Estimation de paramètres par projection optimale dans un environnement local

Albert BIJAOUI, Alejandra RECIO-BLANCO, Patrick DE LAVERNY

Université de Nice Sophia Antipolis
UMR CNRS 6202 Cassiopée
Observatoire de la Côte d'Azur, BP 4229, 06304 Nice Cedex 04, France
bijaoui@oca.eu,arecio@oca.eu,laverny@oca.eu

Résumé – Dans le cadre de la mission Gaia de l'ESA, qui doit être lancée en 2012, les spectres de plusieurs millions d'étoiles seront observés. La théorie des atmosphères stellaires, associée au modèle instrumental, permet de construire des spectres théoriques très proches de la réalité. Ces modèles dépendent de plusieurs paramètres physiques qu'il s'agit d'estimer avec la meilleure précision possible. Pour cela on cherche à minimiser la distance observation - modèle sur une grille discrète. Un premier algorithme (MATISSE) a été développé en déterminant les paramètres par projection dans le cadre d'une multiregression linéaire locale. Dans un second algorithme (PEOPLE) les paramètres sont déterminés par une interpolation à noyau dans l'espace du signal. Les méthodes mises en oeuvre sont bien adaptées à la comparaison de nombreuses observations à une bibliothèque de modèles, dont le calcul est laborieux. Avec PEOPLE, l'utilisation du noyau d'Epanechnikov permet de relier l'environnement dans l'espace des paramètres à la corrélation des modèles. Grâce à ce résultat on peut affiner la stratégie de construction de la bibliothèque de modèles permettant l'apprentissage.

Abstract – The ESA Gaia mission will be launched in 2012. During this mission its RVS spectrograph will observe several millions of stellar spectra. The theory of stellar atmospheres, combined with the instrumental model, allows one to build libraries of synthetic spectra very similar to the observed ones. These models depend on different physical parameters. Our purpose is to get their best estimates from the observations, by minimizing the distance between the observation and the models, computed on a discrete grid. The MATISSE and PEOPLE algorithms perform this task by projecting the spectra on specific vectors, associated to a local environment. The methods are well adapted to the huge number of estimations to be done. The use of an Epanechnikov kernel in PEOPLE allows one to connect the environment in the parameter space to the model correlation. The model library can then be more easily sampled taking into account this property.

1 Observations, modèles et paramètres.

La mission Gaia de l'ESA doit être lancée en 2012 afin d'approfondir notre connaissance sur la formation, la constitution et l'évolution de notre Galaxie. Parmi les instruments à bord, le spectrographe RVS doit permettre l'acquisition de plusieurs millions de spectres stellaires, avec pour premier objectif la détermination des vitesses radiales des étoiles [5]. Les spectres contiennent aussi des informations essentielles sur la physique de l'atmosphère des étoiles et l'un des objectifs sera d'en extraire les paramètres physiques correspondants. Il s'agit essentiellement de la température effective à la surface, de la gravité de surface, de l'abondance relative des éléments chimiques par rapport à l'hydrogène. D'autres paramètres physiques, comme la proportion relative des éléments alpha, qui nous informent sur les échelles de temps de la formation stellaire, ou la vitesse de rotation de l'étoile contribuent significativement à la modélisation des spectres.

Pour cela les observations doivent être comparées à des spectres synthétiques calculés à partir de la théorie des atmosphères stellaires et du modèle d'observation, intégrant les caractéristiques instrumentales et les propriétés du bruit. La comparaison d'observations et de modèles obtenus par simulation est de nos jours une activité commune en astrophysique. Chaque modèle est caractérisé par un jeu de paramètres Θ. En tenant compte de la chaîne instrumentale, il se ramène à un signal S_Θ. Celui-ci doit être comparé aux données expérimentales O. Le bruit instrumental conduit à une probabilité conditionnelle $p(O/S_\Theta)$. Dans le cadre de cette communication, le bruit sera considéré comme étant blanc gaussien. Dans ces conditions, l'application du principe du maximum de vraisemblance conduit à estimer les paramètres du meilleur modèle associé à une observation en minimisant la distance entre O et S_Θ.

La méthode des moindres carrés a conduit à de très nombreux travaux [1]. Dans le cas de modèles analytiques non linéaires, les paramètres peuvent être déterminés par la méthode de Gauss-Newton [2], avec des corrections :

$$\delta\Theta = (J^T J)^{-1} J^T (O - S_\Theta), \qquad (1)$$

où J est le Jacobien $[\partial S(l,\Theta)/\partial\theta_i]$. Cet algorithme converge dans le cas convexe. Depuis les années 70 divers algorithmes, comme les algorithmes génétiques [3] ou le recuit simulé [4], ont été proposés pour forcer la convergence dans le cas non convexe.

Le spectrographe RVS permettra l'acquisition d'une dizaine de millions de spectres d'étoiles sur 971 éléments spectraux (spectrels). Une partie de l'analyse de ces données consistera à déterminer des paramètres physiques liés aux atmosphères de ces étoiles. Pour cela on dispose d'une bibliothèque de modèles associés à un échantillon de paramètres. Le calcul de ces modèles est très consommateur en temps de calcul. De plus, les modèles évoluent au fur et à mesure des progrès de l'astrophysique stellaire et de l'amélioration de la précision des nombreuses constantes associées. Ces modèles se présentent sous

la forme de tableaux de données, l'application directe de l'algorithme de Gauss-Newton ne serait alors possible qu'avec un échantillonnage suffisamment serré pour évaluer le Jacobien.

La recherche directe du minimum de distance entre l'observation et les différents modèles est souvent effectuée. Cela conduit à des calculs très fastidieux si le nombre de modèles est élevé. L'application de l'algorithme Nedler-Mead [6] permet d'accélérer la convergence sur la grille des paramètres échantillonnés [7]. Néanmoins, le pas de la grille limite la précision des résultats.

Nous avons proposé une autre voie, permettant d'interpoler autour d'un point de la grille. Cette approche est basée sur la détermination des paramètres par projection. Dans la méthode MATISSE [8], le paramètre est déterminé par régression multilinéaire, l'environnement étant déterminé par les points de la grille les plus proches. Dans la méthode PEOPLE l'utilisation de l'interpolation à noyau de Nadaraya-Watson [10] permet de faire le lien entre échantillonnage et corrélation [9].

2 La méthode MATISSE.

On considère un environnement autour d'un point de la grille. Les modèles sont considérés à moyenne nulle. On soustrait également la moyenne du spectre observé. Les modèles et les observations sont normalisés ensuite de manière à avoir une énergie égale à 1. Les paramètres des modèles sont aussi réduits à une moyenne nulle dans cet environnement. θ_i est estimé par régression multilinéaire, c'est-à-dire par projection :

$$\hat{\theta_i} = \sum_l B_i(l) O(l). \tag{2}$$

B_i est supposé être une combinaison linéaire des modèles $j \in (1, J)$ de l'environnement considéré :

$$B_i(l) = \sum_{j=1,J} \alpha_{ij} S_j(l). \tag{3}$$

Ce qui conduit à estimer le vecteur α_i par l'inversion du système :

$$\Theta_i = C\alpha_i, \tag{4}$$

où $C = [c_{jj'}]$ est la matrice de corrélation et Θ_i le vecteur des paramètres i des spectres. Pour une matrice inversible on obtient :

$$\alpha_i = C^{-1}\Theta_i, \tag{5}$$

ce qui conduit pour le vecteur d'apprentissage :

$$\hat{\Theta}_i = SB = S(S\alpha_i) = CC^{-1}\Theta_i = \Theta_i. \tag{6}$$

Généralement la matrice est mal conditionnée. Une inversion itérative du type Landweber [11] permet de régulariser l'inversion. Ceci est d'autant plus important que les données sont bruitées.

Compte tenu de la non linéarité des modèles par rapport aux paramètres on procède en deux étapes (Figure 1). Une estimation préliminaire des paramètres est réalisée par projection avec les vecteurs $B_\Theta^{(0)}$ calculés sur une sélection des spectres sur toute la grille. Puis on effectue la projection sur les vecteurs $B_\Theta^{(f)}$ construits avec les spectres de l'environnement des paramètres estimés. Quelques itérations (moins d'une dizaine) sur cette seconde étape peuvent parfois être nécessaires pour affiner l'estimation.

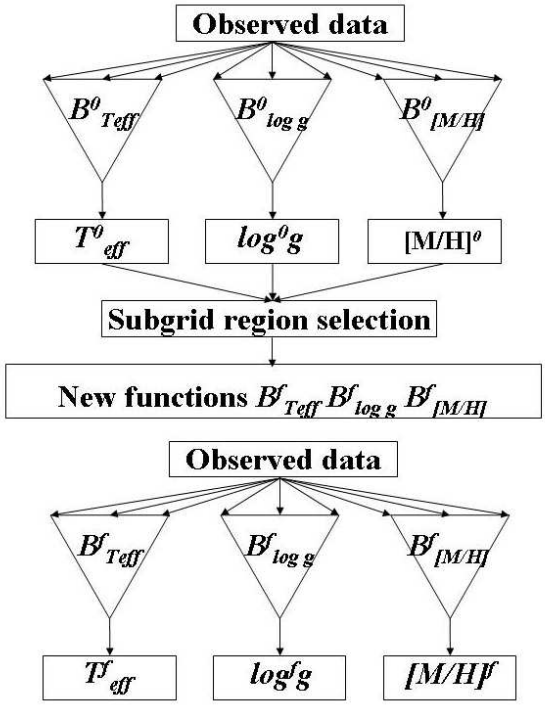

FIG. 1 – Schéma de l'algorithme MATISSE pour l'estimation des paramètres stellaires.

3 La méthode PEOPLE.

Le problème d'estimation peut aussi être perçu comme un problème d'interpolation. Dans l'espace du signal où chaque point a pour coordonnées les valeurs des spectrels, chaque modèle correspond à un point pour lequel les valeurs des paramètres sont connues. Pour le point correspondant à l'observation on applique la moyenne pondérée de Nadaraya-Watson [10] :

$$\hat{\theta}_i(O) = \frac{\sum_j K(\frac{|O-S_j|}{a})\overline{\theta}_{ij}}{\sum_j K(\frac{|O-S_j|}{a})}, \tag{7}$$

$K(d)$ est un noyau qui dépend de la distance d dans l'espace des spectrels. a est le paramètre d'échelle du noyau. L'application de cet algorithme avec les paramètres θ_{ij} ne restitue pas en général les valeurs aux noeuds de la grille. Les quantités $\overline{\theta}_{ij}$ doivent donc être telles qu'on doit retrouver les paramètres des modèles pour chacun d'entre eux soit :

$$\theta_{ij} = \frac{\sum_j K(\frac{|S_i-S_j|}{a})\overline{\theta}_{ij}}{\sum_j K(\frac{|S_i-S_j|}{a})}, \tag{8}$$

Considérons le noyau d'Epanechnikov [12] :

$$K(x) = \frac{3}{4}(1-x^2) \quad \text{si} \quad |x| \leq 1 \quad \text{autrement} \quad K(x) = 0. \tag{9}$$

En tenant compte de la normalisation des spectres, les poids K peuvent s'écrire :

$$w_{jj'} = H(c_{jj'} - c), \tag{10}$$

où c est un seuil de corrélation et $H(x)$ la fonction d'Heaviside. Le paramètre θ_i est donc estimé par la relation :

$$\hat{\theta}_i = \frac{\sum_j H(\sum_l O(l)S_j(l) - c)\overline{\theta}_{ij}}{\sum_j H(\sum_l O(l)S_j(l) - c)}. \tag{11}$$

Si on définit $e(O)$ comme l'ensemble des spectres S_j tels que $c_j > c$, la relation (11) peut s'écrire :

$$\hat{\theta}_i = \frac{\sum_l O(l)[\sum_{j\in e(O)} S_j(l)\overline{\theta}_{ij}] - c\sum_{j\in e(O)} \overline{\theta}_{ij}}{W}, \quad (12)$$

où $W = \sum_{j\in e(O)} W_j$. Il en résulte :

$$\hat{\theta}_i = \sum_l O(l) B_i^e(l) - \theta_i^e \quad (13)$$

où $B_i^e(l) = \frac{1}{W}\sum_{j\in e(O)} S_j(l)\overline{\theta}_{ij}$ et $\hat{\theta}_i^e = \frac{c\sum_{j\in e(x)} \overline{\theta}_{ij}}{W}$.

Dans le cas de MATISSE, l'environnement est défini par l'échantillonnage de la grille. Avec PEOPLE l'environnement résulte du seuil de corrélation. On retrouve que l'estimation se fait par projection sur des vecteurs spécifiques à chaque paramètre dans un environnement lié au point de la grille considérée. Toutefois on évite le calcul des vecteurs correspondants car l'estimation des paramètres peut être effectuée directement après inversion de la matrice de corrélation seuillée.

4 Tests sur des spectres stellaires.

Un ensemble de 1386 spectres ayant 971 spectrels ont été calculés. 3 paramètres physiques varient : la température effective de l'étoile T, sa gravité de surface $\log g$ et un indice dit de métallicité $[M/H]$, relatif à l'abondance des éléments plus lourds que l'hélium. Sur la Figure 2 on a représenté les spectres relatifs aux paramètres solaires avec une petite variation pour chacun d'entre eux.

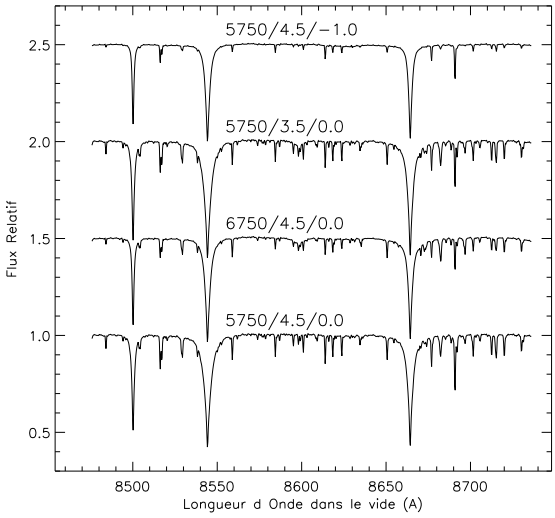

FIG. 2 – Spectres synthétiques correspondant aux paramètres solaires ($T = 5750$, $g = 4,5$, $[M/H] = 0$) et à de petites variations.

Les fonctions B calculées dans MATISSE sont tracées sur la Figure 3 pour le point de la grille correspondant au modèle solaire. Elles montrent les poids utilisés pour estimer chacun des paramètres physiques. Ces fonctions sont presque orthogonales.

Pour PEOPLE différents seuils de corrélation ont été examinés. La précision augmente avec le seuil. Le choix est borné par la nécessité d'avoir pour chaque observation un environnement suffisamment peuplé. Sur la Figure 4 on a représenté

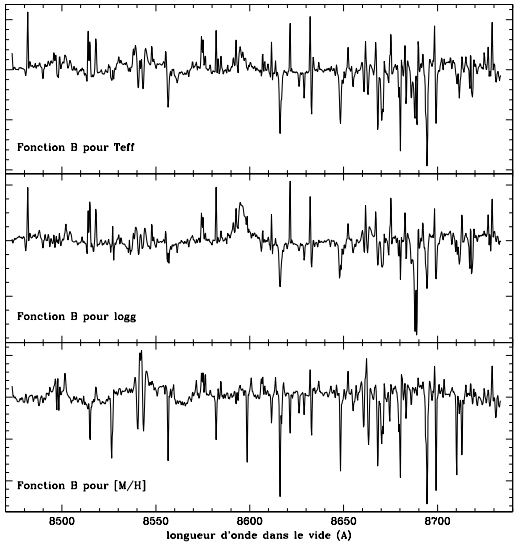

FIG. 3 – Fonctions B permettant l'estimation des paramètres pour des modèles proches du modèle solaire.

la matrice de corrélation inter-spectres seuillée à $0,995$. Cette matrice est très peu peuplée, mais pour chaque spectre il y a suffisamment de spectres pour une bonne estimation des paramètres physiques dans son environnement.

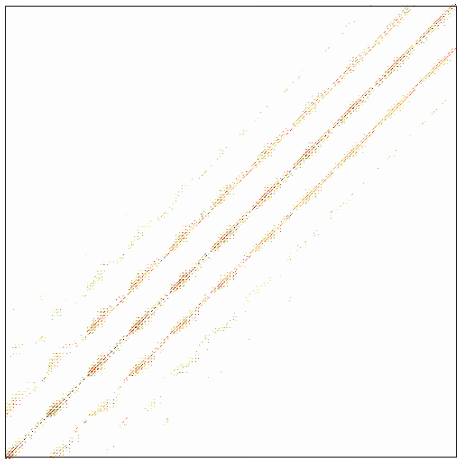

FIG. 4 – Matrice d'intercorrélation des spectres seuillée à $0,995$. La structure en bandes de cette matrice est liée à l'échantillonnage des paramètres.

Dans la Table 1 les résultats obtenus avec différentes expériences sont présentés :

PEOPLE Le seuil de corrélation choisi est de $0,995$. Il permet d'avoir une matrice suffisamment peuplée pour déterminer correctement les paramètres pour l'ensemble de la grille.

MATISSE B0 Toute la dynamique de la grille est utilisée pour calculer les fonctions B. L'estimation est donc très sensible aux non linéarités.

MATISSE B0+Bf Après localisation, une nouvelle projection est faite pour obtenir les valeurs correctes.

MATISSE avec rejet L'algorithme précédent fonctionne parfaitement sur 99% de la grille. On rejette les spectres tels

Méthode	T-max	$\sigma(T)$	$\log g$-max	$\sigma(\log g)$	μ-max	$\sigma(\mu)$
PEOPLE $c = 0,995$	122	13	0.42	0.06	0.17	0.03
MATISSE B0	1402	451	2.95	0.69	1.32	0.26
MATISSE B0+Bf	1709	104	6.	0.19	1.34	0.08
B0+Bf 99% étoiles	0.	0.	0.	0.	0.	0.

TAB. 1 – Résultats expérimentaux sur la grille de spectres.

que la distance entre le spectre observé et le spectre restauré est trop grande. Ces spectres sont situés au bord de la grille. Pour ces spectres l'algorithme MATISSE diverge en raison de la non convexité de la fonction distance.

La grille d'échantillonnage a un pas en température de 250 degrés, de $0,5$ dex en logarithme de gravité de surface, et de $0,5$ dex en indice de métallicité. Il apparaît clairement que la méthode PEOPLE permet de retrouver avec une très grande précision les valeurs de l'échantillon d'apprentissage en une seule passe. Pour la méthode MATISSE, deux passes sont nécessaires. Pour quelques spectres en bord de grille la restitution est très mauvaise, mais pour 99% d'entre entre eux les paramètres de la grille d'apprentissage sont parfaitement restitués.

Dans notre exposé le bruit des mesures n'est introduit qu'initialement pour justifier la recherche d'une distance euclidienne minimum. Nous avons examiné l'effet du rapport signal sur bruit sur la précision des mesures [8]. Nous avons montré que les estimateurs étaient bien consistants, avec des variances expérimentales conformes aux prédictions théoriques (Figure 5).

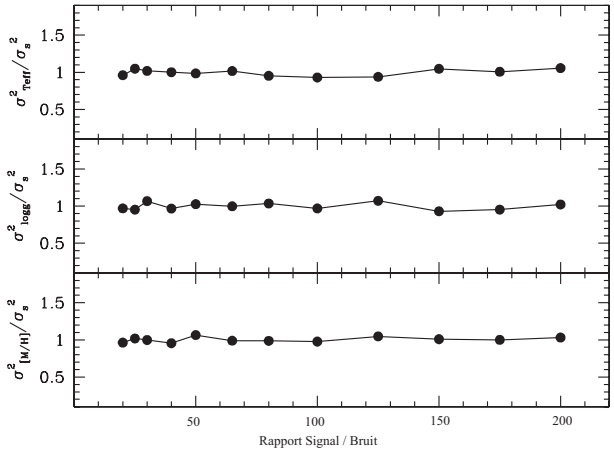

FIG. 5 – Variation du rapport entre la variance des mesures et la variance théorique en fonction du rapport entre l'amplitude du continu du spectre et l'écart-type du bruit. L'étude portait sur trois paramètres physiques, la température effective et la gravité de surface et l'abondance des éléments plus lourds que l'hélium.

5 Conclusion.

La méthode MATISSE est bien adaptée à l'ajustement de modèles quand il s'agit de grands nombres d'observations, avec des modèles non exprimables de manière analytique et ayant peu de paramètres. Avec la méthode PEOPLE on relie la qualité de l'estimation au nombre de modèles fortement corrélés à l'observation. Ceci permet de déduire une stratégie optimale d'échantillonnage de la grille compte tenu du but poursuivi.

Grâce à la méthode PEOPLE on peut localiser avec une bonne précision les bons paramètres en une passe. Elle peut se substituer à la première phase $B0$ de la méthode MATISSE, la phase Bf conduisant ensuite à la meilleure estimation.

Références

[1] J. Wolberg, *Data Analysis Using the Method of Least Squares : Extracting the Most Information from Experiments*, Springer, Francfort, 2005.

[2] A. Björck, *Numerical methods for least squares problems*, SIAM, Philadelphia, p.260, 1996.

[3] J.H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.

[4] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, M.P., *Optimization by Simulated Annealing*, Science, 200, 671-680, 1983.

[5] M.I. Wilkinson et 40 auteurs, *Spectroscopic survey of the Galaxy with Gaia - II. The expected science yield from the Radial Velocity Spectrometer*, Mon. Not. Royal Astro. Soc. 359, 1306-1335, 2005.

[6] J. Nelder, R. Mead, *A simplex method for function minimization*, Computer J., 7, 308-313, 1965.

[7] C. Allende Prieto, *Stellar atmospheric parameters : the four-step program and Gaia's radial velocity spectrometer*, Classification and discovery in large astronomical surveys, ed. C.A.L. Bailer-Jones, 47-53, AIP conf. 1082, 2008.

[8] A. Recio-Blanco, A. Bijaoui, P. de Laverny, *Automated derivation of stellar atmospheric parameters and chemical abundances : the MATISSE algorithm*, Mon. Not. Royal Astro. Soc. 370, 141-150, 2006.

[9] A. Bijaoui, A. Recio-Blanco, P.de Laverny, *Parameter Estimation from an Optimal Projection in a Local Environment*, Classification and discovery in large astronomical surveys, ed. C.A.L. Bailer-Jones, 54-60, AIP conf. 1082, 2008.

[10] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, p.35, Springer, 2001.

[11] L. Landweber, *An iterative formula for Fredholm integral equations of the first kind.*, Am. J. Math., 73, 615-624, 1951.

[12] V.A. Epanechnikov, *Theory Probab. Appl.*, 14, 153-158, 1969.