. Il parametro della profondità di transito deve essere stimato ora specificatamente per ogni lunghezza d'onda. In questa fase dunque i parametri sono 283, numero delle componenti dello spettro. Essi vengono definiti come distribuzioni normali troncate tra 0 e 0.5, il cui valore atteso risulta essere uguale per tutti ed è dato dal parametro medio ottimizzato

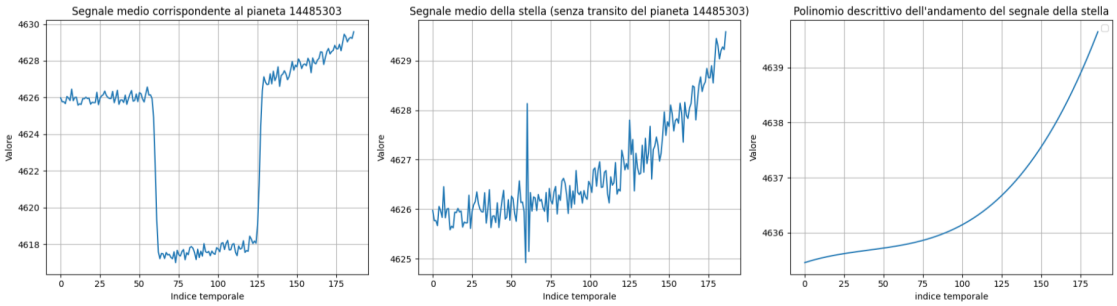

Figura 3.9: Sulla destra è rappresentato il segnale medio per il pianeta 14485303, al centro il segnale medio dopo la rimozione del transito del pianeta, sulla sinistra il polinomio che descrive al meglio il drift stellare.

trovato nella fase precedente. I parametri di profondità non possono assumere valori negativi e per questo è stato necessario inserire il troncamento della distribuzione. La composizione dei vari parametri costituisce la prior del modello. Se indichiamo con s_λ il fattore di profondità di transito per la lunghezza d'onda λ, quantità concettualmente analoga a s precedentemente presentato, allora in formule si ha:

$$s_\lambda \sim \mathcal{N}(s, \sigma^2) \quad \text{con} \quad 0 \leq s_\lambda \leq 0.5, \quad \forall \lambda \in \{1, \ldots, 283\}$$

Osservando la distribuzione dei target per il dataset di train, si vede come la distribuzione sia pressoché gaussiana con una varianza molto piccola e da ciò ne deriva la scelta della distribuzione. La prior creata risulta essere molto informativa, sia per la distribuzione scelta, sia perché viene assegnata ad essa una varianza iniziale anch'essa molto piccola, in modo tale che i valori non si discostino troppo dal valore medio.

La funzione di likelihood riprende la funzione obiettivo della fase precedente: per ogni lunghezza d'onda vengono registrati i residui quadratici generati dal segnale vero e proprio corretto durante il transito tramite il parametro della prior, e il drift stellare riscalato per la lunghezza d'onda presa in analisi. In particolare viene definito un ulteriore parametro q_λ per ciascuna componente dello spettro che rappresenta quanto il segnale debba essere riscalato affinché il drift stellare sia comparabile con la lunghezza d'onda presa in analisi. Vediamo dunque che nuovamente si calcola il segnale della stella per ogni lunghezza d'onda eliminando il transito del pianeta. La tecnica utilizzata è analoga a quella descritta per il segnale medio stellare.

$$S_{\text{stella}}(t,\lambda) = \begin{cases} S(t,\lambda) & t < i-\delta \quad \text{oppure} \quad t > e+\delta \\ S(t,\lambda)\cdot\left(1+\frac{1}{3}s_\lambda\right) & i-\delta \le t \le i \quad \text{oppure} \quad e \le t \le e+\delta \\ S(t,\lambda)\cdot\left(1+\frac{2}{3}s_\lambda\right) & i < t < i+\delta \quad \text{oppure} \quad e-\delta < t < e \\ S(t,\lambda)\cdot(1+s_\lambda) & i+\delta \le t \le e-\delta \end{cases}$$

La likelihood L_λ viene definita tramite la seguente espressione:

$$L_\lambda = -\left(S_{\text{stella}}(t,\lambda) - (P_{\text{stella}}(t) + q_\lambda)\right)^2$$

dove P_{stella} è il drift stellare trovato durante l'analisi del segnale medio, traslato in seguito della quantità q_λ.

La posterior invece è definita come la somma dei residui quadratici della likelihood e della log pdf della prior.

$$Pos_\lambda = log\, pdf(s_\lambda) + L_\lambda$$

Metodi di campionamento Montecarlo, algoritmi come il campionamento di Gibbs o Metropolis–Hastings, prevedono un tempo computazionale eccessivo dato l'alto numero di parametri da stimare e poiché l'obiettivo risulta essere soltanto quello di stimare il valore migliore della posterior e non l'intera distribuzione, si è deciso di utilizzare la tecnica del MAP, presentata nel capitolo 2. Tramite nuovamente il metodo Nelder-Mead si trova il set di parametri che ottimizzano la posterior, con un costo computazionale decisamente più contenuto rispetto ad altri approcci.

Data la mancanza della distribuzione della posterior per i motivi sopra citati, per la stima dell'incertezza si ricorre ad un'approssimazione: Si suppone che in un intorno del MAP la distribuzione della posterior sia pressoché gaussiana. Data tale assunzione, può essere utilizzato il calcolo della varianza di una gaussiana introdotta nel capitolo 2 sezione 2.3. Infine si individua l'intervallo di confidenza al 95%. Tale valore trovato per ciascuna lunghezza d'onda coincide con l'incertezza associata al parametro individuato dal MAP, stima invece della componente dello spettro atmosferico per una determinata lunghezza d'onda.

Dato l'alto rumore associato al segnale suddiviso per le componenti dello spettro, i valori stimati presentano un'elevata variabilità, discostandosi in certi casi eccessivamente dal valore medio stimato in precedenza. Lo step conclusivo prevede dunque per ogni lunghezza d'onda una composizione pesata del valore medio uguale per tutte e il valore stimato per ciascuna componente. In questo modo il valore stimato dalla fase 3 determina più una tendenza della profondità di transito per

ciascuna componente, invece di un valore da tenere in considerazione in termini assoluti. Discorso analogo viene effettuato per la stima dell'incertezza, anch'essa pesata con un valore comune prefissato, trovato empiricamente.

3.4 Correzione bias e offset tramite Discesa del gradiente

Dopo aver ottenuto le predizioni dal modello bayesiano, si applica un'ulteriore fase di correzione ai valori stimati. In particolare, l'obiettivo di questa ultima fase è ridurre l'errore tra le predizioni e i valori di riferimento, utilizzando una matrice che stima i fattori moltiplicativi di correzione e un vettore che salva invece le quantità di cui traslare le predizioni. I due oggetti in questione vengono appresi attraverso un processo di ottimizzazione basato sulla discesa del gradiente. In questa ultima parte, il dataset di train risulta fondamentale in quanto si ha il suo spettro target e si può applicare una vera e propria fase di apprendimento. L'aggiornamento iterativo della matrice e del vettore viene effettuato minimizzando una funzione di costo definita come la somma degli errori quadratici tra le predizioni e i valori reali. Tale matrice individua eventuali bias frequenti per ciascuna componente dello spettro durante la fase di predizione del modello bayesiano. La matrice risulta essere diagonale, per cui ogni elemento sulla diagonale corrisponde al fattore moltiplicativo di correzione per una specifica componente dello spettro, ma non vengono prese in considerazione eventuali correlazioni tra le varie componenti. Le componenti del vettore corrispondono invece agli offset di correzione per ciascuna lunghezza d'onda. Essi vengono infine applicati alle predizioni del dataset di test, come ultima correzione di tipo lineare. Dunque se indichiamo con $\bar{s}$ il vettore che raccoglie le profondità di transito trovate in precedenza per ciascuna lunghezza d'onda, $\bar{s}_{target}$ lo spettro atmosferico target fornito per il dataset di train, D la matrice diagonale che raccoglie le correzioni moltiplicative da applicare a ciascuna lunghezza d'onda, q il vettore che raccoglie gli offset per ciascuna componente, abbiamo che il problema di minimo risulta essere:

$$D_{opt}, q_{opt} = arg\,min_{D,q}\,||\bar{s}_{target} - (\bar{s} \cdot D + q)||_2$$

Il vettore finale di predizioni $\bar{s}_{corrected}$, in seguito all'ottimizzazione di D e q sarà fornito dalla seguente:

$$\bar{s}_{corrected} = \bar{s} \cdot D_{opt} + q_{opt}.$$

Capitolo 4

Visualizzazioni e risultati

Il modello presentato nel capitolo 3 è stato valutato mediante lo score presentato al capitolo 2, in particolare tramite la formula della Log verosimiglianza gaussiana 2.4 per il calcolo della precisione della predizione per ciascuna lunghezza d'onda, le quali vengono combinate tramite la 2.5 per tutto il dataset di test.

Esso ha permesso di giungere allo score di 0.5730, punteggio che in classifica finale risulta al 140º posto su 1152 partecipanti.

Nel seguente capitolo vengono riportati i singoli contributi allo score delle diverse fasi del progetto e alcune visualizzazioni grafiche delle analisi effettuate.

4.1 Test Preprocessing

Per il preprocessing dei segnali sono stati utilizzati i file di calibrazione per correggere le distorsioni strumentali. In particolare si è visto nella sezione 3.1 come siano stati utilizzati i file dei dead frame, dark frame, read frame, linear correction e come siano stati individuati gli hot frame. Vediamo dunque il contributo di ognuno di essi per il raggiungimento dello score finale.

Sono stati eseguiti i seguenti test:

- Dead test: Modello il cui preprocessing non contiene l'eliminazione dei dead frame, sostituiti invece nel modello completo da un'interpolazione lineare rispetto ai frame vicini.

- Hot test: Modello il cui preprocessing non contiene l'eliminazione degli hot frame, anch'essi sostituiti invece nel modello completo da un'interpolazione lineare rispetto ai frame vicini.

- Dark test: Modello il cui preprocessing non contiene la correzione del rumore oscuro descritto dai dark frame.
- Linear correction test: Modello il cui preprocessing non contiene la correzione lineare del segnale dovuta alla perdita capacitiva di lettura dello strumento nel tempo.
- Read test: Modello il cui preprocessing non presenta l'eliminazione dei read frame, fotogrammi che contengono il rumore elettronico legato alla lettura da parte del sensore.
- Complete test: Modello che presenta il preprocessing finale descritto nella sezione 3.1, scelto per la valutazione ufficiale.

Le parti successive, quelle dell'individuazione della finestra di transito, del modello bayesiano e dell'addestramento con il gradient descent per la stima di bias e offset, sono rimaste invariate, come precedentemente spiegate nel capitolo 3. Ognuna delle casistiche test è stata eseguita 5 volte su 200 pianeti per il training, campionati casualmente dai 673 originali, e testata su 100 pianeti scelti anch'essi casualmente tra i rimanenti.

Nella tabella 4.1 vengono riportati gli score per ognuno dei test sopra elencati, ottenuti tramite la valutazione ufficiale presentata nel capitolo 2, in particolare ricordiamo la formula 2.5.

	DEAD TEST	**HOT TEST**	**DARK TEST**	**LINEAR CORRECTION TEST**	**READ TEST**	**COMPLETE TEST**
1° TEST	0.5066	0.5778	0.5773	0.5560	0.5873	0.5889
2° TEST	0.5269	0.5966	0.5962	0.5730	0.5961	0.5972
3° TEST	0.4863	0.5460	0.5462	0.5460	0.5460	0.5466
4° TEST	0.5199	0.5871	0.5875	0.5663	0.5876	0.5890
5° TEST	0.5023	0.5733	0.5725	0.5547	0.5773	0.5731
SCORE MEDIO	**0.5084**	**0.5761**	**0.5759**	**0.5592**	**0.5789**	**0.5790**

Tabella 4.1: Risultati dei test relativi alla fase di preprocessing, valutati con lo score fornito dalle formule 2.4 e 2.5.

Possiamo osservare come i contributi più importanti siano dati dall'utilizzo dei file che salvano i frame morti e i file che correggono la lettura del segnale in modo lineare. Senza la correzione dei frame morti, infatti, lo score medio peggiora del

12%, sottolineando l'importanza di questi ultimi nella fase di preprocessing. La rimozione dei frame morti risulta fondamentale nella fase di individuazione del transito del pianeta in quanto, essendo tale stima basata sulla pendenza della curva e quindi sulla derivata prima della funzione, la presenza di frame scuri può fare in modo che l'algoritmo individui un calo improvviso nella luminosità del segnale dovuto proprio alla presenza di tali frame e che individui proprio in queste oscillazioni improvvise il punto di inizio del transito. Anche nel caso in cui non venga corretta la lettura non lineare del numero di elettroni tramite il polinomio inverso salvato nei file di Linear correction, possiamo osservare grazie al Linear correction test come lo score sia il 3.4% peggiore rispetto a quello di riferimento del Complete test. Gli altri file di calibrazione migliorano il risultato ma il contributo risulta molto piccolo nell'ordine dei millesimi dello score.

Data la poca differenza nello score per alcuni dei test sopra analizzati, si è deciso di verificare che le predizioni ottenute dai diversi preprocessing siano statisticamente differenti. Vediamo più nel dettaglio come funziona il test e quali sono le conseguenze che si possono trarre dalla sperimentazione.

Il test t è un test statistico utilizzato per confrontare le medie di due gruppi e determinare se la loro differenza sia statisticamente significativa. Detto in altri termini, il test t verifica se vale l'ipotesi nulla, H_0, che sostiene che le due medie siano uguali tra loro, oppure l'alternativa, H_1, corrispondente all'ipotesi, invece, che le due medie siano diverse.

In particolare il test calcola la statistica t, che misura la differenza tra le due medie in unità di deviazione standard delle differenze tra i dati. Il valore della statistica t non ha un senso in termini assoluti, ma si può comunque fornirne un'interpretazione:

- valore di t molto alto (positivo o negativo): significa che le differenze tra i due vettori sono grandi rispetto alla variabilità nei dati, rifiutando così l'ipotesi H_0, per accettare invece l'ipotesi alternativa H_1.

- valore di t vicino a 0: significa che i due vettori risultano essere molto simili, suggerendo la validità dell'ipotesi nulla H_0.

L'altro strumento molto utile nelle analisi statistiche è il $p - value$, misura della probabilità usata nel test di ipotesi. Il p-value rappresenta la probabilità che la variabilità nelle medie dei dati campione sia il risultato di pura casualità. Un p-value piccolo porta a rifiutare l'ipotesi nulla H_0, in quanto la probabilità che le medie risultino uguali per puro frutto della casualità è piccola. La soglia tipica del rifiuto dell'ipotesi nulla è 0,05. In sintesi, dunque, se il p-value è inferiore a 0,05, l'ipotesi nulla viene rifiutata in favore dell'ipotesi alternativa, secondo cui almeno una media sarebbe diversa dal resto. Se, invece, il $p - value > 0.05$ si conclude che

le differenze tra i gruppi non siano statisticamente sufficienti per affermare che le distribuzioni siano diverse, per cui si accetta l'ipotesi nulla.

Nel caso in analisi si è deciso di confrontare a coppie i valori della profondità di transito predetti con i vari tipi di preprocessing per vedere se la differenza nello score sia dovuta a pura casualità o se invece sia statisticamente significativa. In particolare i valori ottenuti con il test che presentava il preprocessing completo sono stati confrontati con i valori generati da tutti gli altri test. Nella tabella 4.2 possiamo osservare i risultati della statistica t e il corrispettivo p-value dati dal confronto tra le due coppie di predizioni date dal test con il preprocessing completo e le predizioni di ciascun preprocessing precedentemente testato.

Possiamo chiaramente osservare che i confronti tra il complete test e ciascun preprocessing diano valori di statistica test in modulo elevati e, di conseguenza, il p-value ottenuto risulta essere inferiore alla soglia di 0.05. Ciò conferma l'ipotesi per cui le soluzioni generate dai vari tipi di preprocessing generino predizioni statisticamente differenti tra loro. Questo ci permette di concludere che le varie tecniche di preprocessing introdotte contribuiscano effettivamente a migliorare le predizioni e il rispettivo score, e che ciò non sia frutto della pura casualità dei test effettuati.

4.2 Test del Modello di inferenza Bayesiana

Per osservare le distribuzioni dei dati provenienti dalle due stelle forniti dal dataset di partenza sono state estratte diverse features dai segnali corrispondenti ai pianeti suddivisi tra le due stelle e ne sono stati fatti test t per capire se le distribuzioni dei segnali potessero essere statisticamente simili.

Nell'immagine 4.1 possiamo osservare graficamente le distribuzioni suddivise per stella 0 e stella 1 delle features: media lungo l'asse spettrale, varianza lungo l'asse spettrale, skewness spettrale e temporale, kurtosis spettrale e temporale, i quantili al 5% spettrali e temporali, i quantili al 95% spettrali e temporali e infine il rapporto tra segnale e rumore (SNR). Sono stati presi in considerazione per tutte le distribuzioni soltanto i dati provenienti dal sensore AIRS-CH0.

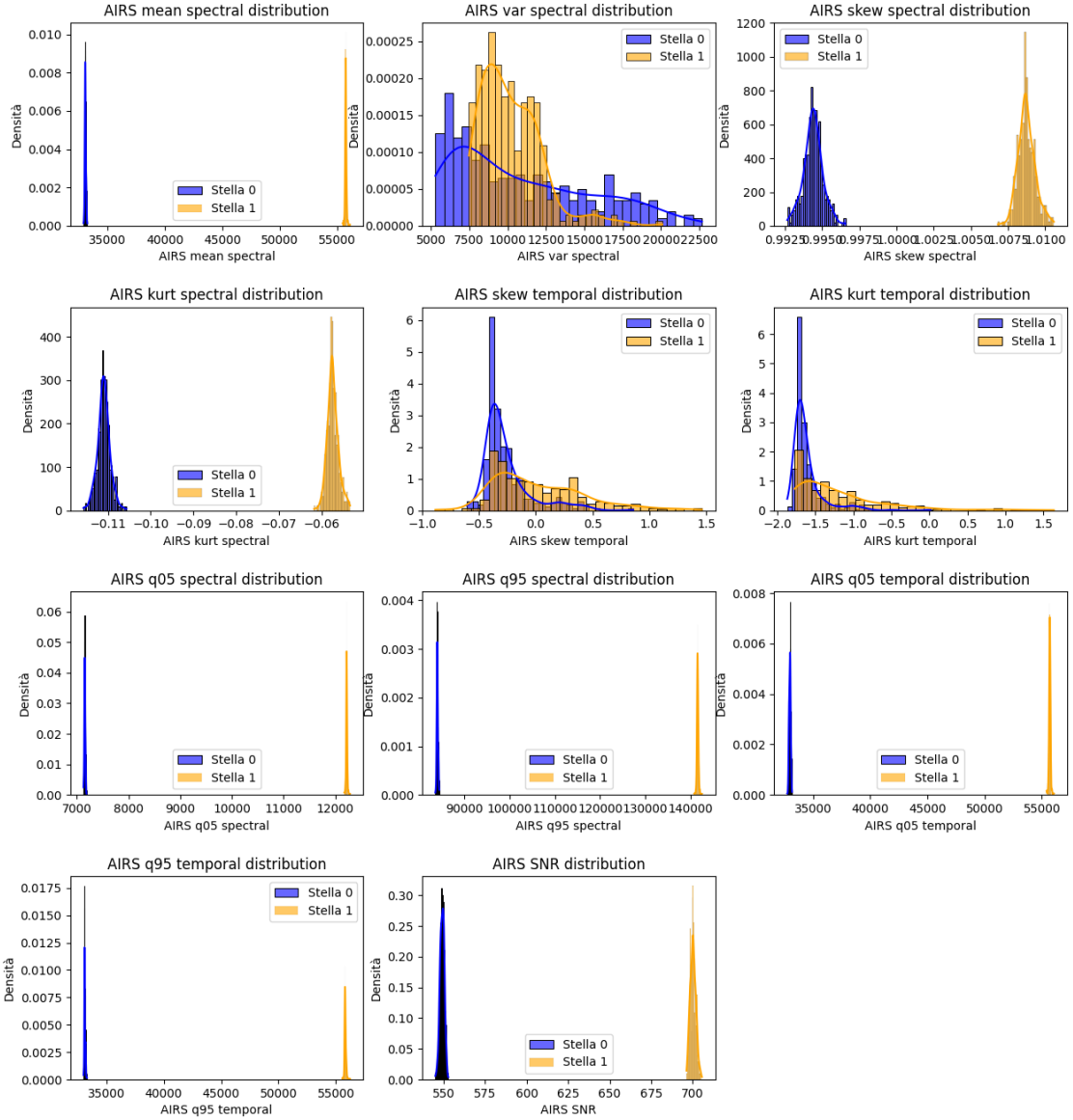

Figura 4.1: Distribuzioni di varie features estratte dal dataset di train suddiviso per la stella 0 e stella 1.

	COMPLETE - DEAD TEST		COMPLETE - HOT TEST		COMPLETE - DARK TEST		COMPLETE - LINEAR CORRECTION TEST		COMPLETE - READ TEST	
	Statistica t-test	**p-value**	**Statistica t-test**	**p-value**	**Statistica t-test**	**p-value**	**Statistica t-test**	**p-value**	**Statistica t-test**	**p-value**
1° TEST	-118.9845	≈ 0.00 (<1e-5)	-95.0936	≈ 0.00 (<1e-5)	-105.4448	≈ 0.00 (<1e-5)	91.3472	≈ 0.00 (<1e-5)	-7.5648	≈ 0.00 (<1e-5)
2° TEST	-117.8765	≈ 0.00 (<1e-5)	-89.0045	≈ 0.00 (<1e-5)	-96.6349	≈ 0.00 (<1e-5)	92.5967	≈ 0.00 (<1e-5)	-4.7997	≈ 0.00 (<1e-5)
3° TEST	-120.8709	≈ 0.00 (<1e-5)	-109.1765	≈ 0.00 (<1e-5)	-106.2401	≈ 0.00 (<1e-5)	93.3878	≈ 0.00 (<1e-5)	-6.6601	≈ 0.00 (<1e-5)
4° TEST	-115.9865	≈ 0.00 (<1e-5)	-105.6519	≈ 0.00 (<1e-5)	-106.0023	≈ 0.00 (<1e-5)	92.2387	≈ 0.00 (<1e-5)	-6.9560	≈ 0.00 (<1e-5)
5° TEST	-118.9765	≈ 0.00 (<1e-5)	-94.9962	≈ 0.00 (<1e-5)	-103.3261	≈ 0.00 (<1e-5)	91.8473	≈ 0.00 (<1e-5)	-7.8264	≈ 0.00 (<1e-5)

Tabella 4.2: Risultati del t-test (statistica test e p-value) ottenuti dal confronto delle predizioni del complete test rispetto agli altri preprocessing testati.

Da tali analisi si è osservato come i dati associati ai pianeti delle due stelle fossero molto differenti, con relative distribuzioni nettamente diverse sotto diversi punti di vista, quali media, mediana, varianza, quantili e struttura delle code.

Effettuare dunque inferenza bayesiana direttamente sul dataset di test è risultata essere la scelta migliore, nonostante fosse fornito lo spettro atmosferico target del dataset di train. Lavorare con pianeti provenienti da due sole stelle del train non ha permesso di creare modelli di apprendimento supervisionato robusti e che fossero completamente generalizzabili alle stelle ignote utilizzate durante il test, date le loro distribuzioni molto diverse le une dalle altre.

Per risolvere il problema di minimo della funzione obiettivo introdotta in 3.3, si è scelto di utilizzare il metodo di Nelder-Mead. Vediamo ora alcune rappresentazioni grafiche estratte durante i test sulla convergenza del modello. Nella figura 4.2 possiamo osservare come converge il valore della funzione obiettivo e il valore della profondità di transito estratta per tre diversi pianeti con il Nelder-Mead method.

Possiamo osservare come il metodo giunga a convergenza con poche iterazioni (≈ 30), e come esso sia stabile al variare di quest'ultime. Vediamo infatti che il metodo inizia a stabilizzarsi verso la soluzione più corretta già dopo 20 iterazioni e non presenta oscillazioni anomale. Questo indica che il metodo di Nelder-Mead durante l'ottimizzazione procede verso un minimo locale stabile in modo efficiente, senza rimanere bloccato in minimi locali poco significativi. L'assenza di picchi suggerisce inoltre che il problema di ottimizzazione è ben condizionato e che la scelta dei parametri iniziali risulta essere adeguata per garantire una convergenza efficiente.

Per la scelta della prior sono state prese in considerazione 3 diverse distribuzioni: normale troncata, gamma e beta, delle quali vediamo una rappresentazione grafica nell'immagine 4.3. Esse costituiscono tutte prior informative, i cui valori si concentrano attorno allo zero, come ci si aspetta dalla distribuzione dei target osservata.

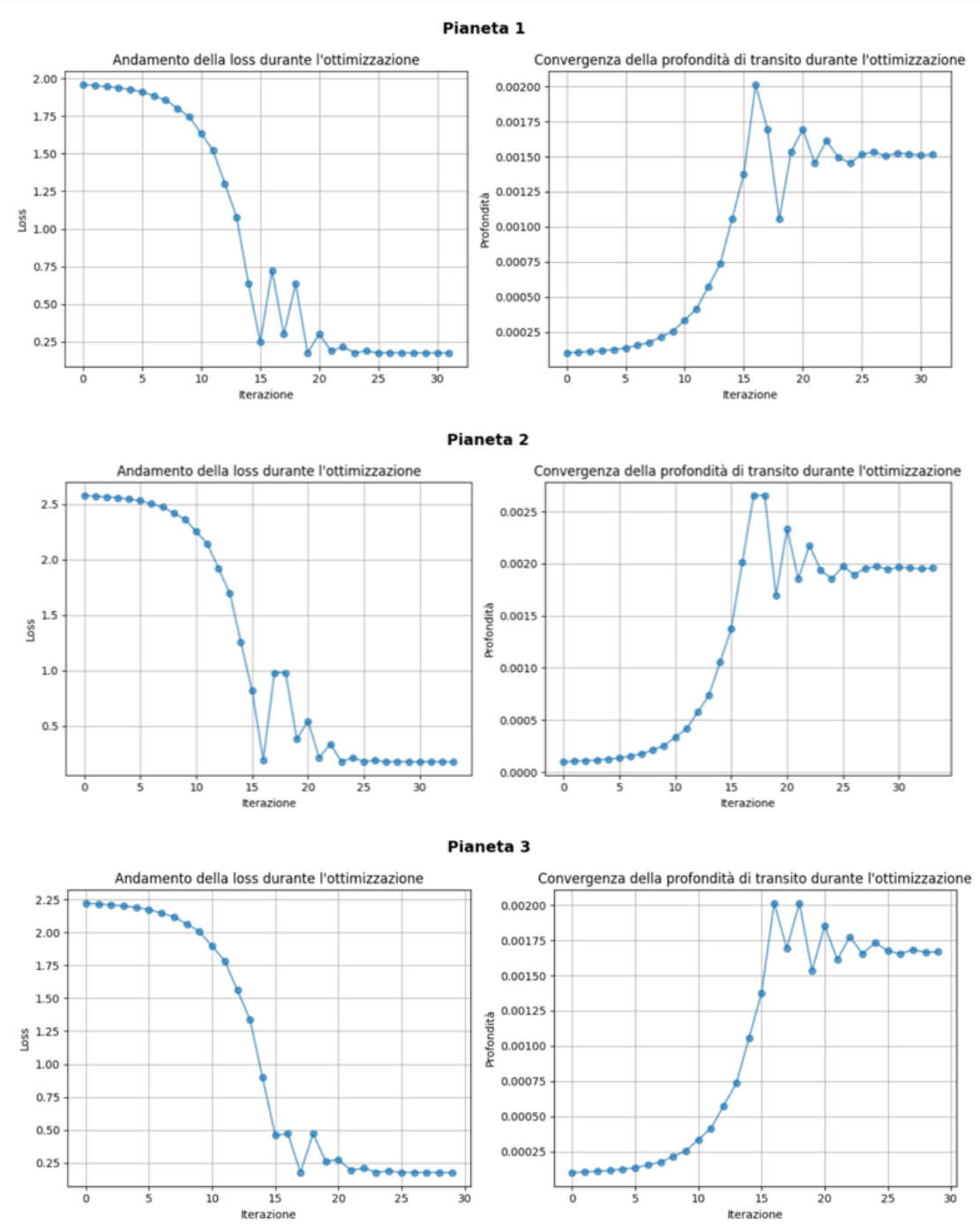

Figura 4.2: Rappresentazione grafica della loss e della profondità di transito media su tutte le lunghezze d'onda durante l'ottimizzazione della funzione obiettivo con il metodo Nelder-Mead, per tre diversi pianeti.

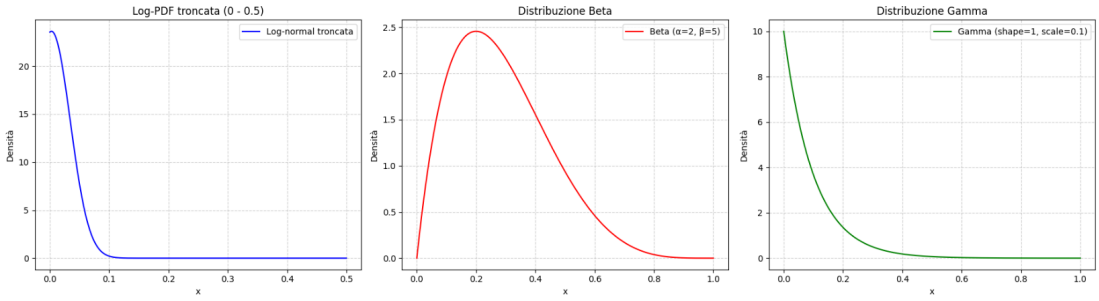

Figura 4.3: Rappresentazione grafica delle funzioni di densità di una normale ($\mu = 0.02$ $\sigma = 0.01$), di una Beta ($\alpha = 2$ e $\beta = 5$) e di una Gamma (shape = 1, scale = 0.1)

Nella tabella 4.3 possiamo osservare gli score ottenuti dai modelli di inferenza bayesiana in cui è implementata una diversa distribuzione di prior ogni volta, tra le tre segnalate precedentemente. Il test è stato eseguito per 5 volte su un campione casuale di 200 pianeti sui 673 possibili e testato su 100 pianeti scelti casualmente dai rimanenti.

	TRUNC NORM (0 – 0.5)	**BETA** (ALPHA = 2, BETA = 5)	**GAMMA** (SHAPE = 1, SCALE = 0.1)
1° TEST	0.5972	0.5877	0.5891
2° TEST	0.5889	0.5873	0.5773
3° TEST	0.5890	0.5877	0.5876
4° TEST	0.5466	0.5460	0.5460
5° TEST	0.5731	0.5726	0.5725
SCORE MEDIO	**0.5790**	**0.5763**	**0.5747**

Tabella 4.3: Score relativi ai 5 test effettuati per ciascuna distribuzione di prior presa in analisi (normale troncata (0, 0.5), Beta (2,5) e Gamma (1, 0.1)).

Nell'ultima riga della tabella in particolare possiamo osservare lo score medio ottenuto sui 5 test effettuati. Si può notare come tutte siano prior molto valide che permettono di giungere a risultati simili, nell'ordine delle centinaia. La prior coincidente con la normale troncata comunque risulta essere la più promettente, dato il suo dominio più limitato rispetto alle altre due distribuzioni.

Data la poca differenza nello score ottenuta per le varie casistiche di prior si è deciso di effettuare un test-t per vedere se le differenze tra i valori predetti tra le tre prior fossero statisticamente differenti.

Nel caso in analisi si è deciso di confrontare a coppie i valori della profondità di transito predetti. In particolare i valori ottenuti con la prior Normale troncata sono stati confrontati prima con i valori generati dalla prior Beta e poi dalla prior Gamma. Nella tabella 4.4 possiamo osservare i risultati della statistica t e il corrispettivo p-value per le due coppie di dati per ogni test precedentemente effettuato.

	TRUNC NORM - BETA		**TRUNC NORM - GAMMA**	
	Statistica t-test	**p-value**	**Statistica t-test**	**p-value**
1° TEST	-10.9904	≈ 0 (<1e-5)	-7.5524	≈ 0 (<1e-5)
2° TEST	-8.3546	≈ 0 (<1e-5)	-4.8135	≈ 0 (<1e-5)
3° TEST	-9.7860	≈ 0 (<1e-5)	-6.6506	≈ 0 (<1e-5)
4° TEST	-8.9864	≈ 0 (<1e-5)	-5.2614	≈ 0 (<1e-5)
5° TEST	-9.6248	≈ 0 (<1e-5)	-6.3513	≈ 0 (<1e-5)

Tabella 4.4: Risultati del t-test (statistica test e p-value) ottenuti dal confronto delle predizioni nei casi di prior: Normale Troncata - Beta, Normale Troncata - Gamma.

Si può chiaramente vedere come il valore della statistica test t sia elevato in modulo, quantità che presuppone già una differenza statistica tra i vari gruppi, confermata dal valore molto piccolo di p-value ottenuto. Esso risulta infatti minore di 0.05, soglia scelta precedentemente, al di sotto della quale i due vettori possono considerarsi statisticamente differenti.

4.3 Test correzione bias e offset tramite Discesa del gradiente

L'ultimo step del modello ha visto l'implementazione di un algoritmo di allenamento del dataset di train tramite la Discesa del gradiente [29] per l'individuazione di coefficienti moltiplicativi e additivi di correzione per ciascuna componente dello spettro. In questa sezione ci si concentra sulla validazione statistica della stabilità della matrice che riporta eventuali bias e il vettore che contiene invece offset per

ogni lunghezza d'onda, per verificare che la loro stima sia robusta e non frutto della casualità.

L'approccio utilizzato per l'analisi si articola nei seguenti step:

1. Divisione del dataset in sottoinsiemi casuali (80% training set, 20% validation set)

2. Esecuzione dell'algoritmo di Discesa del gradiente sul sottoinsieme di training e validazione sul resto del dataset.

3. Esecuzione dei primi due step 30 volte e analisi delle distribuzioni dei valori finali per dimostrare che essi siano coerenti tra le diverse iterazioni e che giungano all'individuazione degli stessi valori di bias e offset.

4. Esecuzione del test ANOVA (analysis of variance) per vedere se i valori della matrice e del vettore ottenuti nei diversi bootstrap siano statisticamente diversi o se possano essere considerati stabili tra le varie iterazioni.

Nell'immagine 4.4 possiamo osservare i valori assunti dai coefficienti moltiplicativi della matrice di correzione per ciascuna componente dello spettro con la relativa varianza associata sotto forma di error-bar e, allo stesso modo, le componenti del vettore di correzione che determinano invece una traslazione dello spettro predetto per ciascuna lunghezza d'onda con relativa incertezza.

Possiamo osservare come i valori stimati per ciascuna componente dello spettro nelle varie iterazioni del bootstrap siano essenzialmente sempre gli stessi. Sia per i coefficienti di correzione dei bias sia soprattutto per i valori del vettore degli offset si può notare come la varianza associata sia molto piccola, evidenziando la poca variabilità tra i vari bootstrap.

Il test statistico ANOVA [30] (Analysis of variance) ha come obiettivo quello di individuare differenze statistiche tra i gruppi presi in analisi. In particolare usa come ipotesi nulla H_0 il fatto che le medie dei vari gruppi considerati siano uguali e ipotesi alternativa H_1 che almeno una media sia diversa dalle altre.

Il test in particolare si concentra sul confronto tra due tipi di varianze:

- Varianza tra gruppi: calcola la differenza tra la media generale e la media di ciascun gruppo

- Varianza intra gruppi: calcola la differenza in ciascun gruppo tra i singoli valori e la media del gruppo di appartenenza.

Viene calcolato infine l'F-statistic dato dal rapporto delle due varianze sopra elencate. Se l'F-statistic è grande significa che la varianza tra i gruppi è molto più

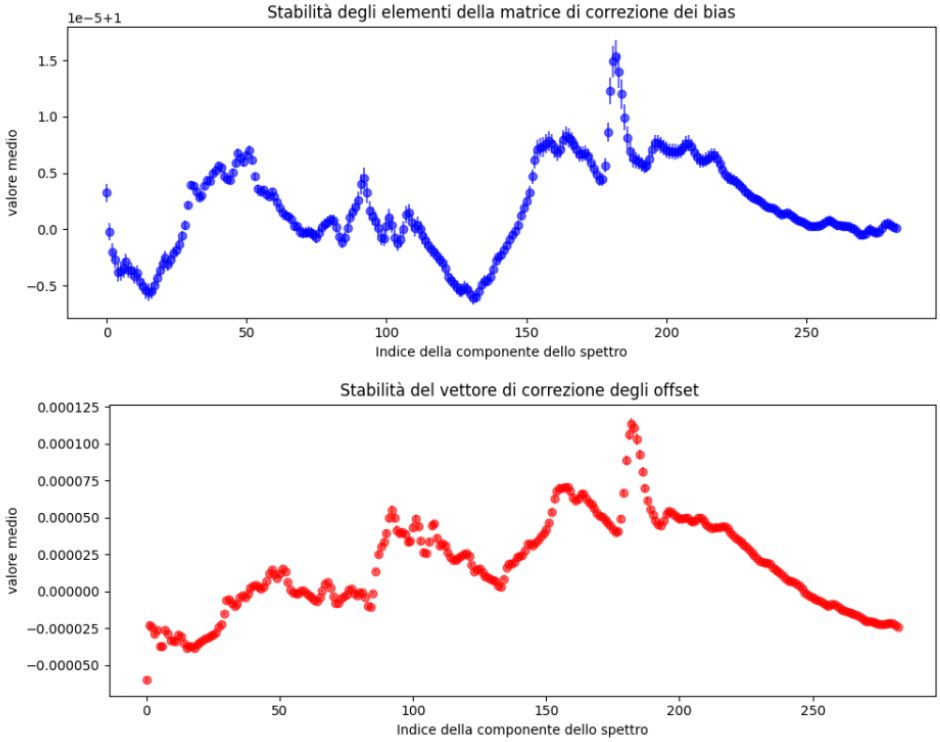

Figura 4.4: Sopra: Valori assunti dalla matrice di bias per ciascuna componente dello spettro con relativa varianza. Sotto: Valori assunti dal vettore degli offset per ciascuna componente dello spettro con relativa varianza

alta della varianza interna e perciò le distribuzioni dei gruppi sono da considerarsi diverse; al contrario, se il valore di F-statistic è più piccolo, le distribuzioni sono da considerarsi uguali.

Nel caso preso in analisi, si vuole osservare se i coefficienti diagonali della matrice di bias risultino simili tra i vari bootstrap, dimostrando così che il metodo della Discesa del Gradiente individua sempre le stesse correzioni da applicare alle predizioni. Discorso analogo viene fatto per il vettore degli offset. In particolare si applica l'ANOVA in modo tale che all'interno di un gruppo siano inserite tutte le componenti dei vettori per una specifica lunghezza d'onda generate da tutte le iterazioni del bootstrap. L'obiettivo quindi è quello di capire se la correzione applicata alle varie componenti dello spettro risulta simile tra gli elementi dello stesso gruppo ma diversa dalle correzioni applicate dagli altri gruppi, cioè per le altre lunghezze d'onda. Nella tabella 4.5 vengono riportati i risultati dei test ANOVA per entrambi gli oggetti.

	F-STATISTIC	**P-VALUE**
MATRICE BIAS	824.5045	≈ 0.000 (< 0.05)
VETTORE OFFSET	8346.8788	≈ 0.000 (< 0.05)

Tabella 4.5: F-statistic e p-value da test ANOVA sulle componenti dei coefficienti moltiplicativi della matrice di correzione e il vettore degli offset

Possiamo osservare chiaramente per entrambi gli oggetti presi in analisi come il valore dell'F-statistic sia molto alto, cosa che suggerisce una differenza sostanziale tra le correzioni applicate alle varie componenti, supposizione confermata dal p-value che in entrambi i casi assume un valore sotto alla soglia 0.05. I test ci suggeriscono dunque di rifiutare l'ipotesi nulla e accettare invece l'ipotesi H_1 che sostiene una differenza statisticamente significativa tra le correzioni applicate alle varie componenti e non frutto della semplice casualità.

Capitolo 5

Conclusioni

La spettroscopia di trasmissione si conferma essere una strategia promettente per l'individuazione di nuovi pianeti e per derivarne la composizione atmosferica, in quanto stima in maniera indiretta lo spettro atmosferico tramite la modellazione della profondità di transito. Tale tecnica permette di adattare conoscenze teoriche di dominio riguardanti i segnali stellari ad algoritmi di facile implementazione.

Il preprocessing del segnale, grazie soprattutto alla rimozione di pixel morti e alla correzione lineare della lettura del segnale da parte del sensore, ha permesso l'utilizzo di un algoritmo di individuazione della fase di transito molto semplice, basato solo sullo studio della derivata prima, cosa che non sarebbe stata possibile se il segnale avesse presentato ancora diffetti evidenti o oscillazioni anomale.

Dalle analisi effettuate, inoltre, possiamo confermare che l'utilizzo dell'inferenza bayesiana risulta essere uno strumento importante per la risoluzione di problemi di questo tipo. Per prima cosa, infatti, non necessita di per sè di un ampio dataset di partenza per allenare il modello, poichè i parametri vengono stimati direttamente per il singolo pianeta senza utilizzare informazioni precedentemente apprese durante la fase di training. L'altro aspetto fondamentale è stato anche l'utilizzo di prior informative basate sul segnale medio lungo l'asse delle lunghezze d'onda, in quanto essendo poco rumoroso permetteva di giungere a conclusioni precise, molto utili come punto di partenza per il modello di inferenza. Anche l'utilizzo del MAP si è rivelato fondamentale in casi di questo tipo in cui il numero di parametri da stimare è elevato. Combinato con il metodo di minimizzazione di Nelder-Mead, tale tecnica ha permesso di giungere al punto di massimo della posterior per ciascuna componente dello spettro in poche iterazioni. Allo stesso modo anche l'approssimazione gaussiana che si è scelta per stimare la varianza si è dimostrata efficiente, mantenendo i costi computazionali molto bassi.

Anche l'ultimo step del modello ha rivelato la sua importanza durante la fase di test. Le correzioni moltiplicative e addittive apprese tramite la Discesa del gradiente, nonostante siano per alcune componenti dello spettro molto piccole, sono risultate molto stabili, come confermato dall'analisi della varianza dei campioni generati dal bootstrap e dai test ANOVA e del p-value, che hanno confermato la loro solida validà statistica.

Per quanto riguarda le implementazioni future del modello mostrato in questo progetto, esse potrebbero innanzitutto trattare anche l'utilizzo dei dati provenienti dall'altro sensore FGS1, eventualmente dopo un approfondito processo di rimozione del rumore. Per quanto riguarda invece la parte relativa al modello si è osservato come il MAP di per sé sia un'ottima soluzione, ma non permette di stimare l'incertezza associata, motivo per il quale in questo progetto si è fatta un'assunzione importante per cui la curva viene considerata pressoché gaussiana in un intorno del MAP. Possibili miglioramenti potrebbero riguardare la stima della distribuzione a posteriori tramite nuovi algoritmi che abbiano un costo computazionale ridotto, dato l'alto numero di parametri da stimare. Avere la distribuzione della posterior, o almeno saper campionare da essa, permetterebbe di approfondire in maniera più dettagliata e precisa l'incertezza associata alle soluzioni.

Bibliografia

[1] Aurélien Falco, Tiziano Zingales, William Pluriel e Jérémy Leconte. «Toward a multidimensional analysis of transmission spectroscopy-I. Computation of transmission spectra using a 1D, 2D, or 3D atmosphere structure». In: *Astronomy & Astrophysics* 658 (2022), A41 (cit. a p. 4).

[2] Nikku Madhusudhan, Marcelino Agúndez, Julianne I Moses e Yongyun Hu. «Exoplanetary atmospheres—chemistry, formation conditions, and habitability». In: *Space science reviews* 205 (2016), pp. 285–348 (cit. a p. 4).

[3] access 29/01/2025. URL: `https://www.kaggle.com/competitions/ariel-data-challenge-2024` (cit. alle pp. 5, 12).

[4] Flavio Giobergia, Alkis Koudounas e Elena Baralis. «Reconstructing Atmospheric Parameters of Exoplanets Using Deep Learning». In: *2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT)*. 2023, pp. 1–6. DOI: `10.1109/AICT59525.2023.10313185` (cit. a p. 5).

[5] Alkis Koudounas, Flavio Giobergia e Elena Baralis. «Bad Exoplanet! Explaining Degraded Performance when Reconstructing Exoplanets Atmospheric Parameters». In: *NeurIPS 2023 AI for Science Workshop*. 2023. URL: `https://openreview.net/forum?id=9Z4XZOhwiz` (cit. a p. 5).

[6] Alkis Koudounas, Flavio Giobergia e Elena Baralis. «Ex(o)plain: Subgroup-Level Analysis of Exoplanet Atmospheric Parameters». In: *IEEE Access* 12 (2024), pp. 139773–139788. DOI: `10.1109/ACCESS.2024.3466919` (cit. a p. 5).

[7] Laura Kreidberg. «Exoplanet Atmosphere Measurements from Transmission Spectroscopy and other Planet-Star Combined Light Observations». In: *Handbook of Exoplanets* (2018) (cit. a p. 6).

[8] Florian Debras, Baptiste Klein, Jean-Francois Donati e Thea Hood. «Characterizing exoplanet atmospheres through transmission spectroscopy with SPIRou». In: *Monthly Notices of the Royal Astronomical Society* (2023) (cit. a p. 6).

[9] F. Murgas, G. Chen, E. Pallé, L. Nortmann e G. Nowak. «The GTC exoplanet transit spectroscopy survey X. Stellar spots versus Rayleigh scattering: the case of HAT-P-11b». In: *Astronomy & Astrophysics* 622 (feb. 2019) (cit. a p. 6).

[10] Alexandre Branco, Pedro Machado, Olivier Demangeon, Tomás Azevedo Silva, Sarah A. Jaeggli, Thomas Widemann e Paolo Tanga. «Transmission Spectroscopy Along the Transit of Venus: A Proxy for Exoplanets Atmospheric Characterization». In: *Atmosphere* 12 (2024) (cit. a p. 6).

[11] Hossein Rahmati, Stefan Czesla, Sara Khalafinejad e Paul Mollière. «Transmission spectroscopy of WASP-7 b with UVES». In: *Astronomy & Astrophysics* 668 (nov. 2022) (cit. a p. 6).

[12] J. A. Nelder e R. Mead. «A Simplex Method for Function Minimization». In: *The Computer Journal* 7 (gen. 1965), pp. 308–313 (cit. a p. 6).

[13] Andrew Gelman, John B. Carlin, Hal S. Stern e Donald B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman Hall/CRC, 2003 (cit. alle pp. 8, 10).

[14] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009 (cit. a p. 9).

[15] Lixin Dou e R J W Hodgson. «Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation. I». In: *Inverse problem* 11 (1995) (cit. a p. 9).

[16] Marcelo Pereyra. «Maximum-a-Posteriori Estimation with Bayesian Confidence Regions». In: *Siam J. Imaging Sciences* 10 (2017), pp. 285–302 (cit. a p. 9).

[17] Filip Tronarp, Simo Särkkä e Philipp Hennig. «Bayesian ODE Solvers: The Maximum A Posteriori Estimate». In: *Statistics and Computing* 31, 23 (2021) (cit. a p. 10).

[18] Sergios Agapiou, Martin Burger, Masoumeh Dashti e Tapio Helin. «Sparsity-promoting and edge-preserving Maximum a Posteriori estimators in non-parametric Bayesian inverse problems». In: *Inverse Problems* 34 (feb. 2018) (cit. a p. 10).

[19] Remi Laumont, Valentin De Bortoli, Andres Almansa, Julie Delon, Alain Durmus e Marcelo Pereyra. «On Maximum-a-Posteriori estimation with Plug Play priors and stochastic gradient descent». In: *Journal of Mathematical Imaging and Vision* 65 (gen. 2023), pp. 140–163 (cit. a p. 10).

[20] Joseph B. Nagel e Bruno Sudret. «Spectral likelihood expansions for Bayesian inference». In: *Journal of Computational Physics* 309 (mar. 2016), pp. 267–194 (cit. a p. 10).

[21] H. M. Tabernero, E. Marfil, D. Montes e J. I. González Hernández. «STEPARSYN: A Bayesian code to infer stellar atmospheric parameters using spectral synthesis». In: *Astronomy & Astrophysics* 657 (gen. 2022) (cit. a p. 11).

[22] C. Defay, M. Deleuil e P. Barge. «A Bayesian method for the detection of planetary transits». In: *Astronomy & Astrophysics* 365 (gen. 2001), pp. 330–340 (cit. a p. 11).

[23] I. P. Waldmann, G. Tinetti, M. Rocchetto, E. J. Barton, S. N. Yurchenko e J. Tennyson. «Tau-REx I: A next generation retrieval code for exoplanetary atmospheres». In: *The Astrophysical Journal* 802, Number 2 (apr. 2015) (cit. a p. 11).

[24] Ahmed F. Al-Refaie, Quentin Changeat, Ingo P. Waldmann e Giovanna Tinetti. «TauREx III: A fast, dynamic and extendable framework for retrievals». In: *The Astrophysical Journal* 917, Number 1 (ago. 2021) (cit. a p. 11).

[25] P.G.J. Irwin et al. «The NEMESIS planetary atmosphere radiative transfer and retrieval tool». In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 109, Number 6 (apr. 2008), pp. 1136–1150 (cit. a p. 11).

[26] Siddharth Gandhi, Nikku Madhusudhan, George Hawker e Anjali Piette. «HyDRA-H: Simultaneous Hybrid Retrieval of Exoplanetary Emission Spectra». In: *The Astrophysical Journal* 158, Number 6 (nov. 2019) (cit. a p. 11).

[27] Matthew C. Nixon, Luis Welbanks, Peter McGill e Eliza M.-R. Kempton. «Methods for Incorporating Model Uncertainty into Exoplanet Atmospheric Analysis». In: *The Astrophysical Journal* 966, Number 2 (mag. 2024) (cit. a p. 11).

[28] Ariel Data Challenge 2024. Ultimo accesso: 5 Febbraio 2025. 2024. URL: `https://www.kaggle.com/competitions/ariel-data-challenge-2024/overview` (cit. a p. 22).

[29] Alkis Koudounas, Flavio Giobergia e Elena Baralis. «Time-of-Flight Cameras in Space: Pose Estimation with Deep Learning Methodologies». In: *2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT)*. 2022, pp. 1–6. DOI: `10.1109/AICT55583.2022.10013574` (cit. a p. 39).

[30] Lars St, Svante Wold et al. «Analysis of variance (ANOVA)». In: *Chemometrics and intelligent laboratory systems* 6.4 (1989), pp. 259–272 (cit. a p. 40).