DEA Astrophysique et méthodes associées

DEA Dynamique des Systèmes gravitationnels

PROJET INFORMATIQUE C++ 2001-2002

Frédéric Arenou

UMR 8633 du CNRS et DASGAL, Observatoire de Paris

Revision : 1.6 , Date : 2001/09/19 16 :53 :56

La seule méthode qui permette l'estimation des distances d'étoiles sans aucune hypothèse physique est l'utilisation des parallaxes trigonométriques : l'angle avec lequel est vue l'orbite terrestre est $\approx$ inversement proportionnel à la distance de l'objet observé. D'où la définition du parsec comme la distance d'une étoile ayant une parallaxe annuelle de une seconde d'arc.

À partir de ces distances d'étoiles proches (calibration primaire), on utilise ensuite d'autres méthodes, par exemple la relation période-luminosité de céphéïdes, pour avoir des étalons secondaires.

Le but de ce projet est de montrer, en utilisant une méthode de maximum de vraisemblance sur un échantillon d'étoiles lointaines, que les parallaxes obtenues avec le satellite européen Hipparcos n'ont pas d'erreur systématique.

1 Le problème astrophysique

On va utiliser des étoiles lointaines pour lesquelles on a une estimation de la magnitude absolue grâce à des mesures photométriques. En utilisant la loi de Pogson, la magnitude absolue est définie comme :

$$M_V = m_V + 5\log\pi + 5 - A_V \tag{1}$$

où m_V est la magnitude apparente de l'étoile, A_V est l'absorption interstellaire et π est la parallaxe ; comme la magnitude dépend de la bande dans laquelle on observe l'étoile, l'indice V précise ici que c'est dans le visible ; le $\log$ est décimal.

Grâce à des calibrations faites dans le passé, on connaît la magnitude absolue de certaines étoiles par l'intermédiaire de leur photométrie, et l'absorption également, donc le module de distance que l'on notera ici $t = m_V - M_V - A_V$.

On va se servir de ce module de distance, dont on voit bien qu'il contient de l'information concernant la parallaxe (photométrique) via l'équation 1. On va se servir également de la distribution spatiale de nos étoiles, parce que l'on sait qu'elles sont distribuées exponentiellement au-dessus du plan galactique ; ici encore on voit que cette distribution contient de l'information concernant la distance. Et enfin, bien entendu, on va utiliser la parallaxe trigonométrique mesurée par Hipparcos.

On va donc faire un modèle qui contient toutes ces informations, et on va ajouter à ce modèle un paramètre z, le décalage possible des parallaxes trigonométriques, et un paramètre

k qui indiquera si l'on a correctement estimé la précision de ces parallaxes. Le but est donc de déterminer la valeur de z et k, dont on espère qu'ils valent respectivement 0 et 1 environ, ainsi que la précision sur ces paramètres.

On voit maintenant pourquoi on va utiliser des étoiles lointaines : la parallaxe est très petite, donc la parallaxe trigonométrique mesurée ne contient quasiment que du $\ll$ bruit $\gg$ (les erreurs de mesure), tandis que la véritable information provient des données photométriques et spatiales.

Le test sur le $\ll$ point-zéro $\gg$ des parallaxes est évidemment important, puisqu'il faut savoir si les parallaxes sont non-biaisées avant de les utiliser. Une des applications directe est justement de calibrer les magnitudes absolues, des étoiles proches cette fois-ci.

Le modèle décrit ci-dessous permettrait également de calibrer la magnitude absolue : on n'essaierait pas de déterminer z et k, mais on écrirait la magnitude absolue comme une fonction de certains indices photométriques, et on essaierait de déterminer les paramètres de cette fonction. Ce n'est pas notre propos ici, mais il faut juste penser à avoir une vision générale dans l'implantation pratique que nous allons réaliser : il peut y avoir d'autres paramètres à déterminer. On peut également se dire que l'on n'a pas utilisé les informations cinématiques de chaque étoile, et il faut penser le programme en fonction de développements futurs de ce type.

Pour le principe, on écrit ci-dessous comment on est arrivé aux équations qui sont utilisées, mais on peut ignorer les commentaires et s'en servir simplement comme d'un formulaire. L'explication et la justification des méthodes statistiques adoptées pourra être vue dans le cours de statistiques de D. Pelat [3].

2 Formulation mathématique

Pour chaque étoile, la densité de probabilité (pdf) d'observer la parallaxe trigonométrique π_{H} sachant le module de distance t, la latitude galactique b, et les paramètres inconnus z et k peut s'écrire :

$$\begin{aligned} f(\pi_{\text{H}}|t,b,z,k) &= \frac{g(\pi_{\text{H}},t,b|z,k)}{h(t,b|z,k)} \\ &= \frac{g(\pi_{\text{H}},t,b|z,k)}{\int_{-\infty}^{+\infty} g(\pi_{\text{H}},t,b|z,k)d\pi_{\text{H}}} \end{aligned} \tag{2}$$

où $g(.)$ peut être écrite comme la loi marginale :

$$g(\pi_{\text{H}},t,b|z,k) = \int_0^{+\infty} q(\pi_{\text{H}},t,b|\pi,z,k)p_4(\pi)d\pi \tag{3}$$

avec π la *vraie* parallaxe (hélas inconnue sinon on ne se donnerait pas autant de mal !), et qui est positive puisque c'est l'inverse de la distance. La pdf $q(.)$ peut s'écrire comme le produit des pdf indépendantes :

$$q(\pi_{\text{H}},t,b|\pi,z,k) = p_1(\pi_{\text{H}}|\pi,k,z).p_2(t|\pi).p_3(b|\pi) \tag{4}$$

où les pdfs conditionnelles $p_1 ... p_4$ sont déterminées ci-dessous par les équations 6, 7, 8.

On arrive maintenant à la partie où on indique les lois de probabilités qui sont adoptées. On utilisera en particulier la loi Gaussienne classique (où $\Pi = 3.14\ldots$!)

$$\mathcal{G}_x(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{5}$$

2.1 Loi des parallaxes observées

On peut considérer que les erreurs de mesure des parallaxes trigonométriques Hipparcos $\pi_{\rm H}$ de précision $\sigma_{\rm H}$ sont $\approx$ gaussiennes [1]. On introduit donc nos deux paramètres inconnus dans la loi de probabilité :

$$p_1(\pi_{\rm H}|\pi,k,z) = \left| \begin{array}{l} \frac{\mathcal{G}_{\pi_{\rm H}}(\pi+z,k\sigma_{\rm H})}{\int_{\pi_{\rm H}^-}^{\pi_{\rm H}^+}\mathcal{G}_{\pi_{\rm H}}(\pi+z,k\sigma_{\rm H})d\pi_{\rm H}} \text{ si } \pi_{\rm H}\in[\pi_{\rm H}^-,\pi_{\rm H}^+] \\ 0 \text{ sinon} \end{array} \right. \quad (6)$$

On a pris en compte la possibilité que les parallaxes Hipparcos aient pu être rejetées de l'échantillon si elles étaient en-dehors de l'intervalle $[\pi_{\rm H}^-,\pi_{\rm H}^+]$. Pour simplifier, on ne fera pas de sélection, donc le dénominateur vaudra 1, et on se retrouve avec une gaussienne décalée de z et dilatée d'un facteur k.

2.2 Loi photométrique

On supposera également que les erreurs de la magnitude absolue calibrée M_V sont gaussiennes autour de la vraie valeur ${M_V}'$, c-à-d qu'elles suivent la loi $\mathcal{G}_{M_V}({M_V}',\sigma_M)$. De même, on supposera que les erreurs sur la magnitude apparente et sur l'absorption sont gaussiennes, d'où les lois $\mathcal{G}_{m_V}({m_V}',\sigma_m)$ et $\mathcal{G}_{A_V}({A_V}',\sigma_{A_V})$, où le « prime » désigne la vraie quantité. En conséquence, t est aussi gaussien autour de $t' = {m_V}' - {M_V}' - {A_V}' = -5\log\pi - 5$, et sa variance est $\sigma_t^2 = {\sigma_m}^2 + {\sigma_M}^2 + {\sigma_A}^2$.

On prend ici également en compte une possible censure sur le module de distance en-dehors de l'intervalle $[t^-,t^+]$. Cette fois, on ne peut pas s'en passer, pour une raison simple : c'est notre instrumentation imparfaite qui nous empêche de voir les étoiles les moins brillantes, et qui borne notre intervalle en limite supérieure. Si l'on n'en tient pas compte, alors on va sélectionner préférentiellement les étoiles les plus *intrinsèquement* brillantes (i.e. la magnitude absolue la plus grande), et donc on va biaiser notre estimation : c'est le biais de Malmquist [2]. En résumé, la loi que l'on adopte pour le module de distance est donc :

$$p_2(t|\pi) = \left| \begin{array}{l} \frac{\mathcal{G}_t(-5\log\pi-5,\sigma_t)}{\int_{t^-}^{t^+}\mathcal{G}_t(-5\log\pi-5,\sigma_t)dt} \text{ si } t\in[t^-,t^+] \\ 0 \text{ sinon} \end{array} \right. \quad (7)$$

2.3 Loi spatiale

Dans le dernier terme de l'équation 3, il nous faudra calculer le produit $p_3(b|\pi)p_4(\pi) = p(b,\pi)$. On va supposer que la distribution spatiale de nos étoiles est indépendante de la longitude galactique (i.e. on suppose une distribution uniforme dans le plan), d'où :

$$p(b,\pi) \propto p(r,l,b)|\frac{\partial r}{\partial \pi}| = \frac{1}{\pi^2}p(r,l,b) = \frac{1}{\pi^2}p(X,Y,Z)|J|$$

où $J = r^2\cos b$ est le Jacobien de la transformation des coordonnées cartésiennes héliocentriques (X,Y,Z) vers les coordonnées sphériques associées (r,l,b), où $r = \frac{1}{\pi}$ est la distance à l'étoile. Il est assez réaliste de considérer que Z suit une loi exponentielle avec une échelle de hauteur moyenne h_Z, c-à-d $p(X,Y,Z) \propto \frac{1}{2h_Z}e^{-\frac{|Z|}{h_Z}}$. Finalement, on obtient :

$$p_3(b|\pi)p_4(\pi) \quad \propto \quad \frac{\cos b}{2h_Z\pi^4}e^{-\frac{|\sin b|}{\pi h_Z}} \quad (8)$$

Il y a un coefficient de proportionnalité, mais qui importe peu, car il disparaît dans l'équation 2.

2.4 maximum de vraisemblance

Le principe du maximum de vraisemblance (MLE) est simple : la probabilité d'obtenir notre échantillon est le produit des probabilités des étoiles individuelles, et on va chercher la valeur du couple (z, k) qui maximise ce produit. L'estimateur du maximum de vraisemblance est donc celui qui maximise la log-vraisemblance de notre échantillon de taille n :

$$\ln\ \mathcal{L}(z,k) \quad = \quad \sum_{i=1}^{n} \ln f(\pi_{\mathrm{H}i}|t_i, b_i, z, k) \tag{9}$$

Dit d'une autre manière, c'est cette valeur qui fournit la plus grande probabilité d'avoir observé la parallaxe Hipparcos, sachant les propriétés photométriques et astrométriques de chaque étoile. Cet estimateur qui est asymptotiquement non-biaisé va être calculé numériquement en utilisant les équations (2) à (9).

Une fois obtenue la valeur des paramètres, il faut connaître leur précision (une estimation sans barre d'erreur n'a aucune signification !). L'erreur formelle sur nos deux paramètres et la corrélation entre eux est approchée numériquement en utilisant l'inverse de la matrice d'information de Fisher (nommé le Hessien de $\ln\ \mathcal{L}$) :

$$(\mathbf{V}) \quad = \quad \left(\frac{\partial^2 \ln\ \mathcal{L}(\ldots\theta_i\ldots)}{\partial\theta_i \partial\theta_j} \right)^{-1} \tag{10}$$

évaluée au point (z, k) obtenu précédemment.

2.5 Ajustement et estimation robuste

Une fois trouvés z et k, on va juger de la qualité de l'ajustement que l'on vient d'effectuer.

Pour cela, on prédit la valeur que l'on aurait dû observer par la valeur moyenne (espérance) de π_{H} qui suit la loi $f(\pi_{\mathrm{H}}|t, b, z, k)$. Ceci nous permet donc de calculer les résidus O-C (Observé moins Calculé) des parallaxes par :

$$\delta_i = \pi_{\mathrm{H}i} - \int_{-\infty}^{+\infty} \pi_{\mathrm{H}} f(\pi_{\mathrm{H}}|t_i, b_i, z, k) d\pi_{\mathrm{H}} \tag{11}$$

Si le modèle adopté est correct, et si il n'y a pas de points « aberrants », alors les résidus normalisés $\frac{\delta_i}{\sigma_{\delta_i}}$ doivent être de moyenne nulle et de variance 1, et être indépendants des quantités observées t et b. On fera les graphiques des résidus normalisés en fonction de ces quantités pour en juger. On tracera également leur histogramme.

Comme on ne connaît pas facilement la distribution exacte des résidus normalisés, on a procédé par simulation pour connaître l'intervalle de confiance à 95% dans lequel ils doivent se trouver. Cet intervalle est indiqué au 2.7 ci-dessous.

Si une étoile de notre échantillon a son résidu normalisé en-dehors de l'intervalle de confiance, alors on considère qu'il s'agit d'un point aberrant, et on marque l'étoile qui est la plus aberrante.

Puis on refait tourner l'algorithme en excluant l'étoile marquée, et on teste à nouveau les résidus, et on itère l'opération. Quand plus aucune étoile n'est aberrante dans notre échantillon réduit, on considère que c'est terminé et que les valeurs de k, z obtenues sont les valeurs définitives. Rien ne dit qu'en pratique il y aura effectivement des valeurs aberrantes, mais il vaut mieux le prévoir, sinon les résultats risqueraient d'être fortement influencés par ces valeurs.

2.6 simulations

Pour tester un programme, il faut disposer d'un jeu de données simulées, pour lequel on sait quelles valeurs de paramètres on doit obtenir...et vérifier que le programme redonne bien ce que l'on attend.

Pour faire des simulations, on peut par exemple choisir une $\ll$ vraie $\gg$ parallaxe pour chaque étoile, une $\ll$ vraie $\gg$ magnitude absolue, ajouter du bruit à chacuns, et donc obtenir les données observées.

On peut également utiliser le fait que la pdf *a posteriori* de la vraie parallaxe est $s(\pi|t,b) \propto p_2(t|\pi).p_3(b|\pi).p_4(\pi)$. Pour chaque étoile i, un π_i est tiré de la distribution $s(\pi_i|t_i,b_i)$, puis une parallaxe $\ll$ observée $\gg$ $\pi_{\mathrm{H}i}$ est tirée à partir de $p_1(\pi_{\mathrm{H}i}|\pi_i,k,z)$.

En effectuant un nombre important de simulations, les résultats obtenus devraient non seulement être comptatibles avec les paramètres introduits en entrée, mais de plus leur dispersion devrait être celle qui est indiquée par les erreurs formelles calculées.

Ce sont également ces simulations qui permettent d'obtenir l'intervalle de confiance mentionné précédemment.

2.7 application

En pratique, on conservera toutes les étoiles avec $\sigma_t < 0.35$ et $8.5 < t < 14.5$, de façon à obtenir à la fois des étoiles distantes, et en même temps un échantillon suffisamment important.

On n'appliquera aucune censure sur les parallaxes trigonométriques.

L'échelle de hauteur des étoiles, essentiellement de type B et A, est d'environ $h_Z \approx 100$ pc (parsec).

L'intervalle de confiance à 95% est [-3.77,4.01].

3 Mise en œuvre numérique

Le découpage du projet est indiqué ci-dessous à titre indicatif. Chaque module est à faire par un binôme. On utilisera par exemple RCS pour gérer la configuration des logiciels.

Compte-tenu du temps disponible pour le projet, il est probable qu'il faut donner des priorités : on a écrit [[entre crochets]] les travaux à ne pas effectuer en priorité.

3.1 Modules

prog. principal : enchaînement des opérations,
structures de données,
initialisations,
prototypes,
Makefile,
[[aide en ligne]]

Il y aura un rôle de supervisation des autres modules, et éventuellement d'aide ; par exemple préparer des programmes principaux avec des jeux de données et des Makefile pour tester les autres modules avant leur intégration dans le programme.

E/S : lecture des paramètres,
lecture des données,
écriture des résultats,
[[écriture des résidus]]

graphiques : vraisemblance versus les paramètres
résidus versus b, versus t,
[[histogrammes des résidus normalisés avec gaussienne,]]

Utiliser xmgr en lui donnant un fichier de commande, ou bien lier avec plplot (analogue de pgplot, en `C`) comme librairie

calcul : implantation des densités de probabilité,
[[des dérivées de la vraisemblance]]

fonctions : vraisemblance,
[[résidus et points aberrants]]
utilitaires

On essaiera de simplifier les expressions intervenant dans l'eq. 2

numérique : intégration de fonctions,
maximisation de fonctions,
[[inversion de matrice,]]

Utiliser les Numerical Recipes en C comme librairie. Quelques problèmes pouvant être rencontrés : les bornes d'intégration ($\pi = 0$, cf eq. 8), la réentrance de l'intégration, la précision machine...

tests : tests,
contrôle qualité,
[[simulations]]

Le groupe en charge de cette partie devra juger de façon indépendante si le programme répond bien aux spécifications. On préparera des jeux de données, avec les résultats attendus. Éventuellement on préparera des programmes pour les tests unitaires. On s'attachera enfin à vérifier que le programme est facile à maintenir (lisible, modulaire, documenté). Utiliser les Numerical Recipes en C pour la génération de nombres pseudo-aléatoires.

3.2 Données fournies

– Fichier de configuration contenant : nom du fichier de donnés, nom du fichier résultat, nom des paramètres recherchés, valeur des paramètres en sortie, précision sur ces paramètres, nom des observables, censures sur les observables, leur valeur min et max pour la recherche par maximisation. Exemple :

```
Input_file=H30 #File containing data
Output_file=ajeter #File containing results
```

```
Name_param=OFP,KSP  #Name of all parameters
Init_param=-0.024,1.014 #Initial parameter values
Min_param=-1.,0.8 #Minimum parameter values
Max_param=1.,1.5 #Maximum parameter values
Sigma_param=0.047,0.03 #Parameter standard errors
Find_param=OFP,KSP #Name of parameters to find
Name_obs=PI0,SP0,MAG,SMV,LG2,BG2,IDF #Name of observables
Inf_censor=-0.1,0.,8.5,0.,0.,-90.,-9999 #Lower censorship
Sup_censor=1.,1.,14.5,5.,360.,90.,9999 #Upper censorship
Outlier=-3.77,4.01 #Confidence interval of std residuals
```

– Fichier avec une ligne par étoile, les données étant séparées par des tabulations et contenant les observables dans l'ordre indiqué par le fichier de configuration. La parallaxe (`PI0`) et son erreur (`SP0`) est donnée en mas (0.001"), le module de distance (`MAG`) et sa précision (`SMV`) en magnitude, les coordonnées galactiques (`LG2,BG2`) en degrés décimaux. `IDF` est un identificateur de l'étoile.

3.3 Déclaration de classes

On voit par exemple la structure d'une classe « étoile » : les données (parallaxe observée, précision sur celle-ci, etc) ; un flag pour indiquer si c'est un point aberrant ; la valeur de la vraisemblance individuelle (eq. 2) ; du résidu normalisé.

De même pour une classe « paramètre » : nom, valeur, précision, minimum, maximum, flag indiquant s'il est fixé ou non.

Et enfin pour une classe « échantillon » qui englobe les précédentes.

Références

[1] Arenou F. et al., 1995, *Astron. Astrophys.* **304**, 52

[2] Malmquist, K.G., 1936, *Meddel. Stockholm Obs.* **26**.

[3] Pelat D., *Bruits et signaux*, Cours de l'École Doctorale d'Astrophysique d'Île de France

[4] Press et al., *Numerical Recipes*, ed. Cambridge University Press (en C, ISBN 0-521-35465-X)