

# *Sliced Inverse Regression* pour la détermination des paramètres stellaires fondamentaux

Victor WATSON<sup>1,2</sup>, Jean-François TROUILHET<sup>1,2</sup>, Frédéric PALETOU<sup>1,2</sup> Marwan GEBRAN<sup>3</sup>

<sup>1</sup>Université de Toulouse, UPS-Observatoire Midi-Pyrénées, Irap, Toulouse, France

<sup>2</sup>CNRS, Institut de Recherche en Astrophysique et Planétologie, 14 av. E. Belin, 31400 Toulouse, France

<sup>3</sup>Department of Physics and Astronomy, Notre Dame University-Louaize, PO Box 72, Zouk Mikaël, Lebanon

victor.watson@irap.omp.eu, jean-francois.trouilhet@irap.omp.eu,  
frederic.paletou@irap.omp.eu, mgebran@ndu.edu.lb

**Résumé** – Nous voulons déterminer la valeur d’un paramètre à partir d’un grand vecteur de données associé, sans connaître la fonction permettant de passer d’un espace à l’autre. La méthode *Sliced Inverse Regression* (SIR) permet de projeter un vecteur de données vers un sous-espace qui est cohérent avec la variation du paramètre étudié. Nous proposerons, sur la base du sous-espace obtenu grâce à SIR, une méthode pour estimer de la manière la plus précise possible la valeur que prend le paramètre auquel on s’intéresse.

**Abstract** – We aim at finding the value of an explanatory variable, through its expression in a large data vector, without knowing the link function between the explanatory variable and the data-space. Sliced Inverse Regression (SIR) method, allows the projection of a data-vector onto a subspace consistent with the explanatory variable variation. We suggest a method based on the SIR subspace, that gives the most efficient estimation of an unknown explanatory variable.

## 1 Introduction

L’évolution des détecteurs et de nos capacités de stockage de l’information font que la collecte des données astrophysiques est toujours plus massive. Des projets comme Gaia [1] ou encore le LSST [2] en sont le parfait exemple. Les bases de données ainsi constituées seront alors utilisées pour extraire le maximum d’informations quant à la caractérisation des objets observés. Ici, nous nous intéresserons à la question de la détermination des paramètres stellaires fondamentaux : température effective, gravité de surface et métallicité (ou composition chimique globale) d’une étoile à partir de la mesure de son spectre électromagnétique dans le visible. Cette détermination doit alors être faite sans connaissance d’un lien fonctionnel entre paramètres et données, mais en ayant à disposition soit une base de données de spectres déjà caractérisés, soit des spectres synthétiques. Il devient donc intéressant de mettre en œuvre des méthodes statistiques qui, à partir de telles bases de données, peuvent automatiquement déterminer les quantités physiques désirées sans que la relation entre ceux-ci et les données ne soit connue. SIR [3], avec une application encore très confidentielle dans un contexte astrophysique [4], permet d’obtenir une représentation des données cohérente avec le paramètre sans connaissance de la fonction qui les lie. Nous chercherons donc la manière la plus efficace d’utiliser les composantes du sous-espace (que nous appellerons "directions") obtenues grâce à SIR.

## 2 Sliced Inverse Regression

SIR recherche le sous-espace de représentation d’un vecteur de données  $x$  qui explique au mieux la valeur d’un para-

mètre  $y$  associé. Pour construire ce sous-espace à partir d’une base de données d’objets déjà caractérisés, on commencera par diviser l’espace des données en  $H$  tranches (sans recouvrement et contenant chacune le même nombre d’échantillons) suivant l’évolution du paramètre  $y$  auquel on s’intéresse. Ainsi les échantillons de la base de données sont regroupés en tranches dans lesquelles la valeur du paramètre  $y$  varie peu. Appliquer SIR permet de construire le sous-espace qui maximise la variance entre les tranches tout en gardant une variance globale réduite. De fait, on peut y voir un parallèle avec le fonctionnement de l’analyse factorielle discriminante [5] dans un contexte qui aurait été un contexte de classification.

### 2.1 Principe

Soit  $\mathbf{X}$  la matrice des données dont chaque ligne est un échantillon associé à une valeur du paramètre  $y$ . Nous appellerons  $\mathbf{X}_h$  la matrice des éléments de  $\mathbf{X}$  qui appartiennent à la tranche  $h$ . Pour chaque tranche nous calculerons la moyenne  $\bar{m}_h$  des lignes de  $\mathbf{X}_h$ .

La matrice  $\mathbf{X}_H$  est la matrice des barycentres des tranches, dont chaque ligne est un des vecteur  $\bar{m}_h$ .

$\Gamma$  la matrice de variance-covariance inter-tranches :

$$\Gamma = \frac{1}{H} (\mathbf{X}_H - \bar{\mathbf{X}}_H)^T (\mathbf{X}_H - \bar{\mathbf{X}}_H), \quad (1)$$

et  $\Sigma$  la matrice de variance-covariance totale :

$$\Sigma = \frac{1}{N} (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}), \quad (2)$$

où  $N$  est le nombre d’échantillons de la base de donnée servant à construire le sous-espace,  $\bar{\mathbf{X}}$  est la moyenne des lignes de  $\mathbf{X}$  et

$\bar{\mathbf{X}}_H$  est la moyenne des lignes de  $\mathbf{X}_H$ . Les directions décrivant le sous-espace que l'on recherche sont les solutions de :

$$(\Sigma^{-1}\Gamma - \lambda \mathbf{I}_M)u = 0. \quad (3)$$

Les solutions de l'équation 3 correspondent aux vecteurs propres  $u$  associés aux plus grandes valeurs propres  $\lambda$  de  $\Sigma^{-1}\Gamma$ .  $M$  est la dimension de l'espace de départ.

Un point critique dans la mise en œuvre de SIR, est l'inversion de la matrice  $\Sigma$ . La structure des bases de données, fait que la matrice  $\Sigma$  n'est pas toujours de rang plein, elle est donc mal conditionnée. Lorsque c'est le cas, nous appliquerons, sur les données, une analyse en composante principale (PCA) [6] comme pré-traitement à SIR pour améliorer le conditionnement de  $\Sigma$  en diminuant la dimension de l'espace.

## 2.2 Exemple

Dans cet exemple l'espace de départ est de dimension 4, et nous chercherons à trouver le sous-espace de dimension 1 liant  $x$  à  $y$  sachant que  $y(x) = 2x_1x_2 + x_3 + 0x_4 + \epsilon$  et que les éléments de  $x$  suivent une distribution uniforme  $\mathcal{U}[0, 1]$ . Le bruit  $\epsilon \sim \mathcal{N}(0, 0.01I_4)$

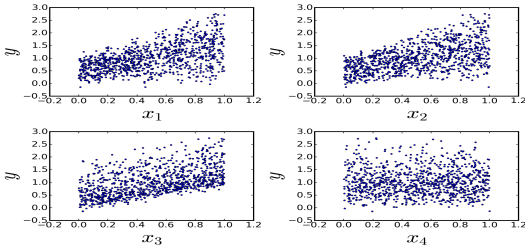


FIGURE 1 – Valeurs prises par  $y$  (ordonnées) par rapport aux réalisations de  $\mathbf{X}$  (abscisses). Il n'est pas évident de trouver un lien entre les réalisations de  $\mathbf{X}$  et la valeur prise par  $y$  à partir de cet espace.

La figure 1 montre, qu'en l'absence de la connaissance de la fonction liant  $x$  à  $y$ , il est nécessaire dans cet exemple de trouver une direction permettant de retrouver la valeur de  $y$ .

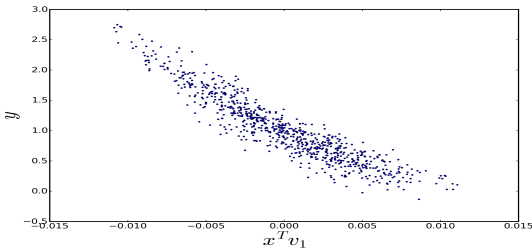


FIGURE 2 – Valeurs prises par  $y$  par rapport aux réalisations de  $\mathbf{X}$  projetées sur la direction associée à la plus grande valeur propre donnée par SIR. Malgré la présence de bruit, on est en mesure d'observer un lien très clair entre les données et la valeur de  $y$

Grâce à la figure 2, on peut voir que SIR, grâce à la prise en compte des variations de  $y$  lors de la recherche de directions pertinentes, est à même de fournir une projection qui montre une forte corrélation avec les données. L'application de SIR nous permet d'obtenir un sous-espace, formé par la ou les directions les plus pertinentes.

## 3 Détermination de paramètres grâce à SIR

Une fois que les données sont projetées il faut donc lier l'espace des données avec celui des paramètres. Le nombre optimal de directions doit donc être déterminé. Pour ce faire, nous pouvons procéder visuellement ou par un protocole de validation croisée, mais seulement dans l'hypothèse où les directions les plus pertinentes sont toujours les premières, et où la dimension optimale du sous-espace est toujours la même. Une fois tous les échantillons de la base projetés, l'échantillon dont on souhaite connaître la valeur du paramètre est à son tour projeté. L'estimateur de la valeur du paramètre est obtenue à partir de ces plus proches voisins dans le sous-espace, par une moyenne des valeurs du paramètre correspondant à ceux-ci. Dans cette partie nous essaierons de retrouver grâce à SIR les informations sur la profondeur et la largeur de raies spectrales gaussiennes bruitées :

$$x(\lambda; h, w) = 1 - h \times e^{-\frac{1}{2}(\frac{\lambda - \lambda_0}{w})^2} + \epsilon \quad (4)$$

$\lambda_0 = 0$  est commune à tous les signaux.  $h$  et  $w$  sont uniformément distribués  $h \sim \mathcal{U}[0.1, 0.9]$ ,  $w \sim \mathcal{U}[1, 3]$  et le bruit  $\epsilon$  suit une distribution gaussienne  $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$ .

Dans ce cas, une PCA (sous-optimale) sera appliquée comme un pré-traitement à SIR de sorte à décroître suffisamment le conditionnement de  $\Sigma$  pour permettre son inversion. La variance du bruit supposée indépendante de la puissance du signal sur chaque composante de l'espace, est valable pour l'application qui nous concerne : des spectres haute résolution à haut rapport signal sur bruit.

### 3.1 Détermination du nombre optimal de voisins à considérer

Pour l'estimation de la valeur du paramètre, une moyenne au sens des  $k$  plus proches voisins a plus de sens qu'une régression linéaire, car pour ce qui concerne notre problème, le lien entre données et paramètres n'a pas de raison d'être linéaire. La considération des plus proches voisins permet de palier le problème des non-linéarités sous l'hypothèse que localement les valeurs du paramètre prises par les individus sont proches. Le choix du nombre de voisins dépend de la structure et de la richesse de la base de données de référence. Si l'on prend un trop grand nombre de voisins, on risque de ne plus pouvoir vérifier l'hypothèse de faible variabilité locale; au contraire si l'on ne considère pas suffisamment de voisins, l'estimation sera très sensible au bruit. Dans le cadre d'une base de données échantillonnée de manière homogène dans les paramètres, il

est possible de déterminer un nombre optimal de voisins quelle que soit la position dans la base (pour avoir la meilleure précision possible compte tenu des non-linéarités). Pour évaluer ce nombre optimal de voisins, nous avons procédé par validation croisée.

### 3.2 Sélection des directions pertinentes

Nous émettons l’hypothèse que suivant la position de l’échantillon projeté dans le sous-espace les directions pertinentes pour sa représentation au sens du paramètre recherché ne sont pas nécessairement les mêmes. La variété portant l’information sur les variations du paramètre ne serait donc pas toujours portée par les mêmes directions du sous-espace. Nous proposons une alternative au sous-espace particulier en sélectionnant les directions jugées pertinentes pour chaque échantillon. Les directions choisies ne seront pas nécessairement celles des vecteurs propres associés aux plus grandes valeurs propres de  $\Sigma^{-1}\Gamma$ , mais celles qui seront les plus informatives au sens de la parcimonie des valeurs prises par  $y$  compte tenu de la projection de l’échantillon à identifier (cf. [7]). Si l’on observe les directions obtenues lorsque l’on considère les largeurs, on obtient les résultats présentés figure 3.

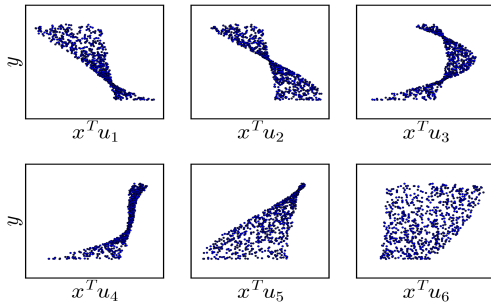


FIGURE 3 – Valeurs des paramètres de largeur des gaussiennes  $y$  par rapport aux valeurs des signaux  $\mathbf{X}$  projetés sur les 6 premières directions données par SIR. On peut observer sur ces différents coefficients de projection, l’expression du lien non-linéaire entre les échantillons de la gaussienne et sa largeur.

La figure 3 montre que suivant la position de la projection d’un échantillon, les directions ne portent pas la même qualité d’information. Par exemple un échantillon se projetant sur les valeurs médianes de la direction  $u_1$  obtient grâce à celle-ci une information très peu précise quant à la valeur de  $y$ . Au contraire si la valeur prise par  $x^T u_1$  est proche de la limite supérieure, on est en mesure, grâce à cette seule direction, de déterminer précisément la valeur de  $y$ . On peut donc choisir suivant la position de l’échantillon dans le sous-espace quelles directions porteront l’information pertinente concernant la valeur de  $y$  associée. Dans le cas du calcul des profondeurs, le lien entre le paramètre et le signal étant linéaire, l’information pertinente est toujours portée par les mêmes directions de SIR. En procédant de la sorte, nous pouvons quantifier l’erreur moyenne en valeur absolue commise lors de la détermination des paramètres, et comparer ce résultat avec ce qu’aurait donné le meilleur sous-espace obtenu par PCA.

TABLE 1 – Erreurs moyennes en valeur absolue lors de la détermination des profondeurs et des largeurs des gaussiennes avec SIR ou la PCA.

	PCA	SIR
Profondeurs	0.0063	0.0025
Largeurs	0.029	0.024

Les erreurs présentées en table 1, sont calculées à partir de la détermination des valeurs de paramètres d’un cinquième de la base de données, le reste servant de base de référence et à la construction du sous-espace. Ces résultats présentés table 1 montrent que SIR permet d’obtenir une meilleure précision que ce que l’on peut attendre d’une PCA telle qu’utilisée dans [8].

## 4 Application à des spectres synthétiques

Nous comparerons les résultats obtenus avec ceux que l’on obtiendrait avec une PCA en suivant le protocole décrit dans [8]. Comme expliqué en section 2.1, une PCA sera appliquée en pré-traitement. La base de données contient 23000 spectres synthétiques et chacun de ces spectres est un vecteur de 8000 points couvrant le domaine spectral 390 – 680 nm (cf. [8]). Le pré-traitement ramènera cet espace à un espace de dimension 50 grâce à une PCA. En termes de valeurs prises par les paramètres (mêmes paramètres que dans [8]), la base est échantillonnée uniformément comme suit :  $T_{\text{eff}} \in [4000 : 8000] K$ ,  $\Delta(T_{\text{eff}}) = 100 K$ ,  $\log(g) \in [4 : 5] \text{ dex}$ ,  $\Delta(\log(g)) = 0.2 \text{ dex}$ ,  $[Fe/H] \in [-1 : 0.5] \text{ dex}$ ,  $\Delta([Fe/H]) = 0.1 \text{ dex}$ , nous nous intéressons aussi à un quatrième paramètre, la vitesse de rotation projetée,  $v \sin(i) \in [0 : 100] \text{ km/s}$ ,  $\Delta(v \sin(i)) = 2 \text{ km/s}$  entre 0 km/s et 20 km/s et  $\Delta(v \sin(i)) = 5 \text{ km/s}$  entre 20 km/s et 100 km/s. ( $\Delta$  représente le pas d’échantillonnage de la base de données) Pour chaque calcul 1000 spectres sont aléatoirement extraits de la base et bruités avec un bruit gaussien centré et d’écart-type  $\sigma = 0.03$  (cohérent avec les données observationnelles qui nous intéressent à terme) pour servir de base de test. On obtient alors avec SIR pour chacun des paramètres des directions telles que illustrés figure 4.

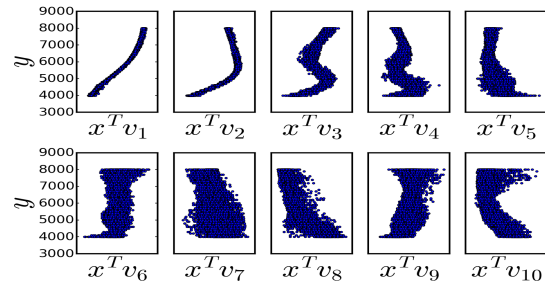


FIGURE 4 – Directions obtenues grâce à SIR pour la détermination des températures effectives ( $T_{\text{eff}}$ )

On peut observer sur la figure 4 que les premières directions sont très informatives, de part le lien que l’on voit apparaître entre  $x^T v_i$  et  $y$ , quant à la valeur prise par  $T_{\text{eff}}$ . Il est intéressant de noter que la direction numéro 5 par exemple n’apporte une

information pertinente que pour le tiers supérieur de  $x^T v_5$ .

On peut comparer les directions de la figure 4 avec les projections sur les composantes obtenues avec PCA en figure 5.

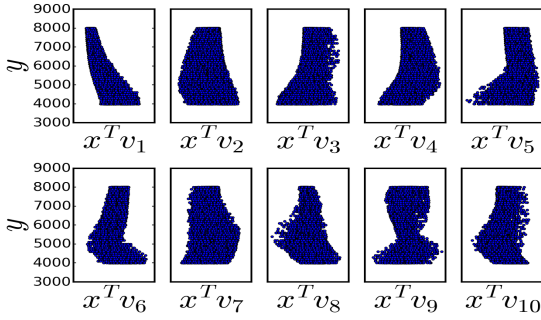


FIGURE 5 – Directions obtenues grâce à la PCA pour la détermination des températures effectives ( $T_{\text{eff}}$ )

On s’aperçoit en regardant la figure 5, que les directions obtenues par PCA traduisent, de part leurs construction, beaucoup moins bien la variation du paramètre  $T_{\text{eff}}$  : cela signifie qu’il faut un plus grand nombre de directions pour avoir une information aussi précise qu’avec SIR. Mais prendre plus de directions signifie aussi plus de bruit. Il est donc peut probable que la PCA seule puisse avoir des résultats aussi précis que ceux atteints grâce à SIR.

Pour cet exemple, la base étant régulièrement échantillonnée pour tous les paramètres, on peut déterminer un nombre de voisins constant à considérer. Une dizaine de voisins permet de rester en dessous du nombre d’échantillons pour une valeur d’échantillonnage du paramètre, nous serons ainsi peu sensibles aux non linéarités. L’expérience montre que entre 10 et 30 voisins, les résultats ne sont pas différentiables

TABLE 2 – Erreurs moyennes en valeur absolue lors de la détermination des paramètres stellaires fondamentaux et du  $v \sin(i)$  sur les spectres synthétiques.

	PCA	SIR
$T_{\text{eff}}$	132 K	61 K
$\log(g)$	0.25 dex	0.10 dex
$[Fe/H]$	0.07 dex	0.03 dex
$v \sin(i)$	1.2 km/s	0.09 km/s

On peut observer table 2 que, comme on s’y attendait, les résultats que l’on obtient avec SIR sont bien plus précis que ceux obtenus grâce à la PCA. L’inconvénient que l’on pourrait trouver à SIR, est que la méthode dans le cas où  $\Sigma$  est mal conditionnée, nécessite un pré-traitement. Elle nécessite aussi que l’on traite indépendamment chaque paramètre, ce qui rend son exécution plus lourde.

## 5 Conclusion

Nous avons vu que SIR permettait de projeter des données sur le sous-espace qui maximise la cohérence de variation entre

elles et un paramètre pré-défini. Cette approche ne nécessite pas la connaissance d’une fonction permettant de passer de l’espace des données à celui du paramètre. Il est nécessaire lorsqu’on l’applique de faire attention au conditionnement de  $\Sigma$ . Pour la détermination de paramètres stellaire, la méthode basée sur SIR se révèle être très efficace sur les spectres synthétiques par rapport à la méthode basée sur PCA. La PCA a l’avantage, quant à elle, de ne nécessiter qu’un espace de projection pour tous les paramètres. Les premiers tests sur des spectres observés montrent une amélioration d’environ 25% sur la température et sur la métallicité par rapport aux résultats de PCA. Cependant les bases de spectres observés étant souvent échantillonnées de manière non-homogène concernant certains paramètres -certaines plages de valeurs de  $\log(g)$  contiennent très peu d’objets par exemple-, nous sommes confrontés à d’autres problèmes, notamment concernant la gravité de surface et la vitesse de rotation projetée. Un développement futur sera d’adapter la méthode pour des bases échantillonnées de manière non-homogène.

## Références

- [1] The Gaia Collaboration. The gaia mission. *Astronomy & Astrophysics*, 595 :A1, 2016.
- [2] M. Juric and T. Tyson. Lsst data management : Entering the era of petascale optical astronomy. *Highlights of Astronomy*, 16 :675–676, March 2015.
- [3] K. C. Li. Sliced Inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86 :316–327, 1991.
- [4] Caroline Bernard-Michel, Sylvain Douté, M Fauvel, Laurent Gardes, and Stephane Girard. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using regularized sliced inverse regression. *Journal of geophysical research*, 114, 2009.
- [5] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*, ISSN 1940-6517 Wiley series in probability and mathematical statistics Volume 544 : Applied probability and statistics Wiley-Interscience paperback series. John Wiley & Sons, 2004.
- [6] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.
- [7] V. Watson, J.-F. Trouilhet, F. Paletou, and S. Girard. Inference of an explanatory variable from observations in a high-dimensional space : Application to high-resolution spectra of stars. 2017 IEEE International workshop ECMSM, 2017.
- [8] F. Paletou, T. Böhm, V. Watson, and J.-F. Trouilhet. Inversion of stellar fundamental parameters from Espadons and Narval high-resolution spectra. *Astronomy & Astrophysics*, 573, 2015.