

Fusion de données par filtrage adapté pour la détection d'exoplanètes en imagerie directe

Jules DALLANT Éric THIÉBAUT Maud LANGLOIS Olivier FLASSEUR

Univ. de Lyon, Univ. Lyon1, ENS de Lyon, CNRS, CRAL UMR 5574, 69230 Saint-Genis-Laval, France

Résumé – Détecter des exoplanètes par imagerie directe est une tâche extrêmement ardue en raison du très haut contraste entre ces sources et leur étoile hôte. L'algorithme PACOME combine des observations à des époques différentes d'un même système extrasolaire en prenant en compte le mouvement képlérien des sources qu'il cherche à détecter et permet d'atteindre une sensibilité encore inégalée. Dans ce papier, nous étendons le formalisme de cette méthode à une approche par filtrage adapté pour en dériver un critère optimal interprétable en terme de rapport signal sur bruit. Nous illustrons l'efficacité de la méthode sur des simulations réalistes de sources ajoutées dans des données réelles.

Abstract – Detecting exoplanets by direct imaging is an extremely difficult task because of the extremely high contrast between them and their host star. The PACOME algorithm combines observations at different epochs of the same extrasolar system by taking into account the Keplerian motion of the sought sources and achieves unprecedented sensitivity. In this paper, we extend the formalism of this method to a matched filter approach to derive an optimal criterion interpretable in terms of signal-to-noise ratio. We illustrate the efficiency of the method on realistic simulations of sources added in real data.

1 Introduction

Seulement une douzaine de planètes extrasolaires ont à ce jour été détectées par imagerie directe. Ceci s'explique par la présence de termes de nuisance (bruits et fuites de lumière stellaire spatialement corrélées) qui dominent les signaux des objets d'intérêt. Il est donc nécessaire de recourir à des méthodes de démixage de sources afin d'extraire le signal utile des exoplanètes. Les méthodes les plus courantes sont basées sur l'imagerie différentielle angulaire (ADI) et spectrale (ASDI) qui font usage de la rotation apparente du champ de vue lors de la séquence d'observations (et/ou de la variété spectrale des données) pour construire un modèle des termes de nuisance qui est soustrait aux données, les sources faibles étant détectées dans les résidus ainsi obtenus. En particulier, l'algorithme PACO [8] apprend localement la statistique des nuisances à partir des corrélations spatio-spectrales. Afin d'améliorer la sensibilité de détection, nous avons développé l'algorithme PACOME [3] qui tire partie du formalisme statistique de PACO, basée sur le maximum de vraisemblance. PACOME combine différentes observations d'un même système extrasolaire en maximisant une fonction de coût tenant compte du mouvement orbital des sources qu'il cherche à détecter.

Dans ce papier, nous étendons l'approche de PACOME au formalisme du filtrage adapté qui permet l'obtention d'un score de détection interprétable en terme de rapport signal à bruit (RSB), nous discutons des garanties statistiques de cette approche et nous démontrons l'efficacité de la méthode sur des sources simulées dans des données réelles.

2 Dérivation d'un RSB multiépoque

Nous présentons le formalisme du filtrage adapté permettant de combiner linéairement différentes observations ASDI de manière optimale. Il permet simultanément de détecter des compagnons orbitant autour d'étoiles et de dériver les paramètres décrivant leurs orbites.

2.1 Modèle direct des observations

Les données ASDI classiques sont des séries temporelles d'images coronographiques, réparties sur plusieurs canaux spectraux, et observées à différentes dates appelées *époques*. La position apparente $\theta_t(\mu)$ sur le ciel d'une source ponctuelle décrivant un mouvement képlérien à l'époque d'observation t dépend seulement de ses éléments orbitaux μ et de la masse totale du système. Nous modélisons la trame k d'un jeu de données $\mathbf{r}_{t,\ell,k} \in \mathbb{R}^N$ prise à l'époque t , au canal spectral ℓ comme dans [3] :

$$\mathbf{r}_{t,\ell,k} = \alpha_{t,\ell} \mathbf{h}_{t,\ell}(\theta_t(\mu)) + \mathbf{f}_{t,\ell,k}, \quad (1)$$

avec $\mathbf{h}_{t,\ell}(\theta) \in \mathbb{R}^N$ la fonction d'étalement de point (FEP) hors axe centrée à la position θ et identique pour toutes les trames k de la même époque, $\alpha_{t,\ell}$ l'intensité de la source (valant 0 en l'absence de source) à l'époque t et au canal spectral ℓ et $\mathbf{f}_{t,\ell,k}$ les termes de nuisance.

Le nombre de photons collectés par chaque pixel du détecteur étant suffisant, nous supposons que les termes de nuisance suivent une distribution gaussienne multivariée [4, 5]. Les images des différents canaux spectraux et époques sont mutuellement indépendantes et quasi-identiquement distribuées :

$$\mathbf{f}_{t,\ell,k} \sim \mathcal{N}(\bar{\mathbf{f}}_{t,\ell}, \sigma_{t,\ell,k}^2 \mathbf{C}_{t,\ell}), \quad (2)$$

avec $\bar{\mathbf{f}}_{t,\ell}$ et $\mathbf{C}_{t,\ell}$ l'espérance et la matrice de covariance typique du terme de nuisance $\mathbf{f}_{t,\ell,k}$, et où $\sigma_{t,\ell,k}^2$ sont des facteurs de pondérations temporo-spectraux rendant compte localement de la qualité inégale des images [4]. Il est montré dans [7] et [8] qu'à partir de l'estimation de la matrice de précision $\mathbf{W}_{t,\ell,k} = \sigma_{t,\ell,k}^{-2} \mathbf{C}_{t,\ell}^{-1}$ (évaluée localement par PACO à l'échelle de petits patches) et des termes :

$$\begin{cases} a_{t,\ell}(\theta) = \sum_k \mathbf{h}_{t,\ell}(\theta)^\top \mathbf{W}_{t,\ell,k} \mathbf{h}_{t,\ell}(\theta), \\ b_{t,\ell}(\theta) = \sum_k \mathbf{h}_{t,\ell}(\theta)^\top \mathbf{W}_{t,\ell,k} (\mathbf{r}_{t,\ell,k} - \bar{\mathbf{f}}_{t,\ell}), \end{cases} \quad (3)$$

l'estimateur du maximum de vraisemblance de $\alpha_{t,\ell}$ s'exprime :

$$\hat{\alpha}_{t,\ell}^+(\boldsymbol{\mu}) = \arg \max_{\alpha_{t,\ell} \geq 0} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{[b_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))]_+}{a_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))}, \quad (4)$$

où $[x]_+$ dénote la partie positive de x et $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu})$ la log-vraisemblance des données connaissant les flux $\boldsymbol{\alpha}$ et les paramètres orbitaux $\boldsymbol{\mu}$ de la source. En effet, le flux étant nécessairement positif (ou nul si la source est absente), il est avantageux de lui imposer une contrainte de positivité [6]. On peut montrer que l'estimateur, au sens du maximum de vraisemblance, des éléments orbitaux $\hat{\boldsymbol{\mu}}$ permettant de combiner optimalement les signaux d'un point source potentiel le long de son orbite dans différentes données s'écrit [3] :

$$\hat{\boldsymbol{\mu}} = \arg \max_{\boldsymbol{\mu}} \left\{ \mathcal{L}(\boldsymbol{\mu}) = \sum_{t,\ell} \frac{([b_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))]_+)^2}{a_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))} \right\}. \quad (5)$$

2.2 Formalisme du filtrage adapté

Dans le cas général où des données $\mathbf{d} \in \mathbb{R}^n$ sont mesurées dans des conditions reproductibles paramétrées par $\boldsymbol{\phi} \in \mathbb{R}^p$ (e.g. conditions expérimentales, paramètres de l'objet, etc.), il est possible d'écrire que :

$$\mathbf{d} = \mathbf{m} + \mathbf{z}, \quad (6)$$

avec $\mathbf{m} = \mathbb{E}\{\mathbf{d} | \boldsymbol{\phi}\}$ l'espérance des mesures dans les conditions données $\boldsymbol{\phi}$ et $\mathbf{z} \in \mathbb{R}^n$ un terme stochastique.

Considérons une quantité mesurable $\eta \in \mathbb{R}$ dépendant linéairement des données tel que :

$$\eta = \mathbf{q}^\top \mathbf{d}, \quad (7)$$

avec $\mathbf{q} \in \mathbb{R}^n$ le vecteur des coefficients de la combinaison linéaire. On appelle *filtre adapté* [1] le filtre linéaire $\hat{\mathbf{q}}$ maximisant le RSB de cette combinaison linéaire. Il s'écrit :

$$\hat{\mathbf{q}} = \xi \mathbf{S}^{-1} \mathbf{m}, \quad (8)$$

avec $\mathbf{S} = \mathbb{E}\{\mathbf{z} \mathbf{z}^\top | \boldsymbol{\phi}\} = \text{Cov}\{\mathbf{d} | \boldsymbol{\phi}\}$ la covariance du terme stochastique \mathbf{z} dans des conditions données $\boldsymbol{\phi}$ et $\xi \in \mathbb{R}_*$ un facteur arbitraire. Dans ce cas, le meilleur RSB possible de η dans les conditions $\boldsymbol{\phi}$ s'écrit [1, 11] :

$$S/\mathcal{N}(\eta | \boldsymbol{\phi}) = \max_{\mathbf{q}} \frac{\mathbb{E}\{\eta | \boldsymbol{\phi}\}}{\sqrt{\text{Var}\{\eta | \boldsymbol{\phi}\}}} = \sqrt{\mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m}}. \quad (9)$$

2.3 Application à la détection multiépoque

Nous appliquons le formalisme du filtre adapté décrit ci-dessus au modèle direct des observations ASDI en considérant les données résiduelles :

$$\mathbf{d}_{t,\ell,k} = \mathbf{r}_{t,\ell,k} - \bar{\mathbf{f}}_{t,\ell}, \quad (10)$$

où $\mathbf{d}_{t,\ell,k}$ est supposé avoir une distribution gaussienne d'espérance $\alpha_{t,\ell} \mathbf{h}_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))$ et de variance $\sigma_{t,\ell,k}^2 \mathbf{C}_{t,\ell}$.

En supposant les données résiduelles provenant de différentes époques, trames et canaux spectraux mutuellement indépendantes, la quantité maximisant le RSB vaut [1, 11] :

$$\begin{aligned} \eta &= \xi \sum_{t,\ell} \alpha_{t,\ell} \sum_k \mathbf{h}_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))^\top \mathbf{W}_{t,\ell,k} \mathbf{d}_{t,\ell,k} \\ &= \xi \sum_{t,\ell} \alpha_{t,\ell} b_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu})). \end{aligned} \quad (11)$$

Compte tenu des données et des indépendances supposées, le RSB maximal correspondant s'obtient par l'Éq. (9) :

$$\begin{aligned} S/\mathcal{N}(\boldsymbol{\mu}) &= \sqrt{\sum_{t,\ell,k} \alpha_{t,\ell}^2 \mathbf{h}_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))^\top \mathbf{W}_{t,\ell,k} \mathbf{h}_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))} \\ &= \sqrt{\sum_{t,\ell} \alpha_{t,\ell}^2 a_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))}. \end{aligned} \quad (12)$$

Si l'on considère l'estimateur du maximum de vraisemblance de $\alpha_{t,\ell}$ de l'Éq. (4), cette valeur théorique peut être approximée par l'estimateur suivant :

$$S/\mathcal{N}(\boldsymbol{\mu}) \simeq \sqrt{\sum_{t,\ell} \frac{([b_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))]_+)^2}{a_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))}}. \quad (13)$$

En présence d'une source, ce RSB multiépoque combine optimalement les informations contenues dans les données le long de la trajectoire apparente du compagnon.

Nous retrouvons que ce RSB multiépoque optimal est exactement la racine carrée du critère dérivé par approche du maximum de vraisemblance en Éq. (5) [3]. Maximiser l'un ou l'autre par rapport à $\boldsymbol{\mu}$ donne ainsi exactement la même solution. Le critère de combinaison est donc optimal à la fois dans le sens du maximum de vraisemblance et en terme de RSB. Il permet la détection de sources encore indétectables à chaque époque individuelle et fournit simultanément une estimation de leurs éléments orbitaux. Ce critère est non convexe et présente de nombreux extrema locaux, le maximiser nécessite l'utilisation de méthodes d'optimisation globale. L'algorithme PACOME (voir Sec. 3) a été développé dans ce but.

2.4 Garanties statistiques

Dans l'expression de l'estimateur de l'intensité de la source $\hat{\alpha}_{t,\ell}^+$, le terme $a_{t,\ell}$ est déterministe et seul le terme $b_{t,\ell}$ fluctue. Étant donné notre modèle statistique et l'indépendance mutuelle des trames, la variance de $b_{t,\ell}$ vaut :

$$\text{Var}\{b_{t,\ell}(\boldsymbol{\theta})\} = a_{t,\ell}(\boldsymbol{\theta}). \quad (14)$$

De façon remarquable, $b_{t,\ell}(\boldsymbol{\theta})$ et $a_{t,\ell}(\boldsymbol{\theta})$ constituent une statistique suffisante pour détecter une source potentielle. En approximant la variance de $\hat{\alpha}_{t,\ell}^+$ donnée dans l'Éq. (4) par celle de l'estimateur non contraint, il vient :

$$\text{Var}\{\hat{\alpha}_{t,\ell}^+(\boldsymbol{\mu})\} \approx \frac{\text{Var}\{b_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))\}}{a_{t,\ell}^2(\boldsymbol{\theta}_t(\boldsymbol{\mu}))} = \frac{1}{a_{t,\ell}(\boldsymbol{\theta}_t(\boldsymbol{\mu}))}. \quad (15)$$

Étant donné l'estimateur du flux et de sa variance [8], [5], [4], le RSB temporo-spectral est calculé indépendamment pour chaque époque aux positions $\boldsymbol{\theta}$ comme :

$$S/\mathcal{N}_{t,\ell}(\boldsymbol{\theta}) = \frac{\mathbb{E}\{\hat{\alpha}_{t,\ell}(\boldsymbol{\theta})\}}{\sqrt{\text{Var}\{\hat{\alpha}_{t,\ell}^+(\boldsymbol{\theta})\}}} \approx \frac{b_{t,\ell}(\boldsymbol{\theta})}{\sqrt{a_{t,\ell}(\boldsymbol{\theta})}}, \quad (16)$$

et suit une distribution normale en l'absence de source. Grâce à cette propriété, ce RSB monoépoque est directement interprétable en termes de probabilité de détection et de probabilité de fausse alarme.

En remarquant que le RSB multiépoque de l'Éq. (13) est la racine de la somme quadratique des RSB monoépoques positifs, il peut lui aussi être interprété en termes de probabilité

de détection multiépoque. En effet, la distribution statistique du critère correspond à la somme des carrés de $L = M$ distributions gaussiennes rectifiées et indépendantes, où L représentent respectivement le nombre total d'époques et de canaux spectraux. À notre connaissance, cette distribution n'a pas d'expression analytique, nous recourons donc à une approche de type Monte-Carlo pour construire la distribution du critère et nous en déduisons empiriquement, à niveau de confiance $\alpha \in [0; 1]$ donné, la valeur de la fonction quantile $Q_M(1 - \alpha)$ [2] associée. Cette dernière représente la valeur du seuil de détection pour le RSB multiépoque de l'Eq. (1) en dessous duquel se situe une proportion α des données en l'absence de compagnon. Les fonctions de densité de probabilité du RSB multiépoque calculées par cette procédure sont présentées dans la Fig. 1. Quelques valeurs de la fonction quantile évaluée à différents niveaux de confiance et pour plusieurs degrés de liberté sont données en Fig. 2.

FIGURE 1 : Densités de probabilité théoriques du RSB multiépoque en l'absence de source.

FIGURE 2 : Seuil de détection du RSB multiépoque pour différents degrés de liberté et niveaux de confiance.

3 L'algorithme PACOME

L'algorithme PACOME [3] échantillonne le RSB multiépoque sur une grille multidimensionnelle représentant l'espace de recherche des éléments orbitaux. Pour cela, les positions projetées 2-D subpixeliques sont calculées pour chaque nœud de l'espace des paramètres et pour toutes les époques. Les termes a_i et b_i sont interpolés à ces positions afin d'évaluer le RSB multiépoque. Les quelques meilleures solutions orbitales sont ensuite raffinées par optimisation locale via la méthode MLMB [9]. Le RSB multiépoque maximal obtenu est finalement comparé à la limite de détection multiépoque (à un niveau de confiance donné) estimée numériquement via la distribution statistique empirique du critère pour décider d'une éventuelle détection.

4 Application sur données simulées

Nous évaluons les performances de PACOME sur un jeu de données semi-synthétiques construit par la simulation réaliste de sources synthétiques ajoutées aux données réelles du système exoplanétaire HR 8799 observé le 2016-11-18 avec l'instrument SPHERE-IRDIS [10] opérant au Very Large Telescope. Nous avons simulé deux fausses époques totalisant 5:12 heures d'observation espacées de 50 ans. Quatre sources (désignées par (1), (2), (3) et (4)) y ont été ajoutées selon le modèle direct de l'Eq. (1) sur des orbites connues aux époques correspondantes en utilisant la FEP hors axe mesurée lors de l'observation. Le flux injecté des sources est choisi tel qu'elles restent indétectables (aux probabilités de fausses alarmes usuelles dans le domaine) par le traitement non-conjoint de chaque époque d'observation. Seule la source (1) a un RSB monoépoque légèrement supérieure à cette limite empirique sur 18.

Nous cherchons à détecter en aveugle les quatre sources avec PACOME. Au total, $N_{orb} = 43 \cdot 10^{10}$ orbites différentes ont été explorées afin de couvrir au mieux l'espace des paramètres des éléments orbitaux. Le niveau de confiance de détection multiépoque a été fixé à $\alpha = 10^{-7}$ ce qui correspond, à 18 degrés de liberté, au seuil $Q_{18}(1 - \alpha) \approx 7.4$.

Les estimées de trouvées par PACOME pour chaque source sont données dans la Tab. 1 et comparés à la vérité terrain, au sens de la racine de l'erreur quadratique moyenne (REQM) entre les positions projetées 2-D sur le capteur.

Les solutions orbitales obtenues sont très proches de la vérité terrain (REQM: 0.75 pixels pour chacune des quatre sources, voir Tab. 1) à l'exception de l'élément orbital de la source (4), qui est significativement différent. Ceci révèle une dégénérescence connue du paramètre. L'orbite projetée estimée est en revanche bien identique à celle simulée puisque leurs positions projetées sur le détecteur aux époques données sont quasi-égales (REQM: 0.2 pixels).

Leurs valeurs du RSB multiépoque sont toutes significativement plus élevées que la limite de détection empirique multiépoque à $\alpha = 10^{-7}$ tandis que la grande majorité des points sources ne sont pas détectées individuellement ($\alpha < 5$). Plus précisément, les RSB monoépoques moyens des solutions optimales sont de 4:0, 2:2, 2:4, 2:3 tandis que leurs RSB multiépoques sont de 18:5, 9:9, 10:5, 10:2 (de la source (1) à (4) respectivement). Cela signifie que passer de mono à multiépoque entraîne un gain moyen en RSB d'environ 4.5 facteur quasi égal à la racine carrée du nombre de degrés de liberté ($\sqrt{18} \approx 4.2$) ce qui illustre bien l'optimalité de PACOME. La Fig. 3 met en valeur ce gain en comparant différentes cartes de RSB monoépoque et multiépoque centrées autour des solutions orbitales estimées par PACOME. Tous les signaux dans le champ qui ne se trouvent pas aux positions de la source se combinent faiblement (des couleurs bleues aux couleurs blanches, en dessous du niveau de confiance fixé). En n, la Fig. 4 représente les solutions orbitales trouvées par PACOME (en rouge) ainsi que les 10^3 autres meilleures orbites optimisées (en orange).

5 Conclusions

Nous avons montré que le formalisme du ltrage adapté appliqué à PACOME permet de dériver un critère de détection

