Un Arbre de Décision Oblique pour l'Estimation des Paramètres sur une Grille de Modèles

Albert BIJAOUI, Georges KORDOPATIS, Alejandra RECIO-BLANCO, Patrick DE LAVERNY, Christophe ORDENOVIC

Laboratoire Cassiopée UMR CNRS 6202
Observatoire de la Côte d'Azur, BP 4229, 06304 Nice Cedex 04, France
Albert.Bijaoui@oca.eu, gkordo@oca.eu, arecio@oca.eu
laverny@oca.eu, Christophe.Ordenovic@oca.eu

Résumé – L'analyse de spectres stellaires obtenus dans le cadre de grands relevés nécessite de déterminer rapidement et avec la meilleure précision possible les paramètres physiques associés à un ensemble très vaste d'observations bruitées à partir d'une grille de modèles. Pour résoudre ce problème, nous proposons une méthode basée sur la construction d'un arbre de décision oblique. Le sous-ensemble de modèles associés à chaque noeud est partitionné en deux en comparant le coefficient de projection de chaque modèle sur un vecteur à la médiane des coefficients de projection pour l'ensemble des modèles. Nous proposons une construction particulière du vecteur de projection permettant une meilleure exploration de l'espace des paramètres. L'expérimentation présentée sur notre grille de modèles montre la très grande efficacité de la méthode pour identifier rapidement le meilleur modèle, même pour de faibles rapports signal à bruit.

Abstract – The analysis of stellar spectra obtained in the context of large surveys requires to quickly determine with the best possible accuracy their physical parameters. These parameters are associated to a wide set of noisy observations from a grid of models. For that purpose, we propose a method based on the construction of an oblique decision tree. Each model of the subset attached to a given node is projected on a specific vector. The subset is then partitioned in two parts by comparing the projection coefficient of each model to the coefficient median of the subset. We also introduce a particular construction of the projection vector for a better exploration of the parameter space. The experiments performed for our grid of models show a great efficiency of the method to quickly identify the best model, even at low signal to noise ratios.

1 Estimation des paramètres de spectres stellaires.

Le spectrographe RVS de la mission spatiale Gaia de l'ESA permettra l'acquisition de quelques dizaines de millions de spectres d'étoiles sur 971 éléments spectraux [1]. Une partie de l'analyse de ces données consistera à déterminer des paramètres physiques (au moins la température effective, la gravité de surface et la métallicité moyenne) liés aux atmosphères de ces étoiles. Les modèles d'atmosphère évoluent au fur et à mesure des progrès de l'astrophysique stellaire et de l'amélioration de la précision des nombreuses constantes associées. Comme leur calcul est très consommateur en temps machine, l'analyse des spectres ne peut se faire que par une comparaison rapide à des modèles pré-calculés sur une grille associée à un échantillon de paramètres. Sur la figure 1 nous avons représenté les spectres simulés aux caractéristiques de l'instrument RVS pour 4 jeux de paramètres atmosphériques. Le premier correspondrait au spectre d'une étoile de type solaire, les trois autres à des spectres d'étoiles ayant l'un des paramètres atmosphériques légèrement différent, la variation correspondant au pas d'échantillonnage de la grille de modèles.

Sous l'hypothèse d'un bruit gaussien, l'estimation s'effectue dans le cadre de la méthode des moindres carrés. Le problème posé a les spécificités suivantes :

- Les modèles ne sont pas définis sous forme d'une relation analytique. Pour obtenir une approximation pour un jeu de paramètres quelconques une interpolation multi dimensionnelle est nécessaire.
- Le calcul d'un modèle est élevé en temps de calcul. Les pas d'échantillonnage ont été choisis dans le cadre d'un compromis entre le coût global du calcul de ces modèles et la précision qu'on peut atteindre avec les estimateurs choisis.
- En raison des non-linéarités, la fonction distance d'un spectre simulé par rapport aux autres spectres n'est pas toujours convexe. Dans certains cas de dégénérescence, des jeux de paramètres très différents peuvent conduire à des spectres simulés très semblables. Il est donc essentiel que l'algorithme choisi conduise à éviter de tomber dans un minimum secondaire.
- L'instrument fournira des dizaines de millions de spectres. L'algorithme choisi devra être efficace, robuste et rapide.

La recherche directe du minimum de distance entre le spectre observé et les spectres de la grille de modèles peut s'effectuer par balayage complet de la grille. Cette opération permet d'être

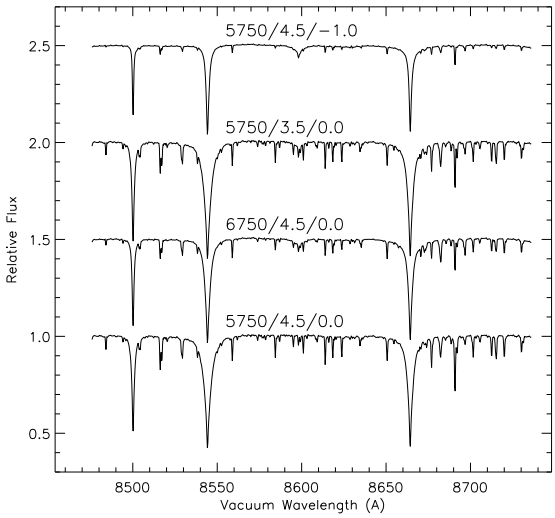

FIG. 1 – Spectres synthétiques correspondant aux paramètres solaires ($T = 5750$, $g = 4,5$, $[M/H] = 0$) et à de petites variations.

sûr de déterminer le minimum global, mais cela conduit à des calculs très fastidieux. En outre, la précision est limitée par le pas de la grille. Nous avons examiné plusieurs méthodes de détermination moins gourmandes comme les algorithmes de Nelder-Mead par déformation d'un simplex [2] [3], de Gauss-Newton lié à une linéarisation itérative [4] [5], MATISSE basée une une regression linéaire locale [6] ou PEOPLE utilisant l'interpolation à noyau de Nadaraya-Watson [7] [8]. Toutes ces méthodes ne garantissent pas la convergence vers le minimum absolu. L'introduction de variantes stochastiques a permis d'améliorer la convergence au prix d'un coût de calcul supérieur à celui du balayage systématique de la grille.

Dans un espace à faible dimension, la recherche du minimum de distance s'effectue rapidement grâce à une décomposition de type arbre kd [9]. Dans un espace de grande dimension, la recherche devient moins efficace que le balayage systématique de la grille [10]. Nous présentons une méthode de reconnaissance basée sur un arbre de décision de type oblique [11], mais qui, comme un arbre kd, est lié à une décomposition dyadique de l'espace.

2 Arbre de décision oblique de type kd.

Structure de l'arbre. Soit $\{S\}$ l'ensemble des spectres de la grille. L'apprentissage conduit à introduire un arbre de décision binaire formé de nœuds n à chacun desquels est associé un sous-ensemble de spectres $\{S_n\}$. Nous cherchons à séparer en deux de manière aussi égale que possible $\{S_n\}$ afin que les sous-ensembles résultants $\{S_n^{(1)}\}$ et $\{S_n^{(2)}\}$ soient aussi distants que possible.

Pour effectuer la séparation, nous projetons les spectres de $\{S_n\}$ sur un vecteur V_n, obtenant ainsi n coefficients $\{c_n\}$. Comme pour l'arbre kd, nous considérons ensuite la médiane γ_n de ces coefficients. On partitionne $\{S_n\}$ en deux sous-ensembles $\{S_n^{(1)}\}$, de N_1 spectres et $\{S_n^{(2)}\}$ de N_2 spectres en comparant chaque coefficient de projection à la médiane. N_1 et N_2 sont alors identiques, à une unité près. En itérant, on obtient, en $N\log_2(N)$ opérations, un arbre de décision binaire de $\log_2(N)$ niveaux.

Règle de minimisation de l'énergie par nœud. L'arbre est complètement défini par le choix des vecteurs de projection. Pour l'arbre kd, on choisit l'une des composantes du vecteur, mais avec 971 pixels toute l'information sur les spectres ne peut être exploitée. Il est naturel de choisir la combinaison linéaire conduisant à la plus grande variance du coefficient de projection. Il s'agit alors de la première composante principale de la matrice de covariance des spectres. Cela nécessite $N/2$ diagonalisations de matrice, ce qui devient fastidieux pour une grille de plusieurs milliers de modèles.

Après plusieurs expérimentations, nous avons introduit un critère basé sur la minimisation de la somme de l'énergie associée aux deux nouveaux nœuds. Soit $\overline{S}_n^{(}1)$ le spectre moyen de $\{S_n^{(1)}\}$ et soit $\overline{S}_n^{(}2)$ celui de $\{S_n^{(2)}\}$. L'énergie de chaque ensemble i est :

$$E_i = \sum_n |S_n - \overline{S}_n^{(}i)|^2 \tag{1}$$

avec $S_n \in \{S_n^{(i)}\}$.

Il est aisé de montrer (théorème de Huyghens) que l'énergie associée aux deux nouveaux nœuds est :

$$E = E_0 - \sum_{i=1,2} N_i \sum_n |\overline{S}_n^{(}i)|^2. \tag{2}$$

où E_0 correspond à l'énergie au nœud n. Nous allons chercher à minimiser E, ce qui correspond à maximiser :

$$F = \sum_{i=1,2} N_i \sum_n |\overline{S}_n^{(}i)|^2. \tag{3}$$

Pour calculer F, il faut partir d'un vecteur, effectuer les projections et calculer la médiane des coefficients résultants. On calcule alors le spectre moyen de chaque sous-ensemble, puis le carré de la distance de chaque spectre par rapport à sa moyenne et on somme le tout.

Le vecteur des différences. La détermination du vecteur de projection qui minimise F ne peut donc pas résulter d'une simple dérivation. La procédure que nous proposons est une heuristique qui est basée sur la variation de F associée à la permutation de deux spectres T_1 et T_2 entre les sous-ensembles $\{S_n^{(1)}\}$ et $\{S_n^{(2)}\}$. On montre que la variation d'énergie est :

$$\begin{aligned}\Delta F = &((-2+\frac{1}{N_1})E_1 + (-2+\frac{1}{N_2})E_2 + \\ &2(T_2-T_1).(\frac{N_2-1}{N_2}\overline{S_n}^{(2)} - \frac{N_1-1}{N_1}\overline{S_n}^{(1)}) + \\ &(\frac{1}{N_1}+\frac{1}{N_2})|T_2-T_1|^2\end{aligned} \tag{4}$$

Les deux premières quantités sont indépendantes des points échangés. Quant à la dernière, elle ne dépend pas des centres.

La variation est donc liée au produit scalaire :

$$G = (T_2 - T1).(\frac{N_2 - 1}{N_2}\overline{S_n}^{(2)} - \frac{N_1 - 1}{N_1}\overline{S_n}^{(1)}). \quad (5)$$

Comme N_1 et N_2 diffèrent très peu, on peut voir que G fait apparaître le vecteur des différences. Ceci nous a amené à tester l'utilisation de ce vecteur pour partitionner le sous-ensemble associé à un nœud. Comme ce vecteur dépend de la partition effectuée, une méthode itérative de construction est nécessaire. La construction de l'arbre de décision ainsi défini, est très rapide.

L'identification d'un spectre de la grille s'effectue en descendant l'arbre en $\log_2(N)$ projections. À chaque nœud, on choisit le bon embranchement en déterminant le coefficient de projection sur le vecteur V_n qui lui est associé.

3 Reconnaissance en présence de bruit.

La reconnaissance des spectres de la grille est très rapide et parfaite en l'absence de bruit. Par contre, l'algorithme brut s'est révélé très sensible au bruit. Si le bon parcours n'est pas choisi dans les premiers niveaux, l'algorithme peut fournir un spectre très différent du spectre réel.

Pour être assuré de choisir à la fin le meilleur modèle, nous procédons de manière probabiliste. Le choix de la branche à chaque nœud est effectué en affectant un poids w_n à la la branche choisie. w_n est calculé avec une fonction d'activation basée sur le noyau d'Epanechnikov [12], correspondant à une parabole tronquée. c_n désigne le coefficient de projection du spectre observé sur le vecteur associé au nœud. γ_n est la médiane des coefficients de projection du sous-ensemble des modèles de ce nœud. Une seule branche est choisie si l'écart entre c_n et γ_n est statistiquement trop élevé. Dans le cas contraire, les deux branches sont choisies avec des poids complémentaires w_n et $1 - w_n$.

À chaque noeud de l'arbre, on attribue un poids global p_n nul, sauf pour la racine. À chaque niveau de l'arbre, on examine les nœuds associés à un poids non nul. On calcule les poids associés aux deux branches correspondantes (w_n et $1 - w_n$) et on les multiplie par p_n pour obtenir les poids globaux associés aux deux nouveaux noeuds. Si w_n n'est pas nul, on augmente d'une unité le nombre de noeuds à examiner au niveau suivant. Ce nombre est un peu réduit en éliminant les modèles ayant un poids devenu négligeable.

En fin d'exploration de l'arbre, plusieurs feuilles peuvent ainsi être identifiées comme candidates pour être le plus proche voisin. Cet ensemble dépend d'un paramètre d'échelle de la fonction d'activation. S'il est trop petit, on n'identifie que très peu de candidats. En cas de bruit, il est clair qu'en raison de la dispersion de c_n on peut parfaitement rater le plus proche voisin. Si le paramètre est trop grand, le nombre de noeuds à analyser à chaque niveau augmente exponentiellement, rendant l'algorithme inefficace.

Après exploration de l'arbre, on calcule la distance du spectre observé O aux J modèles retenus S_j, $j \in (1, J)$. Les paramètres du spectre $\hat{\theta}$ sont déterminés par la relation de Nadaraya-Watson [7] :

$$\hat{\theta} = \frac{\sum_j K(\frac{|O-S_j|}{a})\theta_j}{\sum_j K(\frac{|O-S_j|}{a})} \quad (6)$$

Nous avons utilisé initialement le noyau d'Epanechnikov :

$$K(x) = \frac{3}{4}(1 - x^2) \qquad |x| < 1 \quad (7)$$

$K(x)$ est nul à l'extérieur de cet intervalle. a est un paramètre d'échelle déduit de la répartition des distances aux spectres retenus. Un gain en précision a été obtenu avec un noyau de plus fort exposant ($k \approx 32$) :

$$K(x) = \alpha(1 - x^2)^k \qquad |x| < 1. \quad (8)$$

Des résultats de précision similaire ont été obtenus avec un noyau Gaussien, avec un paramètre d'échelle adapté.

4 Application.

Cet algorithme, baptisé DEGAS (DEcision tree alGorithm for AStrophysics), a été appliqué pour la paramétrisation de spectres obtenus avec le spectrographe FLAMES de l'ESO, pour une résolution et pour un domaine de longueurs d'onde proches de ceux du RVS de Gaia [13]. Des tests ont été effectués sur un catalogue de 20000 spectres synthétiques avec quatre niveaux de bruit. L'arbre a été apprenti avec 2905 spectres échantillonnés régulièrement en température effective (Teff) et en gravité de surface (log(g)) de surface, mais avec un pas variable en métallicité ([M/H]). Les valeurs 100, 50, 20, 10 du rapport signal à bruit (RSB), défini comme le rapport du continu du spectre sur l'écart-type du bruit, ont été considérées pour la reconnaissance. Nous avons comparé les résultats de DEGAS avec ceux obtenus avec l'algorithme MATISSE basé sur une régression linéaire locale.

Ces expériences ont montré que le spectre synthétique le plus ressemblant au spectre bruité était bien retrouvé par MATISSE et DEGAS, tant que le niveau de bruit restait suffisamment faible (RSB $\geq$ 50). Dans ce cas, les erreurs finales sur les paramètres atmosphériques stellaires obtenues par les deux méthodes sont très faibles et semblables, de l'ordre de 60K pour Teff et de 0.1dex pour log(g) et pour [M/H]. Ces valeurs sont inférieures d'un facteur de près de 4 au pas d'échantillonnage des paramètres dans la grille. Elles correspondent à une précision bien supérieure à celle nécessaire pour l'interprétation astrophysique des résultats.

Lorsque le RSB diminue, l'algorithme DEGAS reste plus robuste que MATISSE, conduisant à des erreurs nettement inférieures. Ainsi, à RSB$\approx$ 10, les précisions obtenues par MATISSE pour des étoiles typiques du disque mince de notre Galaxie (naines, riches en métaux) sont respectivement de 382K, 0.69dex et 0.34dex pour la Teff, le logg et [M/H]. Ces erreurs ne sont que de 278K, 0.48dex et 0.21dex dans le cas de DEGAS, soit une amélioration de 30% pour la gravité et la température effective, et de 40% dans le cas de la métallicité globale.

5 Conclusion.

Nous avons présenté un nouvel algorithme de détermination de paramètres de modèle basé sur la classification, comme pour l'algorithme CART de Breiman *et al.* [14], mais adapté à un espace signal de grande dimension. Cet algorithme est basé sur la construction d'un arbre de décision oblique.

Contrairement aux algorithmes d'ajustement basés sur l'optimisation, DEGAS permet d'avoir une exploration complète de l'espace des paramètres échantillonnés. Ainsi, après construction de l'arbre, on détecte facilement les nœuds pour lesquels la variance associée aux paramètres est trop grande. Ceci permet d'identifier les spectres qui se ressemblent beaucoup, alors que les paramètres physiques sont très différents.

DEGAS est avant tout un algorithme de détermination rapide des plus proches voisins dans un espace de grande dimension. L'estimation des paramètres est ensuite effectuée avec la relation d'interpolation de Nadaraya-Watson. Ceci introduit un biais qu'on peut corriger par une méthode inverse [8]. Ceci n'a d'intérêt que pour de hauts RSB. Il suffit alors d'utiliser une méthode locale, comme l'algorithme de Gauss-Newton ou MATISSE, pour raffiner la mesure. DEGAS aura permis alors d'identifier rapidement la région du minimum global, ce qui en soi reste l'un des problèmes les plus délicats de l'ajustement de modèles.

Les codes de MATISSE et de DEGAS ont été développés en F90 et ensuite adaptés en Java. Ils sont maintenant intégrés dans la chaîne de traitement des données Gaia au CNES, constituant le cœur de l'algorithme *Generalized Stellar Parametrizer - spectroscopy* pour être exploités dans le cadre de la mission.

Références

[1] M.I. Wilkinson and 40 authors, *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.

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

[3] 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.

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

[5] C.A.L. Bailer-Jones, The ILIUM forward modelling algorithm for multivariate parameter estimation and its application to derive stellar parameters from Gaia spectrophotometry, Mon. Not. Royal Astro. Soc., 403,(2010), 96-116

[6] 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.

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

[8] 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.

[9] H. Samet. *The Design and Analysis of Spatial Data Structures*, pages 66-80, Addison-Wesley, Reading, MA, 1990.

[10] J. E. Goodman, J. O'Rourke and P. Indyk (Ed.) *Nearest neighbors in high-dimensional spaces*. dans *Handbook of Discrete and Computational Geometry*, chap. 29, 2ième ed., CRC Press, 2004.

[11] R. White, Astronomical applications of oblique decision trees, Classification and parametrization of unresolved galaxies with Gaia, Classification and discovery in large astronomical surveys, AIP Conference Proceedings, 1082, (2008), 37-43.

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

[13] G. Kordopatis, A. Recio-Blanco, P. de Laverny, A. Bijaoui, V. Hill, G. Gilmore,R.F.G. Wyse, and C. Ordenovic. *Automatic stellar spectra parametrisation in the IR Ca II triplet region*. soumis Astronomy & Astrophysics, 2011.

[14] L. Breiman, J.H. Friedman, , R. A. Olshen, C. J. Stone, Classification and regression trees, Monterey, CA : Wadsworth & Brooks/Cole Advanced Books & Software, 1984.