Amélioration et validation du code de caractérisation d'amas de galaxies PROF-CL pour la mission Euclid

Pour le diplôme de Master 1 Astronomie et Astrophysique, "Sciences de l'Univers et Technologies Spatiales". Observatoire de Paris

Wilfried MERCIER
Sous la direction de Gary MAMON

18 juin 2018

Institut d'Astrophysique de Paris
Observatoire de Paris

Je tenais à remercier l'Observatoire de Paris et l'IAP pour m'avoir permis d'effectuer ce stage, notamment les équipes techniques qui ont été très sympathiques tout au long du stage. Un grand merci à tous les stagiaires avec qui j'ai pu échanger et pour toutes ces discussion intéressantes que ce soit en astropysique ou non. Je tenais aussi à remercier l'ensemble des intervenants lors des séminaires du vendredi matin pour tous ces sujets passionants qui m'ont donné de nombreuses idées.

Enfin je souhaitais remercier chaleureusement mon encadrant, Gary Mamon, pour toutes les discussions qu'on a pu avoir au niveau du stage mais aussi sur d'autres sujets d'astrophysiques variés comme vos sujets de recherche actuels. Merci d'avoir pris le temps de répondre à toutes les questions que je me posais et dès fois même plus que je n'en demandais. Merci aussi pour les nombreuses explications et précisions sur le fonctionnement du système universitaire et pour les conseils de lecture.

Résumé

Une nouvelle version du code de caractérisation d'amas de galaxies PROF-CL est codée prenant en compte les amas académiques et ceux d'AMICO. Le code fonctionne en mode circulaire et/ou elliptique avec et sans fond. De nouvelles équations sont calculées pour la probabilité de position des galaxies dans le cas elliptique. Le code principal a été modifié pour pouvoir améliorer la convergence des algorithmes de minimisation non-contraints en ajoutant dans le *likelihood* deux fonctions pénalité différentes ; une étant plus générale que l'autre. Une méthode de détermination du rayon caractéristique par séparation médiane est implémentée et comparée aux méthodes par maximum de vraisemblance. Le code est testé sur les mocks académiques et sur ~ 550 amas AMICO. On trouve des performances correctes en terme de biais pour $N \gtrsim 50$ galaxies à la fois en elliptique et en circulaire pour le *log scale radius*, l'ellipticité et le PA. La richesse est légèrement biaisée négativement vers -0.2dex ce qui est probablement dû à une surévaluation du fond. Globalement DE trouve de meilleurs résultats que TNC bien que la différence soit moins notable sur les amas AMICO. Des améliorations futures du code tel que le *deblending* sont discutées en conclusion.

Table des matières

Notations

Nom	Symbole(s)	Description
Profil NFW	$\rho_{\rm NFW}$	Profil de densité universel de matière noire Navarro, Frenk, White
Densité surfacique	$\Sigma(R)$	Densité surfacique projetée sur le ciel
Rayon de pente -2	$r_{\rm s}$, r_{-2}	Rayon pour lequel la pente de ρ_{NFW} vaut -2
	r_{200}	Rayon pour lequel la densité de la sphère vaut 200 fois la densité moyenne de matière dans l'Univers
Rayon viriel	$r_{\rm vir}$	Rayon pour lequel la matière dans la sphère est en équilibre
Rayon projeté	R	Distance au centre d'un amas projetée sur le ciel
Rayon circularisé	$\mathscr{R}$	Distance projetée au centre d'un amas après circularisation
Ellipticité	e	Ellipticité de l'amas égale à $1 - b/a$ avec b le demi-petit axe et a le demi-grand axe
Angle de position	PA	Angle entre le demi-grand axe et le Nord allant du Nord vers l'Est
	$\boldsymbol{\theta}$	Vecteur des paramètres du modèle
Nombre projeté	$N_{\rm p}(R)$	Nombre de galaxies à une distance projetée R du centre de l'amas
Richesse	$N_{\rm eff}$	Somme des probabilités d'appartenance des galaxies à l'amas

Avant propos

Dans la suite les termes halos et amas seront utilisés pour décrire la même chose, à savoir des amas de galaxies. Pour éviter toute confusion le terme halo sera exclusivement utilisé pour désigner les amas de galaxies issus de [**?**] sur la simulation [Springel et al., 2005].

A l'inverse le terme amas référera uniquement soit aux amas "académiques" (mocks académiques) simulés lors du stage pour tester l'efficacité du code, soit aux amas trouvés par l'algorithme de détection d'amas AMICO.

Les expressions latines et la terminologie en anglais seront écrites en italique dans ce qui suit.

L'ensemble du travail et du traitement de données a été effectué en python avec les bibliothèques *numpy*[1] et *astropy*[2]

1. `https://docs.scipy.org/doc/`
2. `http://docs.astropy.org/en/stable/`

1 Euclid, AMICO et PROF-CL

1.1 Projet Euclid

La mission Euclid cherche à mieux comprendre la nature de l'énergie sombre et l'origine de l'expansion accélérée de l'Univers en étudiant notamment le cisaillement cosmique via lentilles gravitationnelles faibles, le *clustering* des galaxies et en comptant les amas.

Pour ce faire, un large catalogue de 15000 deg^2 (*wide survey*) d'objets extra-galactiques à la fois dans les bandes optiques jusqu'à 24.5mag et dans l'infrarouge proche (bandes Y, J, H) jusqu'à 24mag sera disponible, complémenté par deux catalogues plus restreints (*deep surveys*) cumulant 40deg^2 et pouvant aller jusqu'à 26mag environs [Adam et al., 2018]. Tous les objets observés posséderont un redshift photométrique, voir un redshift spectroscopique pour certains d'entre eux. Euclid sera en mesure d'observer des galaxies de masse supérieure à $10^{14}\text{M}_\odot$ et jusqu'à des redshifts de l'ordre de 2.

L'étude conjointe des lentilles gravitationnelles faibles (*Weak Lensing* - WL) et du *clustering* (*Galaxy Clustering* - GC) devrait permettre en particulier d'étudier certaines questions de cosmologie contemporaine, parmi lesquelles se trouvent[Laureijs et al., 2011] :

- la dépendance en redshift du paramètre $w(a)$ où $a = 1+z$ et de sa dérivée première $w_a(a) = dw/da$ de l'équation d'état de l'énergie sombre $p = w(a)\rho$
- le facteur de croissance et sa dérivée (*growth rate*) $f(z) = [\Omega_m(z)]^\gamma$ qui caractérisent l'efficacité à former des structures dans l'Univers au cours du temps, en particulier la mesure de l'exposant γ

Le comptage des amas sera un $3^{\text{ème}}$ moyen (secondaire) de contraindre les paramètres, c'est pourquoi c'est une branche active de la préparation de la mission. Un total de 14 *Processing Functions* sont en cours de développement. Mon stage se porte sur la *Processing Function* PROF-CL qui cherche à caractériser la structure des amas détectés par le détecteur d'amas DET-CL.

1.2 Algorithmes de détection d'amas : AMICO et PZWav

La première étape consiste à détecter les amas de galaxies dans le catalogue d'Euclid. Dans ce contexte, des algorithmes de détection d'amas (*cluster finders*) ont été développés et testés dans un ensemble de 4 challenges (CFC).

Chaque algorithme a été testé sur un catalogue mock[Ascaso et al., 2015] d'une taille de 300deg^2 fournissant des données similaires à celles attendues pour Euclid. En particulier le catalogue était limité à des objets plus brillants que 24mag dans les intervalles considérés de masse et de redshift. Les positions des galaxies étaient fournies avec un redshift photométrique et une probabilité de distribution associée.

Parmi les 8 codes de départ, deux ont été retenus pour leur efficacité et leur complémentarité dans les intervalles de masse et de redshift où ils excellent : PZWav et Adaptative Matched Identifier of Clustered Objects (AMICO).

Seules les données d'AMICO ont pu être utilisées et analysées pendant le stage. Ainsi on ne donnera qu'une description très brève du fonctionnement de PZWav (voir [Adam et al., 2018] et [Eisenhardt et al., 2008] pour plus d'informations sur le fonctionnement de l'algorithme).

1.2.1 PZWav

Cet algorithme découpe l'intervalle de redshift en plusieurs tranches, puis construit à l'aide du champ de densité de probabilité des cartes de densités à différents redshift en lissant celui-ci par ondelettes (*wavelets*). En répétant l'opération avec des densités de probabilité aléatoires, une valeur maximale sur le bruit du fond peut être déduite ainsi que sa variation avec le redshift. Les amas sont alors détectés comme des pics dans les cartes de densité et une valeur du rapport signal-bruit (S/N) est déduite.

1.2.2 AMICO

AMICO est un algorithme de détection d'amas basé sur une méthode de Filtrage Optimale (OF) originellement utilisé pour la détection d'amas sur des données de lentille gravitationnelle faible[Bellagamba et al., 2018]. Les données sont décrites comme étant la somme d'un modèle et d'un fond non-uniforme. La densité surfacique du fond est calculée localement autour de chaque grande structure mais suffisamment loin pour ne pas prendre en compte le signal provenant du modèle.

L'amplitude du signal brut (sans fond) est alors déterminée en appliquant un filtre sur la distribution des galaxies ; les pics dans la carte de densité 3D obtenue correspondent aux amas. Un processus itératif va alors :

- éliminer tous les pixels de l'image avec un rapport signal-bruit S/N inférieur à une certaine limite $(S/N)_{\min}$ typiquement de l'ordre de 2
- chercher le pic avec la plus grande amplitude et l'assigner comme étant un amas
- assigner une probabilité $p(i \in j)$ d'appartenance de la galaxie i à l'amas j pour toutes les galaxies au voisinage de l'amas ainsi qu'une "probabilité de champ" (*field probability*) définie comme $p_{f,i} = 1 - \sum_{j=1}^{N} p(i \in j)$
- retirer de l'image le signal du modèle pour chaque galaxie multiplié par la probabilité de champ
- recommencer depuis l'étape 1 jusqu'à ce que plus aucune galaxie n'ait $S/N > (S/N)_{\min}$

Contrairement à PZWav cet algorithme possède l'avantage de fournir une probabilité d'appartenance à l'amas (*membership*) que l'on pourra incorporer dans le calcul de la vraisemblance (voir section 4). En principe, AMICO permet de mieux détecter les petites structures.

1.3 PROF-CL : code de détermination des propriétés d'amas par maximum de vraisemblance

PROFCL est un algorithme qui cherche à déterminer les meilleurs paramètres des amas détectés par AMICO et PZWav par maximum de vraisemblance. Le code ne reçoit en entrée que les positions des galaxies appartenant aux amas trouvés, ainsi que les probabilités d'appartenance à l'amas pour AMICO.

À chaque galaxie de l'amas détecté va être associée une probabilité de trouver la galaxie à cette position sachant les paramètres considérés. On peut alors définir un *likelihood* comme étant le produit des probabilités de toutes les galaxies de l'amas. Le code va alors se charger de trouver les paramètres qui minimisent l'opposé du logarithme du *likelihood* via plusieurs procédures de minimisation (descente de gradient, algorithme stochastique, simplex, etc...).
Les paramètres testés par PROF-CL sont les suivants :

- Rayon caractéristique et en particulier son logarithme (*log scale radius*) noté $\log_{10}(r_s/\text{arcmin})$
- Angle de position de l'amas en degrés noté PA
- Ellipticité de l'amas $e = 1 - b/a$
- Densité surfacique du fond Σ_{bg} supposée constante en arcmin^{-2}
- Le logarithme du rayon de troncation $\log_{10}(R_{\text{trunc}}/\text{arcmin})$ si un modèle NFW tronqué est testé

1.4 Fonctionnalités déjà incluses et améliorations à apporter

Au moment de commencer le stage, PROF-CL était dans sa version 1.10. Les fonctionnalités principales avaient déjà été programmées mais le code n'avait été que peu testé jusqu'à présent. En principe, le code était capable de fonctionner dans les cas suivants :

- amas circulaire : considère $e, PA, N_{\text{bg}} = 0$ et retourne $\log_{10}(r_{-2})$
- amas elliptique : retourne $\{e, PA, \log_{10}(r_{-2})\}$
- avec fond : retourne en plus $\log_{10}(\Sigma_{\text{bg}})$

Le cas elliptique pouvait être traité par approximation analytique de l'intégrale de densité de surface (cf. section 3.2) ou par Monte-Carlo. Le code ne fonctionnait que sur les amas académiques d'Emmanuel Artis (stagiaire M2 de G. Mamon l'an dernier).
Dans la pratique, certaines parties du code n'étaient pas opérationnelles ou ne retournaient pas les résultats attendus. Il a donc fallu retravailler dessus avant de pouvoir analyser les performances du code. Ci-dessous une liste non-exhaustive des parties du code sur lesquelles il a fallu retravailler/implémenter :

- Ré-écriture de la séparation médiane
- Ré-écriture et implémentation des équations de la probabilité pour des amas elliptiques (potentiellement décentrés)
- Gestion des sorties de limites pour des algorithmes non-contraintes (cf. annexe B)
- Analyse des performances sur les mocks académiques
- Gestion/association via plusieurs fichiers et analyse des données issues d'AMICO

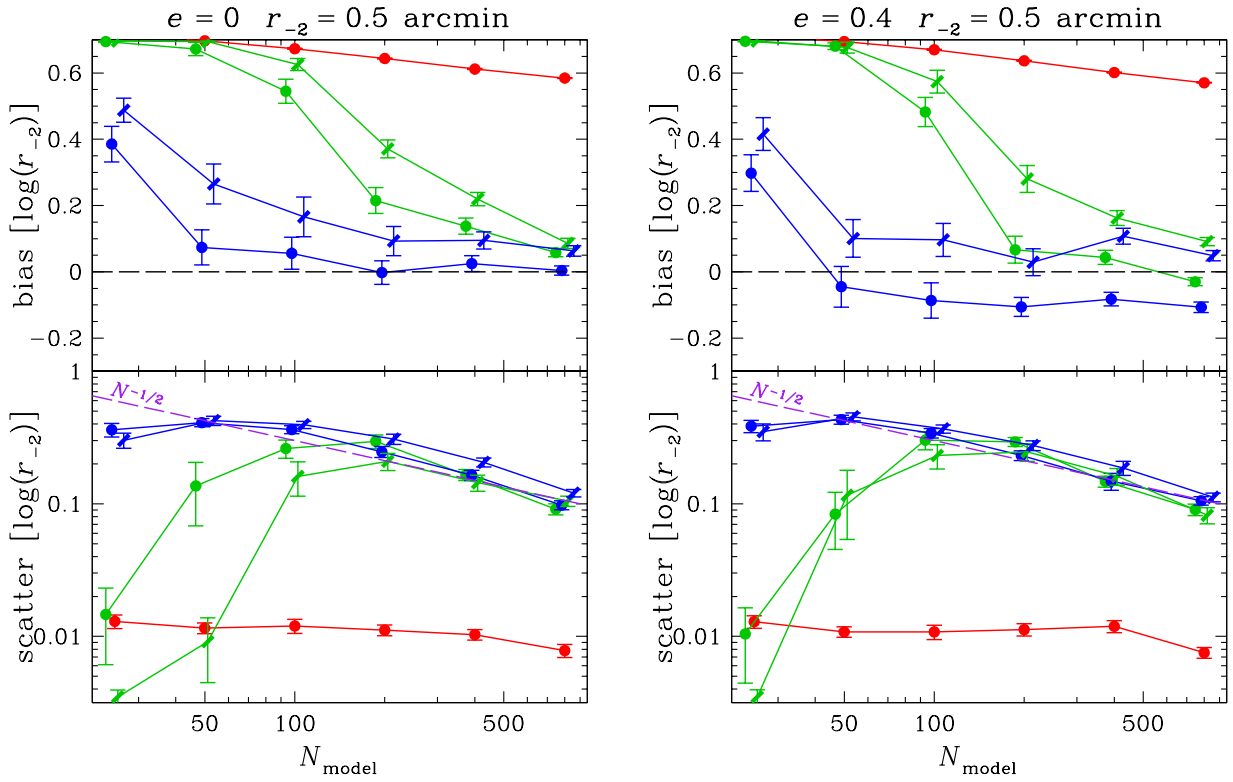


FIGURE 1 – Biais et dispersion en *log scale radius* sur les mocks académiques avec à gauche un amas circulaire et à droite un amas d'ellipticité 0.4. On trouve en vert, bleu et rouge respectivement les méthodes de gradient conjugué tronqué (TNC), d'évolution différentielle (DE) et de séparation médiane. Les cercles et rectangles obliques correspondent aux résultats du *likelihood* avec ajustements circulaire et elliptique, respectivement. DE se comporte globalement le mieux en terme de biais. La séparation médiane a un biais élevé mais comme attendu une dispersion très faible.

2 Étude de la Séparation médiane

2.1 Modélisation et calcul

Avant de commencer à étudier l'efficacité du code par maximum de vraisemblance, une alternative prometteuse semblait résider dans la séparation médiane entre les galaxies. En effet, si l'on considère un amas sans fond constitué de N galaxies, alors à N plus élevé on possède une résolution plus grande pour calculer la médiane des séparations inter-galaxies, ce qui en principe devrait donner intuitivement une loi en puissance $\Delta \propto r_{-2}^{\alpha}$, ou sous forme logarithmique $\log_{10}\Delta = \alpha \log_{10}(r_{-2}) + \beta$.

En pratique on boucle dans le code sur i, j avec $j < i$ et on calcule les $N(N-1)/2$ séparations $\Delta l_{kl} = |l_k - l_l|$. On obtient alors la partie triangulaire supérieure d'une matrice symétrique sur laquelle on peut calculer la séparation médiane.

2.2 Analyse de la séparation médiane

La séparation médiane est comparée dans la Fig.1 avec DE et TNC. Celle-ci est grandement biaisée vers des valeurs positives. Comme attendu on obtient une loi linéaire en log pour $N \gtrsim 50$ avec $\alpha \approx 0.1$ et $\beta \approx 0.67$. Étonnamment, ces valeurs semblent être indépendantes de l'ellipticité de l'amas et changent peu pour des mocks amas+fond avec des valeurs de fond raisonnables.

Le fait que la séparation médiane ne soit basée sur aucun modèle sous-jacent la rend peu sensible aux variations statistiques qui peuvent apparaître lors de la génération aléatoire des mocks. Conséquemment, sa dispersion est très faible. De fait, la séparation médiane semble être une alternative intéressante si l'on arrive à corriger le biais qui semble systématique.

3 Halo de matière noire : Profil NFW et surface de densité

On résume rapidement dans cette section les propriétés 3D et 2D des amas étudiés qui serviront dans le calcul du *likelihood* dans la section 4.

3.1 Profils NFW et NFW tronqués

Vers la fin des années 1990 Navarro, Frenk et White montrèrent à l'aide de simulations à N-corps que les halos de matière noire formés dans des cosmogonies de type CDM possèdent un profil de densité 3D NFW universel (peu dépendant de la masse). Un tel profil est abondamment utilisé dans la littérature pour la modélisation de profil radial d'amas (modèle utilisé dans [Bellagamba et al., 2018]). Le profil NFW est généralement écrit sous la forme[Navarro et al., 1996]

$$\frac{\rho(r)}{\rho_{crit}} = \frac{\delta_{char}}{(r/r_s)(1+r/r_s)^2} \tag{3.1}$$

où r_s représente le rayon de pente -2 et δ_{char} est une surdensité caractéristique que l'on relie au paramètre de concentration $c = r_s/r_v$ via la formule[Mo et al., 2010]

$$\delta_{char} = \frac{200}{3}\frac{c^3}{\ln(1+c) - c/(1+c)} \tag{3.2}$$

3.2 Densité surfacique projetée

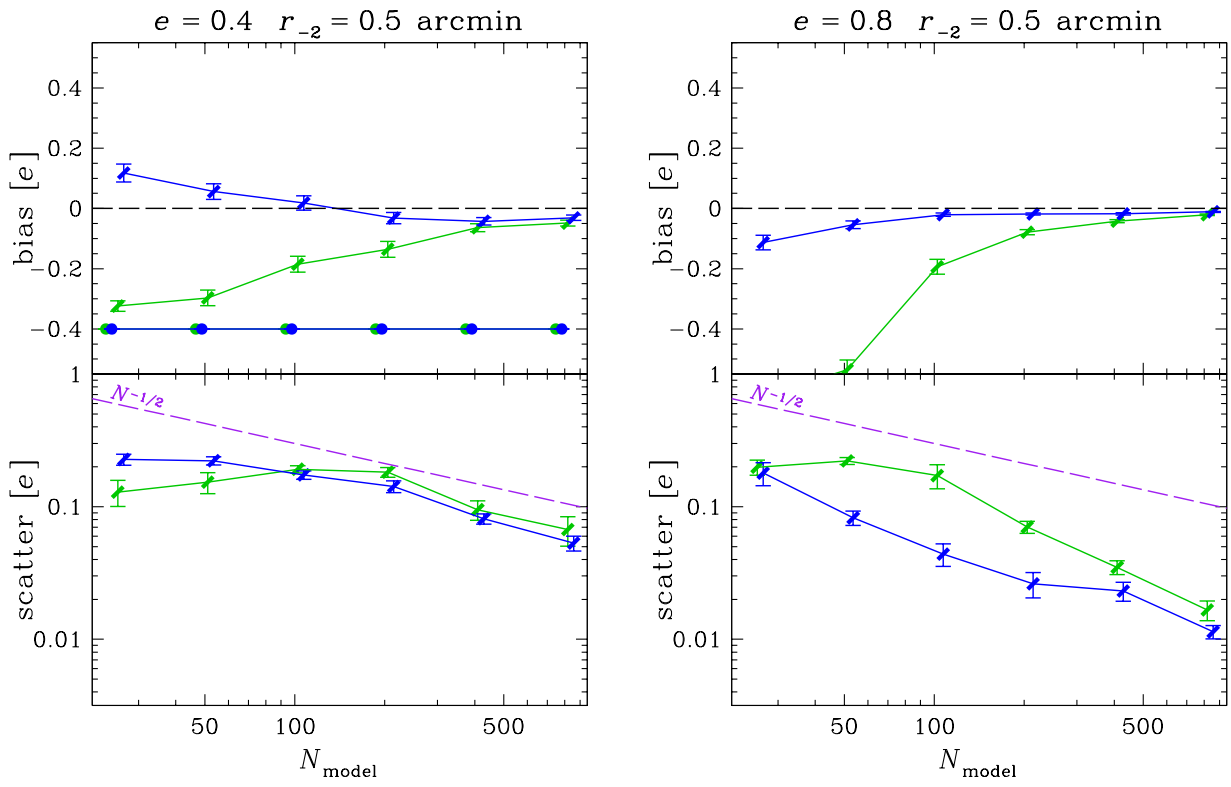
Les données traitées étant les positions projetées des galaxies sur le ciel, il nous faut non pas travailler avec la densité 3D mais avec la densité surfacique projetée. Celle-ci est obtenue en intégrant le profil de densité le long de la ligne de visée

$$\Sigma(R) = \int_{\mathbb{R}} \rho(r) dz \tag{3.3}$$

où r est le rayon 3D (ci-après rayon) et z est la composante selon la ligne de visée. En notant R le rayon dans le plan du ciel (ci-après rayon projeté), on peut réécrire cette dernière équation comme

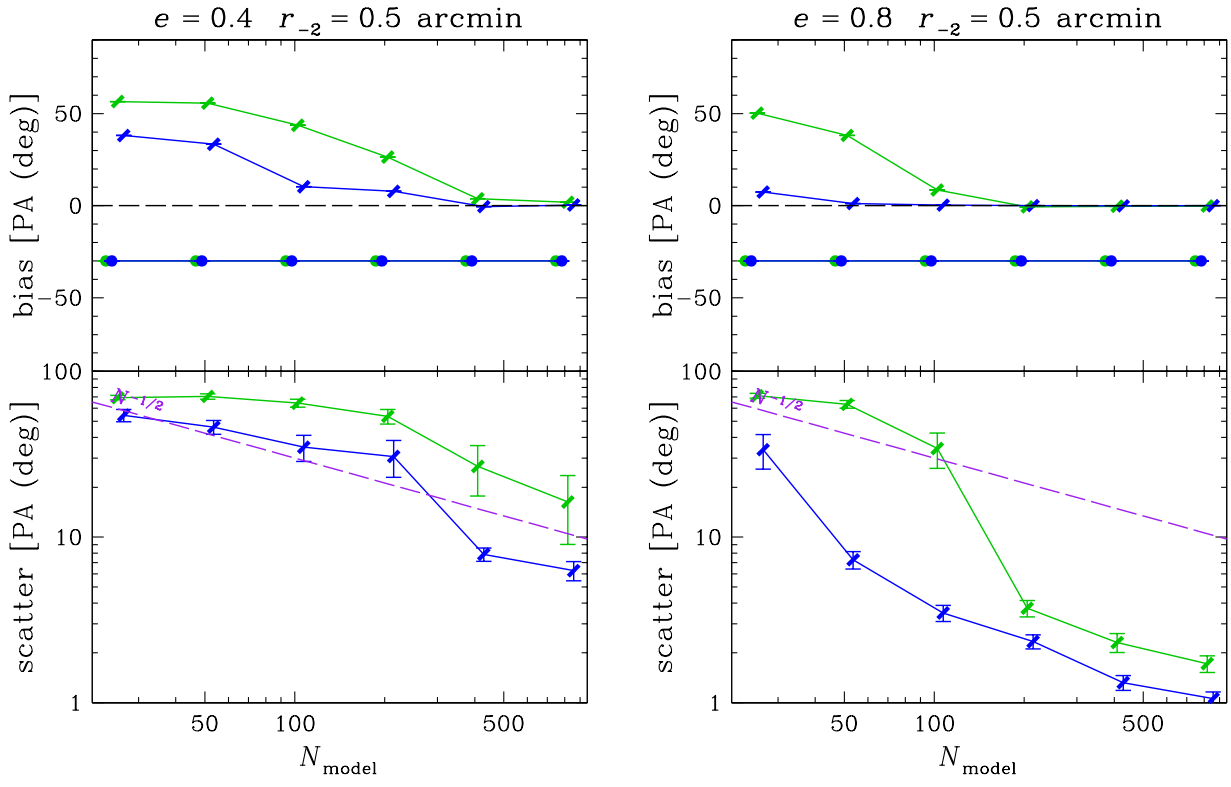
$$\Sigma(R) = 2\int_R^\infty \frac{r\rho(r)dr}{(r^2 - R^2)^{1/2}} \tag{3.4}$$

La solution analytique de cette équation donnée en Annexe A pour un profil NFW est utilisée par PROF-CL pour le calcul du *likelihood*. Les halos auxquels vont être comparées les données d'AMICO ne s'étendent pas jusqu'à l'infini mais sont en réalité tronqués jusqu'à un rayon $\sim r_{\rm vir}$. Cela signifie que Eq.3.4 n'est plus correcte et qu'il faut effectuer le changement $\infty \to R_{\rm trunc}$ au niveau de la borne supérieure. Une solution analytique donnée en Annexe A est aussi utilisée par PROF-CL, mais uniquement dans le but d'étudier les halos (cf. section 7).



(a)

(b)



(c)

(d)

FIGURE 2 – Biais et dispersion en ellipticité et PA pour deux types de mocks académiques elliptiques. Les couleurs et symboles sont similaires à celles de Fig.1. Comme pour le *log scale radius* DE obtient de meilleurs résultats que TNC en termes de biais et de dispersion. On remarque que le PA est mieux trouvé à ellipticité élevée.

4 Calcul du log-likelihood

PROF-CL repose entièrement sur l'idée de trouver les meilleurs paramètres pour chaque halo par maximum de vraisemblance. Dans cette section on présente les équations utilisées par PROF-CL et notamment celles développées lors du stage pour étudier des halos non-sphériques.

4.1 Maximum de vraisemblance pour des amas de galaxies

Jusque dans les années 1980 l'unique méthode employée pour déterminer la densité surfacique projetée des amas était du *binning-fitting*. Chaque image était découpée en anneaux de taille quelconque et le nombre de galaxies dans chaque anneau était compté. Sarazzin montra que ce type de méthode faisait apparaître des artefacts dans le rayon caractéristique des amas dû au choix de la taille des anneaux [Sarazin, 1980]. Il proposa une méthode sans *binning* basée sur le principe de maximum de vraisemblance.

L'idée est d'assigner à chaque galaxie une probabilité de la trouver dans sa position étant donné le modèle considéré et les paramètres testés, puis de construire le *likelihood* comme le produit de ces probabilités. Pour un ensemble de N galaxies de positions $\{\boldsymbol{X_i} = (x_i, y_i)_{1<i<N}\}$, en notant $\boldsymbol{\theta} = \{\theta_i\}$ l'ensemble des paramètres testés, on peut écrire le *likelihood* comme

$$\mathscr{L} = \prod_{i=1}^{N} p(\boldsymbol{X_i}|\boldsymbol{\theta}) \tag{4.1}$$

Techniquement $p(\boldsymbol{X_i}|\boldsymbol{\theta})$ est une densité de probabilité normalisée à 1 telle que $p((X_i)|\boldsymbol{\theta})d^2\boldsymbol{X}$ soit la probabilité de trouver la galaxie i dans l'intervalle $[\boldsymbol{X_i}, \boldsymbol{X_i} + d\boldsymbol{X}]$.

En pratique, puisqu'il est plus simple de minimiser une fonction que de la maximiser et que les probabilités considérées sont très faibles, on considère plutôt l'opposé du logarithme du *likelihood* (*log-likelihood*, ci-après simplement nommé *likelihood*)

$$-\log \mathscr{L} = -\sum_{i=1}^{N} p(\boldsymbol{X_i} \in \text{cluster}) \log \mathrm{p}(\mathrm{X_i}|\boldsymbol{\theta}) \tag{4.2}$$

où on a pris soin de pondérer chaque terme par la probabilité de trouver la galaxie dans l'amas correspondant. Celle-ci est fournie par AMICO mais pas par PZWav ou par les mocks académiques, c'est pourquoi, si la probabilité d'appartenance des galaxies n'est pas connue on la prendra égale à 1.

4.2 Probabilité de position pour un amas sphérique

La probabilité d'appartenance d'une galaxie à un amas sphérique est relativement simple. Dans ce cas il n'est pas question d'ellipticité ou d'angle de position. Ainsi, pour un modèle NFW, les paramètres se réduisent à $\boldsymbol{\theta} = \{r_{-2}, \Sigma_{\mathrm{bg}}\}$ et la probabilité s'écrit[Mamon et al., 2010]

$$p(R|\boldsymbol{\theta}) = \frac{2\pi R[\Sigma(R) + \Sigma_{\mathrm{bg}}]}{N_{\mathrm{tot}}} \tag{4.3}$$

où R représente le rayon projeté, Σ est la densité surfacique du modèle et N_{tot} est le nombre total de galaxies dans l'amas qui, pour un amas dont l'extension angulaire va de R_{min} à R_{max}, s'écrit

$$\begin{aligned} N_{\mathrm{tot}} &= \int_{R_{\mathrm{min}}}^{R_{\mathrm{max}}} 2\pi R \Sigma_{\mathrm{tot}}(R) dR \\ &= N(r_{-2}) \Delta \tilde{N}_p + \pi \Sigma_{\mathrm{bg}} \Delta R^2 \end{aligned} \tag{4.4}$$

avec $\Sigma_{\mathrm{tot}}(R) = \Sigma(R) + \Sigma_{\mathrm{bg}}$ et où on a défini les quantités suivantes

$$\begin{aligned} N_{\mathrm{p}}(R) &= N(r_{-2}) \tilde{N}_{\mathrm{p}}\left(R/r_{-2}\right) \\ \Delta R^2 &= R^2_{\mathrm{max}} - R^2_{\mathrm{min}} \\ \Delta \tilde{N}_{\mathrm{p}} &= \tilde{N}_{\mathrm{p}}(R_{\mathrm{max}}/r_{-2}) - \tilde{N}_{\mathrm{p}}(R_{\mathrm{min}}/r_{-2}) \end{aligned}$$

avec $N_{\mathrm{p}}(R)$ le nombre de galaxies projetées dans l'amas selon le modèle jusqu'à une distance R et $\tilde{N}_{\mathrm{p}}$ l'équivalent adimensionné.

Dans les faits, le terme $2\pi R$ dans Eq.4.3 est constant et n'intervient pas dans la procédure de minimisation, on peut donc se permettre de le retirer du code (ce ne sera plus vrai dans le cas elliptique).

Le facteur de normalisation $N(r_{-2})$ est quant à lui calculé via Eq.4.4 étant donné que l'on connaît à la fois le nombre total de galaxies appartenant à l'amas, son extension spatiale et la densité surfacique du fond testée.

4.3 Probabilité pour un amas elliptique

Dans le cas elliptique les paramètres sont $\boldsymbol{\theta} = \{r_{-2}, \Sigma_{\rm bg}, e, PA\}$. Eq.4.3 n'est plus correcte car l'ellipticité doit être prise en compte. Puisqu'il n'existe pas de modèle elliptique pour la densité surfacique, l'idée va être de "circulariser" l'amas et d'exprimer la probabilité de trouver la galaxies dans cet amas circularisé.

En pratique, on commence par tourner l'amas d'un angle $-PA$ puis on calcule les nouvelles coordonnées locales elliptiques $(u, v^{'})$ pour chaque galaxie comme décrit dans l'Annexe C.

On peut à présent écrire la probabilité de positionnement de la galaxie de manière générale comme

$$p((u,v)|\boldsymbol{\theta}) = \frac{\Sigma_{\rm tot}(u,v)}{\iint_\Gamma \Sigma_{\rm tot}(u,v){\rm dudv}} \tag{4.5}$$

où Γ est le domaine d'intégration de l'amas circularisé tel que la probabilité soit normalisée à 1. Dans Eq.4.5 la densité surfacique totale peut s'écrire comme la somme du modèle et d'un fond constant

$$\Sigma_{\rm tot}(u,v) = \Sigma_{\rm ell}(u,v) + \Sigma_{\rm bg} \tag{4.6}$$

Contrairement au cas circulaire, la densité surfacique du modèle n'est plus isotrope dans le système de coordonnées (u, v), cependant on peut la relier à une densité surfacique isotrope si on considère les coordonnées $(u, v^{'})$ via la relation

$$\Sigma_{\rm ell}(u,v) = \frac{1}{1-e}\Sigma(\mathscr{R}) \tag{4.7}$$

En insérant Eq.4.6 dans le dénominateur de Eq.4.3 on exprime le nombre total de galaxies

$$\begin{aligned} N_{\rm tot} &= \frac{1}{1-e}\iint_\Gamma \Sigma(\mathscr{R}){\rm dudv} + \pi\Delta{\rm R}^2\Sigma_{\rm bg} \\ &= N(r_{-2})[\tilde{N}_{\rm p}(R_{\rm max}/r_{-2}, e) - \tilde{N}_{\rm p}(R_{\rm min}/r_{-2}, e)] + \pi\Delta R^2\Sigma_{\rm bg} \end{aligned} \tag{4.8}$$

où on a défini les quantités suivantes

$$\begin{aligned} \tilde{N}_{\rm p}(R/r_{-2}, e) &= N_{\rm p}(R,e)/N(r_{-2}) \\ N_{\rm p}(R,e) &= \frac{1}{1-e}\int_{-R}^{R}{\rm du}\int_{-\sqrt{{\rm R}^2-{\rm u}^2}}^{\sqrt{{\rm R}^2-{\rm u}^2}}\Sigma_{\rm ell}(\mathscr{R}){\rm dv} \end{aligned}$$

Comme pour le cas circulaire on obtient le facteur de normalisation $N(r_{-2})$ dans le cas elliptique via Eq.4.8. En combinant les résultats précédents on peut alors exprimer la probabilité de position d'une galaxie dans un amas sachant les paramètres $\boldsymbol{\theta}$ comme

$$p((u,v)|\boldsymbol{\theta}) = \frac{1}{N_{\rm tot}}\left[\frac{N(r_{-2})}{\pi r_{-2}^2(1-e)}\tilde{\Sigma}_{\rm ell}(\mathscr{R}/r_{-2}) + \Sigma_{\rm bg}\right] \tag{4.9}$$

5 Amélioration des algorithmes de minimisation

5.1 Gestion des limites

Certains algorithmes comme le gradient conjugué tronqué (TNC) ou Broyden-Fletcher-Goldfarb-Shanno (BFGS) prennent en compte la gestion des limites sur les paramètres de la fonction à minimiser. Dans ce cas il suffit de fournir pour chaque paramètre un intervalle qui va délimiter l'espace des paramètres à explorer. Au contraire certaines méthodes telles que Nelder-Mead (NM) peuvent en principe parcourir complètement l'espace des paramètres et donc converger vers un minimum local pour des valeurs non-physiques.

La solution qui avait été trouvée au début du stage était de tester les valeurs pour chaque paramètre et de retourner $-\log_{10}(\mathscr{L}) \sim 10^{30}$. En principe cette astuce devrait fonctionner pour DE car c'est un algorithme stochastique qui ira tester de nouvelles solutions ailleurs dans l'espace des paramètres.

Pour NM, l'astuce ne fonctionne plus car l'algorithme va avoir tendance à rester localisé dans une même zone. Si l'algorithme est coincé dans un "mur" (cf. Fig.3), i.e. tous les sommets du simplex (cf. Annexe B) retournent $\sim 10^{30}$, celui-ci ne saura plus dans quelle direction se déplacer. La solution à ce problème est d'ajouter une fonction qui va venir pénaliser le *likelihood* de manière continue plutôt qu'une large valeur.

5.2 Solutions mises en place

Soit $\mathscr{P}$ une fonction pénalité et soit $x \in [x_{\min}, x_{\max}]$ un paramètre avec $x_{\min}, x_{\max}$ les valeurs limites. Pénaliser le cas $x < x_{\min}$ revient à chercher une fonction avec les propriétés suivantes :

- $\mathscr{P} \xrightarrow{x \to -\infty} \infty$
- $\mathscr{P}(x_{\min}) = 0$
- $\left.\frac{d\mathscr{P}}{dx}\right|_{x=x_{\min}} = 0$

La dernière condition est ici pour assurer que la pénalité fonctionne aussi sur des méthodes de minimisation demandant des fonctions différentiables. Pour NM cette dernière n'est cependant pas importante. La première fonction pénalité qui fut mise en place dans PROF-CL est

$$\mathscr{P}_1(x) = \begin{cases} p(X) & \text{si } X = x/x_{\min} < 1 \\ p(2-X) & \text{si } X = x/x_{\max} > 1 \\ 0 & \text{sinon} \end{cases} \tag{5.1}$$

où on a défini

$$p(X) = \frac{\pi}{2}(X-1) - \tan\left[\frac{\pi}{2}(X+1)\right] \tag{5.2}$$

Bien que de classe $\mathscr{C}^1$, cette fonction n'adopte le comportement voulu que pour des valeurs de X positives (i.e. $\mathscr{P} \xrightarrow{x \to 0} \infty$). Cela ne pose pas de problèmes, en principe, pour un algorithme comme NM car celui-ci n'ira jamais tester des valeurs négatives de X si le paramètre initial n'est pas trop mauvais.

Cependant l'algorithme Powell (intégré dans la dernière version mais non testé) va quant à lui aller explorer l'espace des paramètres malgré un bonne valeur initiale. Comme Powell ne requiert pas de fonction différentiable et pour éviter tout problème, on intègre dans le code une seconde fonction pénalité seulement $\mathscr{C}^0$ cette fois définie comme

$$\mathscr{P}_2(x) = \begin{cases} \sqrt{|\tilde{X} - 1|} & \text{si } \tilde{X} > 1 \\ 0 & \text{sinon} \end{cases} \tag{5.3}$$

où $\tilde{X} = 2|X - 1/2|$ avec $X = (x - x_{\min})/(x_{\min} - x_{\max})$ permet d'assurer que les cas $x < x_{\min}$ et $x > x_{\max}$ sont traités de manière symétriques.

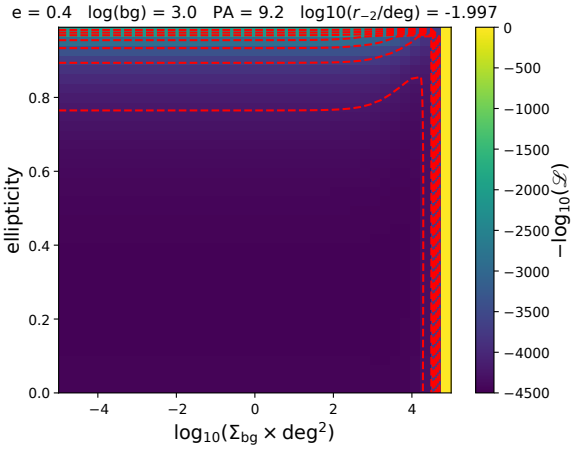


FIGURE 3 – Portion de l'espace des paramètres à PA et r_{-2} fixés. On observe un "mur" au-delà de $\log_{10}(\Sigma_{\rm bg}) \sim 4.7$ lorsque la densité surfacique du fond sort des limites imposées. Le "mur" a été ramené à 0 pour des questions de visibilité. Les traits pointillés représentent les courbes d'iso-likelihood.

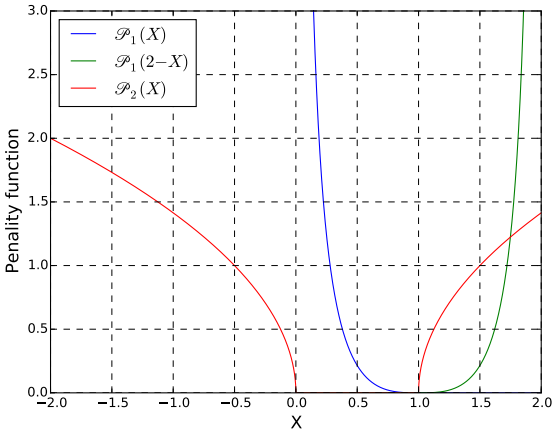
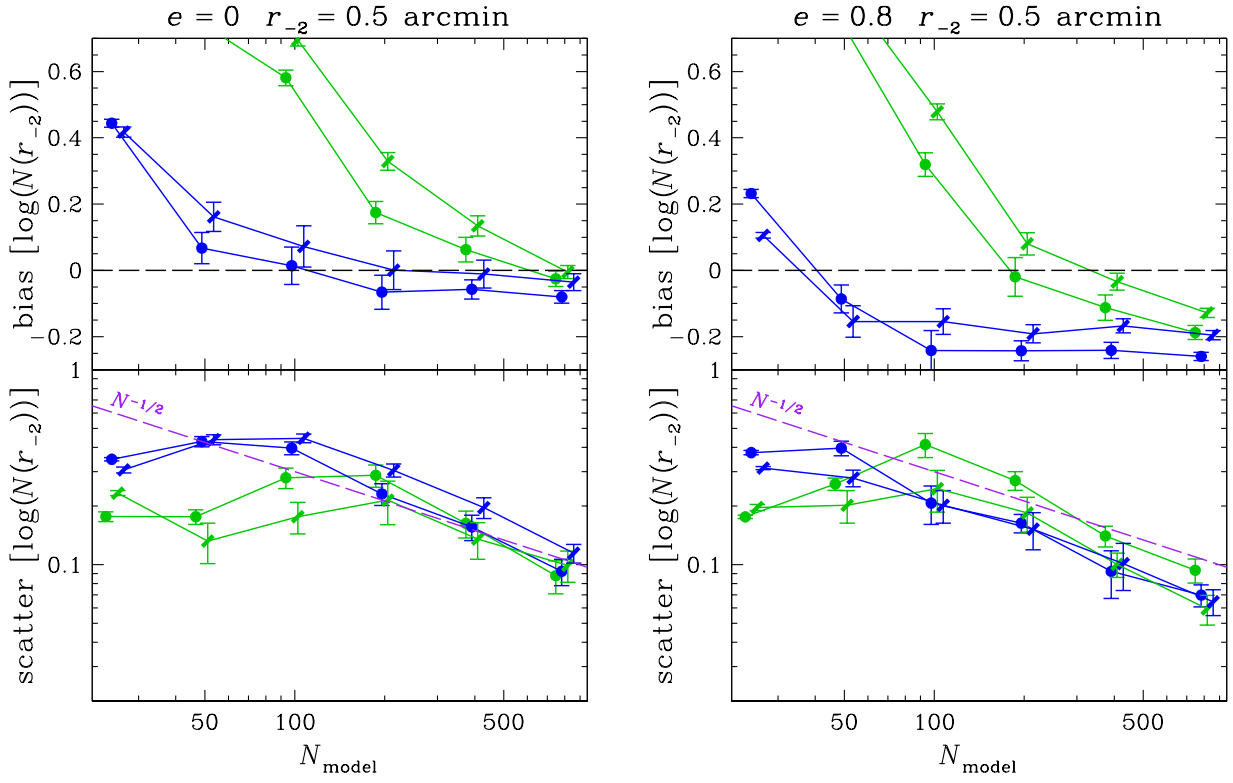


FIGURE 4 – Fonctions de pénalité intégrées dans PROF-CL. En bleu la fonction pénalité n°1 pour $X = x/x_{\rm min} < 1$. En vert son symétrique par rapport à l'axe $X = 1$ pour le cas $X = x/x_{\rm max} > 1$. En rouge la fonction de pénalité n°2 avec $X = (x - x_{\rm min})/(x_{\rm min} - x_{\rm max})$.



(a) (b)

FIGURE 5 – Biais et dispersion en richesse sur les mocks académiques. La richesse est calulée via les relations Eq.4.8 et Eq.4.4. DE est en très bon accord pour le cas circulaire pour des richesses réelles $N \gtrsim 50$ et trouve de meilleurs résultats que TNC, en accord avec les résultats trouvés pour le *log scale radius*. Les deux méthodes semblent converger vers un biais négatif pour de grandes ellipticités.

6 Performances du code sur les amas académiques

PROF-CL a pu être testé sur plusieurs milliers de mocks académiques en modes circulaire et elliptique, pour différentes valeurs de rayon caractéristique, d'ellipticité et de PA, avec et sans fond. Le fond jouant un rôle secondaire dans les résultats obtenus on se concentrera sur le biais et la dispersion du *log scale radius*, de l'ellipticité et du PA. Pour chaque jeu de paramètres $\boldsymbol{\theta} = \{\log_{10}(r_{-2}), e, \mathrm{PA}\}$, 100 amas ont été générés aléatoirement via Monte-Carlo suivant un profil NFW projeté dans un cercle de rayon $R_{\max}/(1-e)$ auquel un fond uniforme a été rajouté. De ces 100 amas, on a pu en tirer le biais moyen et la dispersion des paramètres. Les barres d'erreurs dans les graphes sont obtenues par *bootstrap*.

Globalement DE obtient de meilleurs résultats que TNC. En moyenne, les deux méthodes convergent vers le résultat attendu à la fois en circulaire et en elliptique pour des richesses de l'ordre de $\gtrsim 500$. On peut cependant faire certaines remarques :

— le *log scale radius* est mieux trouvé dans le cas circulaire qu'elliptique pour les deux méthodes, y compris pour des amas d'ellipticité ~ 0.5
— DE calcule une très bonne richesse pour $N \gtrsim 50$ en moyenne mais la dispersion est large
— TNC donne de meilleurs résultats en terme d'ellipticité pour des amas circulaires
— DE sous-estime la richesse à large ellipticité

Le premier point n'est cependant plus vrai pour des ellipticités plus grandes. Au delà de $e \sim 0.5$ on observe l'inverse avec un biais pour le cas elliptique qui converge très rapidement vers 0 tandis que le circulaire reste autour de ~ -0.2dex.

Contrairement à TNC, la dispersion de DE peut être assez facilement corrigée dans la pratique. Cela est dû à l'aspect stochastique de l'algorithme qui va généralement obtenir un résultat excellent pour environs 5 lancers sur le même amas. Un moyen d'améliorer le scatter serait de diminuer la population des solutions dans l'espace des phases (cf. Section B) mais d'augmenter le nombre de lancers de DE. Bien que diminuer la population va réduire la convergence de l'algorithme, le fait de le lancer plusieurs fois va avoir pour effet de lui faire prendre des chemins différents dans l'espace des paramètres et donc d'augmenter le volume cherché. Une seconde manière d'améliorer la convergence et de réduire la dispersion serait de réutiliser à chaque nouveau lancer la meilleure solution des lancers précédents comme valeur initiale.

Le dernier point est étonnant compte tenu du fait que même le résultat en elliptique est biaisé négativement, et ceci malgré des résultats excellents en termes de *log scale radius*, d'ellipticté et de PA. Un tel biais pourrait peut-être provenir d'une sur-évaluation de la densité surfacique du fond, ce qui entraînerait une diminution du nombre de galaxies dans la sphère.

7 Étude de PROF-CL sur les amas AMICO

7.1 Méthodologie pour l'analyse de la performance de PROF-CL sur les données d'AMICO

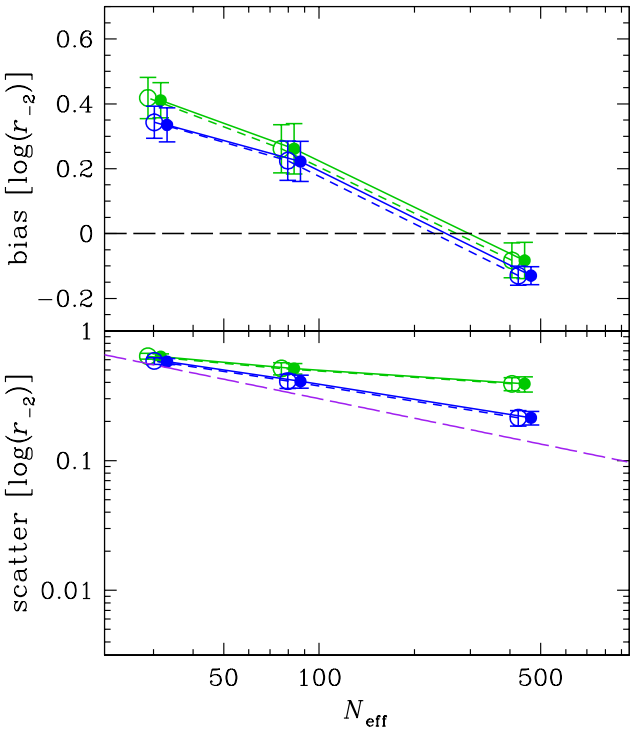


FIGURE 6 – Biais et dispersion en *log scale radius* sur les amas trouvés par AMICO. Le biais est calculé par rapport au rayon trouvé par le NFW tronqué sur le halo correspondant. Pour éviter de donner trop de poids aux données à faible ratio S/N les résultats ont été binnés dans trois intervalles de richesse $[20, 50]$, $[67, 150]$, > 200. DE se comporte légèrement mieux que TNC en terme de biais et de dispersion.

Contrairement aux mocks académiques provenant des codes d'Artis ou de Mamon, les données fournies par AMICO ne permettent pas d'obtenir les paramètres réels des amas qui se sont formés dans la simulation. Le code AMICO, en plus de fournir les identifiants des galaxies et leur probabilité, ainsi qu'un matching galaxies-amas, nous donne des informations supplémentaires au niveau des amas telles que leur richesse[3], leur redshift photométrique, l'erreur sur le redshift et le rapport S/N.

Afin de pouvoir étudier la fiabilité de PROF-CL sur les amas d'AMICO, il nous faut donc déterminer la valeur "réelle" du rayon caractéristique directement sur la simulation. Une solution aurait été de fitter directement sur les données 3D un profil $\rho_{\rm NFW}$. Celles-ci n'étant pas disponibles, la solution trouvée fut d'ajuter le *log scale radius* depuis les données projetées des galaxies du halo. Ces galaxies étant définies dans une zone quasi-sphérique déterminée par l'algorithme *Friends-of-Friends* qui a cherché les halos dans la simulation, il faut non pas ajuster un modèle NFW mais un modèle analogue NFW tronqué [Mamon et al., 2010] (qui rajoute un paramètre supplémentaire). Nous nous contenterons dans ce qui suit d'ajustements circulaires.

7.2 Matching des données AMICO et Euclid

Avant de pouvoir étudier les performances de la dernière version de PROF-CL sur les amas du code AMICO, il est nécessaire d'extraire des différents fichiers d'AMICO et d'Euclid les informations intéressantes. AMICO fournit deux fichiers : `detection.fits` et `association.fits`. Le premier contient une liste d'identifiants des amas détectés ainsi que certaines de leurs propriétés. Le second fournit une corresondance entre les identifiants des amas, ceux des galaxies qui leur appartiennent et les *memberships*. Il faut noter que les galaxies peuvent potentiellement appartenir à plus d'un amas. Un troisième fichier (dit 300deg^2) fait quant à lui le lien de manière bijective entre les identifiants et les coordonnées des galaxies.

7.2.1 Matching des données AMICO

Pour obtenir les positions des galaxies, il est nécessaire d'aller faire correspondre les idenfiants des galaxies dans le fichier association avec le catalogue complet de 300deg^2 de la simulation. Celui-ci faisant plus de 7GB, la difficulté principale a résidé dans le fait de faire correspondre coordonnées, identifiants et *memberships* le plus efficacement possible.

On a commencé par ne stocker dans un tableau que les identifiants des amas sélectionnés ainsi que leur propriétés. On a alors appliqué un masque $\mathscr{M}_1$ au fichier association pour ne garder que les identifiants

3. La richesse calculée par AMICO n'est pas une mesure directe du nombre de galaxies dans l'amas mais plutôt une mesure de l'intensité des pics dans la carte 3D. Il est aussi possible d'estimer la richesse en sommant memberships des galaxies.

et *memberships* des galaxies liées aux amas sélectionnés. La liste des identifiants de galaxies $L_{\rm d}$ issue du fichier association pouvant contenir des doublons, on applique un second masque $\mathscr{M}_2$ qui ne garde que la première occurence de chaque ID. On obtient une nouvelle liste $L_{\rm ID}$ que l'on peut relier à l'autre (en python) par la relation $L_{\rm d} = L_{\rm ID}[\text{indices}]$, où indices est une liste d'indices fournit par $\mathscr{M}_2$.

En ne gardant dans le fichier 300deg^2 que les lignes (contenant ID et positions de galaxies) pour lesquelles les ID de galaxies sont dans $L_{\rm ID}$, et en triant le nouveau tableau par ID croissant, on obtient un nouveau tableau qui a la même structure que $L_{\rm ID}$. Ainsi, tout comme pour cette dernière liste, en passant en paramètre indices pour chaque colonne du tableau on reconstruit un tableau complet liant chaque IDs d'amas et ses propriétés avec les IDs des galaxies, leur position et leur *membership*.

7.2.2 Matching des données d'Euclid

Pour pouvoir étudier les performances de PROFCL avec AMICO, il nous faut aussi être en mesure de fitter un profil NFW tronqué directement sur les halos issus de la simulation. La fragmentation et l'over-merging [4] assez élevés d'AMICO vont avoir tendance à ce que l'algorithme décompose les halos en plusieurs amas ou au contraire associe à un amas plusieurs halos. On s'attend donc à ce que les galaxies associées aux halos ne soient pas les mêmes que celles associées aux amas.

Le matching dans le cas des halos est légèrement plus complexe. AMICO fournit un fichier fournissant un lien entre amas détectés et halos associés. Un amas peut apparaître plusieurs fois dans le fichier s'il y a over-merging et un halo de même s'il y a de la fragmentation. Puisque qu'un amas peut être lié à plusieurs halos, on n'est pas en mesure dans certains cas de savoir à quel halo comparer l'amas. Dans ce cas, on choisit celui le plus proche de l'amas comme point de comparaison. On peut ainsi construire un tableau liant bijectivement amas détecté et halo issu de la simulation.

Une fois le matching amas-halo effectué il faut lier galaxies et halos via un cinquième fichier (dit *blind*) équivalent au catalogue complet à l'exception que cette fois à chaque galaxie ne correspond qu'un seul halo. Si l'on appelle $L_{\rm h}$ la liste des halos bijectivement associée à celle des amas ($L_{\rm a}$), lier les coordonnées des galaxies aux halos revient à lier celles-ci à $L_{\rm a}$ dans un premier temps, puis à $L_{\rm h}$ via la relation $L_{\rm h} \leftrightarrow L_{\rm a}$.

7.3 Analyse des performances du code sur les amas d'AMICO

Pour analyser PROF-CL sur les amas d'AMICO on a sélectionné un ensemble de 1000 amas sur lesquels environs 550 ont pu être associés à des halos. Les amas ont été choisis selon leur *rank* croissant, c'est-à-dire selon leur rapport signal-bruit décroissant, dans 10 intervalles de 100 amas de telle sorte à ce que les résultats obtenus soient aussi bien valables pour des amas très massifs et bien résolus que pour des amas peu riches avec une forte composante de bruit. Par manque de temps, les amas AMICO n'ont pu être analysés qu'en NFW circulaire seulement et les halos correspondant en NFW tronqué circulaire.

Les résultats sur le *log scale radius* sont représentés en Fig.6. Comme pour les mocks académiques, DE donne de meilleurs résultats que TNC mais la différence est bien moins prononcée. Les biais ont des valeurs similaires à ceux des mocks académiques. La dispersion décroît moins rapidement pour les amas AMICO ce qui est d'une part dû au binning effectué et d'autre part à la grande diversité des amas (cf. Annexe D). Ne connaissant pas l'ellipticité réelle des amas, on n'est pas en mesure de pouvoir séparer amas circulaires et elliptiques de manière automatique ce qui induit nécessairement une dispersion plus grande. À cela vient s'ajouter le fait que la fragmentation va avoir tendance à induire un biais négatif dans le rayon caractéristique tandis que *l'overmerging* va introduire une incertitude dans le choix de l'amas sur lequel effectuer la comparaison ce qui peut en principe introduire un nouveau biais. Enfin, il est important de noter que malgré la grande efficacité d'AMICO à détecter des sous-structures de résolution plus faible dans les données issues de la simulation par *Optimal Filtering*, on remarque que dans un certain nombre de cas des sous-structures non détectées par AMICO apparaissent dans les amas. Si ces structures sont bien indépendantes les unes des autres cela va nécessairement introduire statistiquement un biais positif pour le rayon.

Compte tenu des potentielles origines des biais cité ci-dessus, les résultats obtenus sont assez encourageants, bien que des statistiques plus poussées sur plus d'amas seraient utiles.

4. Fraction d'amas associés à plus d'un halo

8 Conclusion et ouverture

Pendant ce stage, j'ai pu participer à l'amélioration et à l'analyse des performances du code - toujours en développement - PROF-CL qui cherche à déterminer les caractéristiques d'amas de galaxies à partir des futures données de la mission Euclid. Ces deux mois très enrichissants furent un moyen d'étendre mes connaissances dans des domaines variés, autant au sujet des simulations cosmologiques à N-corps et des codes semi-analytiques associés que sur les amas de galaxies de manière générale et de leurs propriétés, en passant par tout un ensemble de techniques et algorithmes parmi lesquels certains que je n'ai pu détailler, voire mentionner (spline cubiques, bootstraps, etc).

Le code dans sa version actuelle est fonctionnel à la fois sur les mocks académiques et sur les données AMICO. Les difficultés majeures qui ont pu être rencontrées pendant le stage (séparation médiane non fonctionnelle, gestion des données AMICO, équation de probabilité de position non valable en elliptique, etc) ont pu être entièrement résolues.

Certaines difficultés rencontrées pendant le stage ou certaines parties du code sur lesquelles j'ai pu travailler n'ont pas été mentionnées dans ce rapport soit car elle ne représentent pas la majorité du travail effectué, soit car leur implémentation n'a pas fourni de résultats significatifs sur les performances du code et n'ont donc pas été étudiées suffisamment en profondeur. Qui plus est certaines fonctionnalités supplémentaires comme la possibilité de rajouter le rencentrage comme paramètre supplémentaire dans PROF-CL (en réalité 2 paramètres $RA_{\rm cen}$ et $Dec_{\rm cen}$) n'ont pas non plus été étudiées par manque de temps et de fait ne sont pas citées dans ce rapport. La dépendance du code au fond a été survolée pour des questions de délai, mais aussi car il s'est avéré pendant le stage que son impact restait secondaire sur les autres paramètres.

De nombreuses fonctionnalités peuvent encore être ajoutées/améliorées comme le fait de prendre en compte les couleurs ou les magnitudes des galaxies dans les amas pour potentiellement éliminer les galaxies d'avant-plan ou d'arrière-plan. Néanmoins, parmi les futures fonctionnalités, la plus prometteuse est très certainement le *deblending* (commencée mais pas terminée par manque de temps), c'est-à-dire la possibilité de tester si un amas issu d'AMICO ou de PZWav est en réalité constitué de plusieurs sous-amas. En terme d'équations il ne devrait pas y avoir trop de difficultés, cependant le *deblending* pose de nouveaux problèmes. Par exemple, on devrait s'attendre à ce que, si on augmente les paramètres du modèle, on trouve de meilleurs résultats. La question est alors de savoir quand s'arrêter. Un autre problème est lorsque deux amas ont une masse similaire. En principe on pourrait imaginer que si un amas est plus massif que l'autre on fasse tourner une première fois PROF-CL avec recentrage pour un seul amas afin qu'il trouve approximativement le centre et la taille caractéristique de celui-ci, puis une seconde fois après avoir soustrait le premier amas à disons un rayon de pente -2. Dans le cas de deux amas de même masse ce type d'algorithme ne fonctionnera pas car le code aura du mal soit à trouver l'amas le plus massif (car il n'y en a pas), soit car le centre du premier amas sera très certainement placé à une distance médiane des deux centres, c'est-à-dire entre les deux amas. Or on s'attend en pratique à ce que de telles situations aient lieu. Et le problème peut devenir encore plus compliqué si les deux amas sont au même redshift mais en processus de fusion.

Références

[Adam et al., 2018] Adam, R. et al. (2018). EUCLID : galaxy cluster detection in the wide photometric survey - performance and algorithms selection. *Astronomy and Astrophysics*. Accepted but not yet published.

[Ascaso et al., 2015] Ascaso, B., Mei, S., and Benítez, N. (2015). Apples to apples A^2 - I. Realistic galaxy simulated catalogues and photometric redshift predictions for next-generation surveys. *Monthly Notices of the Royal Astronomy*, 453 :2515–2532.

[Bellagamba et al., 2018] Bellagamba, F., Roncarelli, M., Maturi, M., and Moscardini, L. (2018). AMICO : optimized detection of galaxy clusters in photometric surveys. *Monthly Notices of the Royal Astronomy*, 473 :5221–5236.

[BROYDEN, 1970] BROYDEN, C. G. (1970). The convergence of a class of double-rank minimization algorithms2. the new algorithm. *IMA Journal of Applied Mathematics*, 6(3) :222–231.

[Eisenhardt et al., 2008] Eisenhardt, P. R. M., Brodwin, M., Gonzalez, A. H., Stanford, S. A., Stern, D., Barmby, P., Brown, M. J. I., Dawson, K., Dey, A., Doi, M., Galametz, A., Jannuzi, B. T., Kochanek, C. S., Meyers, J., Morokuma, T., and Moustakas, L. A. (2008). Clusters of Galaxies in the First Half of the Universe from the IRAC Shallow Survey. *The Astrophysical Journal*, 684 :905–932.

[Laureijs et al., 2011] Laureijs, R., Amiaux, J., Arduini, S., Augueres, J. ., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., and et al. (2011). Euclid Definition Study Report. *ArXiv e-prints*.

[Łokas and Mamon, 2001] Łokas, E. L. and Mamon, G. A. (2001). Properties of spherical galaxies and clusters with an NFW density profile. *Monthly Notices of the Royal Astronomical Society*, 321 :155–166.

[Mamon et al., 2010] Mamon, G. A., Biviano, A., and Murante, G. (2010). The universal distribution of halo interlopers in projected phase space. Bias in galaxy cluster concentration and velocity anisotropy ? *Astronomy and Astrophysics*, 520 :A30.

[Merson et al., 2013] Merson, A. I., Baugh, C. M., Helly, J. C., Gonzalez-Perez, V., Cole, S., Bielby, R., Norberg, P., Frenk, C. S., Benson, A. J., Bower, R. G., Lacey, C. G., and Lagos, C. d. P. (2013). Light-cone mock catalogues from semi-analytic models of galaxy formation - I. Construction and application to the BzK colour selection. *Monthly Notices of the Royal Astronomy*, 429 :556–578.

[Mo et al., 2010] Mo, H., van den Bosch, F., and White, S. (2010). *Galaxy Formation and Evolution*. Cambridge University Press.

[Navarro et al., 1996] Navarro, J. F., Frenk, C. S., and White, S. D. M. (1996). The Structure of Cold Dark Matter Halos. *The Astrophysical Journal*, 462 :563.

[Nelder and Mead, 1965] Nelder, J. and Mead, R. (1965). A simplex method for function minimization comput. 7.

[Powell, 1964] Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2) :155–162.

[Sarazin, 1980] Sarazin, C. L. (1980). A maximum likelihood method for determining the distribution of galaxies in clusters. *The Astrophysical Journal*, 236 :75–83.

[Shewchuk, 1994] Shewchuk, Jonathan, R. (1994). An introduction to the conjugate gradient method without the agonizing pain. `http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf`.

[Springel, 2005] Springel, V. (2005). The cosmological simulation code GADGET-2. *Monthly Notices of the Royal Astronomy*, 364 :1105–1134.

[Springel et al., 2005] Springel, V., White, S. D. M., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas, P., Couchman, H., Evrard, A., Colberg, J., and Pearce, F. (2005). Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature*, 435 :629–636.

[Storn and Price, 1997] Storn, R. and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4) :341–359.

A Solutions analytiques de la densité de surface NFW/NFW tronqué

A.1 NFW circulaire

Pour un profil NFW circulaire, la densité surfacique donnée par la solution de Eq.3.4 peut être écrite comme

$$\Sigma_{\rm NFW}(R) = \frac{N(r_{-2})}{\pi r_{-2}^2}\tilde{\Sigma}_{\rm NFW}(R/r_{-2}) \tag{A.1}$$

La densité surfacique adimensionnée $\tilde{\Sigma}_{\rm NFW}$ étant donnée par [Łokas and Mamon, 2001]

$$\tilde{\Sigma}_{\rm NFW}(X) = \frac{1 - {\rm C}^{-1}(1/{\rm X})/|{\rm X}^2-1|^{1/2}}{X^2-1} \tag{A.2}$$

Où l'on a définit

$${\rm C}^{-1}({\rm Y}) = \begin{cases} \arccos(Y) & \text{si } {\rm R} > {\rm r}_{-2} \\ \operatorname{arccosh}(Y) & \text{si } {\rm R} < {\rm r}_{-2} \end{cases} \tag{A.3}$$

A.2 NFW circulaire tronqué

Contrairement aux "mocks académiques" les amas sélectionnés par AMICO pour le Cluster Challenge IV sont tronqués. En effet, une fois que la simulation à N-corps issue de [Springel et al., 2005] a fini de fonctionner, le code *Halo finder* 3D GADGET-2 [Springel, 2005] va aller chercher les amas de galaxies par une méthode de *Friends-of-Friends* (FoF). Cette étape va introduire dans les halos de matière noire un rayon limite de troncation $r_{\rm trunc} \sim r_{\rm vir}$. Les galaxies vont alors être introduites dans la simulation via une variante du code semi-analytique GALFORM [Merson et al., 2013] en utilisant la distribution spatiale des halos de matière noire. Ainsi le rayon de troncation issu du FoF va se retrouver à la fois dans les données finales de la simulation Euclid mais aussi dans les données d'AMICO.

Il est ainsi nécessaire quand on calcule la densité surfacique projeté NFW dans Eq.(3.4) d'effectuer la transformation $\infty \to r_{\rm trunc}$ au niveau de la borne supérieure.
Dans le cas considéré ici où $r_{\rm trunc} = r_{\rm vir}$ la solution de Eq.(3.4) avec la nouvelle borne supérieure s'écrit [Mamon et al., 2010]

$$\Sigma^{\rm trunc}_{\rm NFW}(R, r_{\rm vir}) = \frac{N(r_{-2})}{\pi r_{-2}^2}\tilde{\Sigma}^{\rm trunc}_{\rm NFW}\left(\frac{R}{r_{-2}}, c\right) \tag{A.4}$$

où $c = r_{-2}/r_{\rm vir}$ est la concentration de l'amas et où l'on a défini

$$\tilde{\Sigma}^{\rm trunc}_{\rm NFW}(X,c) = \frac{1}{2\log 2 - 1}\begin{cases} \frac{1}{(1-X^2)^{3/2}}\operatorname{arccosh}\left(\frac{c+X^2}{(c+1)X}\right) - \frac{1}{c+1}\frac{\sqrt{c^2-X^2}}{1-X^2} & \text{si } 0 < X < 1 \\ \frac{\sqrt{c^2-1}(c+2)}{3(c+1)^2} + \frac{(2-c-4c^2-2c^3)(X-1)}{5(c+1)^2\sqrt{c^2-1}} & \text{si } X = 1 \\ \frac{1}{c+1}\frac{\sqrt{c^2-X^2}}{X^2-1} - \frac{1}{(X^2-1)^{3/2}}\arccos\left(\frac{c+X^2}{(c+1)X}\right) & \text{si } 1 < X < c \\ 0 & \text{si } X = 0 \text{ ou } X > c \end{cases} \tag{A.5}$$

B Algorithmes de minimisation

Cinq algorithmes fonctionnels sont implémentés dans la dernière version de PROFCL : TNC, DE, NM, Powell et BFGs. N'ayant travaillé majoritairement que sur les deux premiers algorithmes nous ne rentrerons pas en détail dans les trois derniers.

B.1 Truncated Newtonian Conjugate (TNC)

TNC est une méthode par descente de gradient avec contraintes. Le terme *truncated* signifie qu'il y a un nombre d'itérations maximales au-delà duquel l'algorithme s'arrête. L'algorithme va chercher à résoudre un système du type $A\boldsymbol{x} = \boldsymbol{b}$ où $\boldsymbol{x}$ est la solution cherchée. Si l'on note n la dimension du problème, alors la solution peut être développée dans la base $\{\boldsymbol{c_i}\}$ des n vecteurs mutuellement conjugués de A comme $\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{c_i}$.[5]

L'algorithme va alors boucler sur j de 1 à n et va calculer à chaque étape une nouveau vecteur conjugué aux précédents ainsi que le coefficient x_j associé selon les formules[Shewchuk, 1994]

$$\boldsymbol{c_j} = \boldsymbol{r_j} - \sum_{j<i} \frac{\boldsymbol{p_j}^T A \boldsymbol{r_i}}{\boldsymbol{p_j}^T A \boldsymbol{p_j}} \boldsymbol{p_j} \quad x_j = \frac{\boldsymbol{p_j}^T \boldsymbol{r_j}}{\boldsymbol{p_k}^T A \boldsymbol{p_k}} \tag{B.1}$$

où $\boldsymbol{r_j} = \boldsymbol{b} - A\boldsymbol{x_j}$ est le résidu. La nouvelle solution est alors donnée par $\boldsymbol{x_{j+1}} = \boldsymbol{x_j} + x_j \boldsymbol{p_j}$ et le résidu est mis à jour. Si celui-ci est en dessous de la tolérance demandée ou si le nombre d'itérations maximal est atteint l'algorithme stoppe.

B.2 Differential Evolution (DE)

À l'inverse de TNC, DE est un algorithme stochastique non contraint[Storn and Price, 1997]. Celui-ci va remplir l'espace des paramètres avec un certain nombre de points de l'ordre du nombre de galaxies dans l'amas. À chaque itération il va muter les candidats entre eux afin de construire une nouvelle solution candidate et va appliquer un critère de sélection afin de décider si la nouvelle solution doit être gardée ou non.

Plusieurs types de critères existent. Celui couramment utilisé et employé dans PROF-CL peut être résumé comme suit

- Deux membres m_1, m_2 sont choisis au hasard dans la population et une mutation est calculée comme $\boldsymbol{m} = \boldsymbol{m_{best}} + \eta(\boldsymbol{m_1} + \boldsymbol{m_2})$ où η est un paramètre libre de mutation et $\boldsymbol{m_{best}}$ est le meilleur paramètre entre $\boldsymbol{m_1}$ et $\boldsymbol{m_2}$.
- Une nouvelle solution $\boldsymbol{\theta} = (\{\theta_i\})$ est calculée à partir de $\boldsymbol{m}$ et $\boldsymbol{m_{best}}$. Pour chaque composante θ_i un nombre aléatoire $r_i \in [0, 1[$ est tiré ; si $r_i > \gamma$ avec γ un paramètre libre de recombinaison alors $\theta_i = m_i$, sinon $\theta_i = m_{\mathrm{best,i}}$.
- Une composante aléatoire dans la nouvelle solution est modifiée aléatoirement.
- $\boldsymbol{m_{best}}$ est remplacé par $\boldsymbol{\theta}$ si ce dernier est meilleur. Idem pour le meilleur candidat dans l'ensemble de la population.

Cet algorithme possède les désavantages de demander plus d'itérations que des algorithmes non-stochastiques et de ne pas toujours garantir de trouver une solution optimale au problème.

5. On dit que deux vecteurs $\boldsymbol{a}$ et $\boldsymbol{b}$ sont conjugués par rapport à A s'ils obéissent à la relation $\boldsymbol{a}^T A \boldsymbol{b} = 0$

B.3 BFGS, Powell et NM

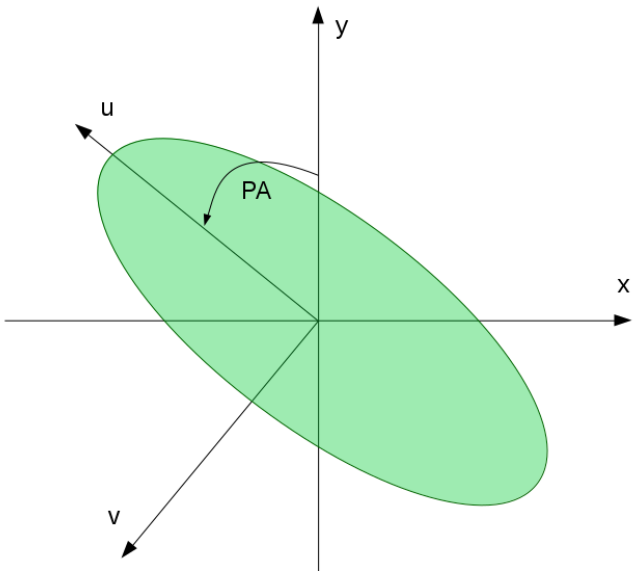
L'algorithme Broyden–Fletcher–Goldfarb–Shanno (BFGS)[BROYDEN, 1970] est une méthode de quasi-Newton non contrainte qui va calculer à chaque étape une approximation de la matrice Hessienne de la fonction à minimiser. Cet algorithme fonctionne en principe sur des fonctions seulement $\mathscr{C}^0$ mais peut potentiellement converger vers un minimum local au lieu d'une minimum global.

La méthode de Powell est un algorithme de minimisation par directions conjuguées[Powell, 1964] qui n'a besoin que de fonctions $\mathscr{C}^0$. Celui-ci va chercher à minimiser la fonction dans une direction à la fois selon k vecteurs, puis va construire une nouvelle solution comme combinaison linéaire de la solution initiale et des déplacements dans les k directions. Cette méthode permet seulement de trouver un minimum local.

L'algorithme Nelder-Mead[Nelder and Mead, 1965] est une méthode d'optimisation heuristique basée sur le principe de simplex. Si l'espace des paramètres est de dimensions n celui-ci va le peupler avec un simplex de dimension $n+1$ ($\mathrm{n}+1-$ polytope). S'en suit une série d'étapes brièvement résumées ci-dessous :

- Si les sommets du simplex sont notés $\boldsymbol{m_i}$ avec $f(\boldsymbol{m_1}) < ... < f(\boldsymbol{m_{n+1}})$ alors on cherche le symétrique $\boldsymbol{m_r}$ de $\boldsymbol{m_{n+1}}$ par rapport au centre de masse des n points restants (réflection) et on remplace l'ancien sommet par le nouveau s'il est meilleur.
- Si c'est la meilleur solution on calcule un développement $\boldsymbol{m_e}$ autour de $\boldsymbol{m_r}$ et on remplace $\boldsymbol{m_r}$ par $\boldsymbol{m_e}$ si ce dernier est meilleur.
- Si la réflection est moins bonne que $\boldsymbol{m_n}$ on calcule une contraction autour de ce point ; si la contraction est meilleur on remplace l'ancien sommet par le nouveau
- On recentre tous les points autour de $\boldsymbol{m_1}$ et on réduit leur distance d'un facteur σ

C Systèmes de coordonnées utilisés dans PROF-CL



Dans le code PROF-CL plusieurs systèmes de coordonnées sont utilisés suivant que l'on considère des amas circulaires ou elliptiques. On en fait ici la synthèse.

C.1 Coordonnées de référence

Le premier jeu de coordonnées utilisé est celui à partir duquel sont défini les positions des amas et des galaxies dans les simulations académiques et semi-analytiques. Les coordonnées utilisées sont les coordonnées équatoriales (x, y) où $\hat{\boldsymbol{x}}$ pointe vers l'Ouest et $\hat{\boldsymbol{y}}$ vers le Nord.

De fait si l'on appelle RA, Dec respectivement l'ascension droite, la déclinaison d'une galaxie alors on a les relations

$$x = -(\mathrm{RA} - \mathrm{RA_{cen}})\cos(\mathrm{Dec})$$
$$y = \mathrm{Dec} - \mathrm{Dec_{cen}}$$

C.2 Coordonnées locales elliptiques centrées

Pour les amas elliptiques centrés il est utile de définir des nouvelles coordonnées (u, v) liées à l'amas définies telles que $\hat{\boldsymbol{u}}$ soit aligné le long du demi-grand axe de l'ellipse en direction du Nord à PA nul et $\hat{\boldsymbol{v}}$ le long du demi-petit axe en direction de l'Est à PA nul.

Ainsi les nouvelles coordonnées s'écrivent

$$u = -x\sin(\mathrm{PA}) + \mathrm{y}\cos(\mathrm{PA})$$
$$v = -x\cos(\mathrm{PA}) - \mathrm{y}\sin(\mathrm{PA})$$

Le rayon projeté réel est naturellement donné par $R = \sqrt{u^2 + v^2}$. Cependant, lors du traitement des amas elliptiques on est amené à "circulariser" l'amas pour pouvoir en extraire les probabilités de position des galaxies. Cela revient en pratique à "décompresser" l'amas selon son demi-grand axe en le divisant par un facteur $1 - e$, i.e. en définissant une nouvelle coordonnée $v^{'} = v/(1 - e)$. Il est alors pratique de définir un rayon elliptique comme

$$\mathscr{R} = \sqrt{u^2 + v^{'2}} = \sqrt{u^2 + \left(\frac{v}{1-e}\right)^2} \tag{C.1}$$

D Exemple d'amas AMICO