SF2A 2007
J. Bouvier, A. Chalabaev, C. Charbonnel (eds)

A ROBUST CLASSIFICATION OF HIGH-REDSHIFT GALAXIES USING SUPPORT VECTOR MACHINES

M. Huertas-Company[1], D. Rouan[1] and L. Tasca[2]

Abstract. We present a new non-parametric method to quantify morphologies of galaxies based on a particular family of learning machines called support vector machines. The method, that can be seen as a generalization of the classical CAS classification but with an unlimited number of dimensions and non-linear boundaries between decision regions, is fully automated and thus particularly well adapted to large cosmological surveys. The source code is available for download at `http://www.lesia.obspm.fr/~huertas/galsvm.html`

To test the method, we use a seeing limited near-infrared (K_s band, $2, 16\mu m$) sample observed with WIRCam at CFHT at a median redshift of $z \sim 0.8$. The machine is trained with a simulated sample built from a local visually classified sample from the SDSS chosen in the high-redshift sample's rest-frame (i band, $0.77\mu m$) and artificially redshifted to match the observing conditions. We use a 12-dimensional volume, including 5 morphological parameters, and other caracteristics of galaxies such as luminosity and redshift. A fraction of the simulated sample is used to test the machine and assess its accuracy.

We show that a qualitative separation in two main morphological types (late type and early type) can be obtained with an error lower than 20% up to the completeness limit of the sample ($KAB \sim 22$) which is more than 2 times better that what would be obtained with a classical C/A classification on the same sample and indeed comparable to space data. The method is optimized to solve a specific problem, offering an objective and automated estimate of errors that enables a straightforward comparison with other surveys. Selecting the training sample in the high-redshift sample rest-frame makes the results free from wavelength dependent effects and hence its interpretation in terms of evolution easier.

1 Introduction

The process of galaxy formation and the way galaxies evolve is still one of the key unresolved problems in modern astrophysics. Many of the physical details remain uncertain, in particular the process and history of mass assembly. One classical observational way to test the models of galaxy formation is to classify galaxies according to morphological criteria, i.e., the organization of its brightness as projected on the sky's plane and observed at a particular wavelength, defined in the nearby Universe (Hubble et al. 1936), and to follow this classification across time (Abraham et al. 1996). However, a major obstacle is still the difficulty in quantifying morphology of high redshift objects with a few simple, reliable measurements. Indeed, with the increasing number of cosmological surveys available today, classical visual classifications become useless and automated methods must be employed. Globally there exist two main approaches: the first one, known as parametric, consists in modeling the distribution of light with an analytic model and fit it to the real galaxy. A commonly used parameter in this approach is the bulge-to-disk (B/D) light ratio that correlates with qualitative Hubble type classifications, and can be obtained by fitting a two-component profile (e.g. Simard et al. 2002). The second approach is called non-parametric and basically consists in measuring a set of well-chosen parameters that correlate with the Hubble type. The main advantage of this method is that it does not assume a particular analytic model and can therefore be used to classify regular as well as irregular galaxies. Abraham et al. 1994 first proposed this method by defining the concentration and asymmetry (C and A) parameters. They showed that plotting those values in a 2D plane, results in a quite good separation between the three main morphological types (early type, late type and irregulars). Subsequent authors modified then the original definitions to make C and A more robust to surface-brighntess selection, centering errors or redshift dependence (Brinchmann et al. 1998; Conselice et al. 2000) and introduced new parameters. In particular a third parameter the smoothness (S) was proposed by (Conselice et al. 2003) and gave its name to the CAS morphological

[1] Observatoire de Paris, LESIA, CNRS UMR 8109, 92195 Meudon, France
[2] Laboratoire d'Astrophysique de Marseille

classification system. More recently Abraham et al. 2003 and Lotz et al. 2004 proposed two new parameters: the Gini coefficient that correlates with concentration and the M20 moment. Each of those parameters brings a different amount of information concerning the galaxy shape. There is no way, however, with classical approaches to use more than 3 parameters simultaneously. Bershady et al. 2000 made a first attempt to do a multi-parameter analysis on a nearby sample using a 4 dimensional space including concentration and asymmetry as well as luminosity and color information. They found indeed correlations between those parameters and defined six 2D planes resulting from the combinations of those parameters. The classification was however done independently in each plane without considering all the information simultaneously. In the framework of the COSMOS consortia (Scoville et al. 2005), Scarlata et al. 2006 have recently made a step forward by proposing a multi-parameter classification scheme (ZEST) based on the positions of galaxies in a three dimensional space resulting from a principal component analysis on a 5 dimensional space. The method uses almost all the information contained in the 5 parameters, but the final calibration is done in 3 dimensions.
In this paper, we propose a generalization of the non-parametric classification that uses an unlimited number of dimensions and non-linear separators, enabling to use simultaneously all the information brought by the different morphological parameters. The approach uses a particular class of learning machines (called support vector machines) that finds the optimal decision regions in a volume using a training set. Here, we build this training set from a local sample that is transformed to reproduce the physical and instrumental properties of the science sample, allowing to use it even on seeing limited observations. The algorithm defines, in an automated way, the optimal decision regions using multi-dimensional hyper-surfaces as boundaries. It allows therefore a straightforward comparison between different science samples. The classification scheme that we propose is intended as a framework for future studies on large cosmological fields.

The paper proceeds as follows: generalities on pattern recognition and in particular on support vector machines (SVM) are described in the next section. In Section 3, we describe the general steps of the proposed method to classify high-redshift objects. We show, in particular, how the training set is built to reproduce the real sample properties (3.1) and we finally describe several tests performed to probe the accuracy of the method (3.2).

We use the following cosmological parameters throughout the paper: $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and $(\Omega_\mathrm{M}, \Omega_\Lambda, \Omega_\mathrm{k}) = (0.3, 0.7, 0.0)$.

2 Generalities on pattern recognition

Suppose a set of observations of a given phenomenon, in which each observation consists of a vector $\mathbf{x_i} \in \mathbb{R}^n, i = 1, ..., l$ and of an associated ”truth” y_i. For instance, in a classical concentration and asymmetry classification plane, $\mathbf{x_i}$ would be a 2D vector whose components are the concentration and the asymmetry, and y_i would be 0 if the galaxy is irregular, 1 if it is disk dominated and 2 if it is bulge dominated. We then call learning machine, a machine whose task is to learn the mapping $\mathbf{x_i} \mapsto y_i$ defined by a set of possible mappings $\mathbf{x} \mapsto f(\mathbf{x}, \alpha)$. A particular choice of α generates what is called a ”trained machine”.

2.1 Support vector machines

Support vector machines are a particular family of learning machines, first introduced by Vapnik et al. 1995 as an alternative to neural networks and that have been successfully employed to solve clustering problems, specially in biological applications. In order to simplify the description of the most important points concerning SVM we will focus on a 2 class classification problem: $\{\mathbf{x_i}, y_i\}, i = 1, ..., l\, y_i \in \{-1, 1\}, \mathbf{x_i} \in \mathbb{R}^d$ without loss of generalization. The basic idea is to find an hyperplane that separates the positive from the negative examples. If this plane exists, the points $\mathbf{x}$ that lie on the hyperplane satisfy $\mathbf{w}.\mathbf{x} + b = 0$, with $\mathbf{w}$ normal to the hyperplane, and $|b|/||\mathbf{w}||$ the perpendicular distance from the hyperplane to the origin. $d_+(d_-)$ will then be the shortest distance from the separating hyperplane to the closest positive (negative) example. The ”margin” is defined to be: $d_+ + d_-$. The algorithm will then simply look for the separating hyperplane with largest margin. One key feature that can be added to solve more complex problems is the use of non linear decision functions. To do so, we map the data to some other (possibly infinite dimensional) Euclidian space H: $\Phi : \mathbb{R}^d \mapsto H$ where the data can be linearly separable by some hyperplane. Since the only way in which the data appear in the training problem is in the form of dot products $\mathbf{x_i}.\mathbf{x_j}$ then the training algorithm would only depend on the data through dot products in H, i.e. on functions of the form $\Phi(x_i).\Phi(x_j)$. If there is a ”kernel function” K such that $K(x_i, x_j) = \Phi(x_i).\Phi(x_j)$ we would never need to explicitly even know what Φ is. In summary, SVM are a particular family of learning machines that: (a) for linearly separable data, simply look for the optimal separating hyperplane between distributions by maximizing the margin, (b) for non separable data a ”tolerance” parameter C must be added which controls the tolerance to errors and (c) for non linear non separable data a kernel function is built that maps the space into a higher dimensional space where the data are linearly separable. Then the Kernel parameters must be adjusted too.

2.2 Application to galaxies

Abraham et al. 1994 proposed the idea of measuring some well-chosen parameters on a galaxy image that can be easily correlated with its morphology. In their paper they introduced the concentration, which basically measures the fraction of light contained in an inner isophote, and the asymmetry, which measures the degree of symmetry of the galaxy. They showed, that plotting those values in a 2D plane results in a quite good separation between the three main morphological populations: early-type, late-type and irregulars. They consequently plotted linear separators to define the regions and classified a set of galaxies with unknown morphology according to their positions in the so-called C/A plane. In other words, they tried to maximize the margins between 3 populations in a 2 dimensional space using linear separators. The same task can be done in a 3 dimensional space (CAS, Conselice et al. 2003) but it becomes simply impossible with more than 3 dimensions. In this sense SVM offer a straightforward generalization of this method since they can separate samples with an unlimited number of dimensions and use non-linear boundaries.

3 The method

When observing objects at higher redshift with a ground-based telescope the S/N decreases, galaxies become poorly resolved and consequently more symmetric and less concentrated (e.g. Conselice et al. 2000). The separation in the C/A plane turns out to be less clear. That's why space data such as HST imaging are widely used for those purposes and classifications based on colors are usually adopted for ground-based data (e.g. Zucca et al. 2006). It is known however (e.g. Arnouts et al. 2007) that a classification based only on colors is highly contaminated by the presence, for instance, of an important population of ”blue” early-type galaxies, specially at high redshift where the red sequence is building up. That is one of the reasons why classifications based on morphological criteria are preferred. Indeed, with the increasing amount of data coming from ground-based surveys becoming available today it would be interesting to know if it is possible to obtain at least a rough morphological classification from these observations. In the following sections we therefore investigate wether the possibilities of using a large number of parameters and non-linear boundaries offered by support vector machines can help to increase the accuracy of ”pure” morphological classifications on high-redshift ground-based data.

3.1 General procedure

The proposed procedure can be summarized in 4 main steps: (a) Build a training set: for that purpose, we select a nearby visually classified sample at a wavelength corresponding to the rest-frame of the high redshift sample to be analyzed. We then move the sample to the proper redshift and image quality and drop it in the high z background. (b) Measure a set of morphological parameters on the sample. (c) Train a support vector based learning machine with a fraction of the simulated sample and use the other fraction to test and estimate errors. (d) Classify real data with the trained machine and correct for possible systematic errors detected in the testing step.

3.2 Testing

In order to test the method, we work on a sample of galaxies observed with WIRCam at CFHT in the near infrared K_s band. The field is part of the Canada-France Hawaii Telescope Legacy Survey (CFHTLS) Deep survey and its near infrared follow-up and it is centered on the COSMOS area (Scoville et al. 2005). We use a cutout of $10^{'} \times 10^{'}$ to perform all the tests. The sample is complete up to $K(AB) = 22$ and the median photometric redshift is ~ 0.8. Images are reduced wit the Terapix pipeline[1] and have a pixel scale of $0.15^{"}$ with a mean FWHM of $0.7^{"}$.

3.2.1 Building the training sample

We use a local catalog of 1472 objects from the Sloan Digital Sky Survey in the i band, which roughly corresponds to the rest-frame of the K-band at $z \sim 1$ and that has been visually classified (Tasca & White 2006).

[1] http://terapix.iap.fr

We first generate a random pair of (magnitude, redshift) values with a probability distribution that matches the real magnitude and redshift distribution of the sample to be simulated. Then, for every galaxy stamp, we proceed in four steps: First, we remove all the foreground stars and all other sources that do not belong to the galaxy itself. Second, we degrade the resolution to reach the one at high redshift: we measure the FWHM at high redshift (f_{hz}), convert it to Kpc using a standard ΛCDM cosmology and deduce the resolution the local galaxy must have (f_{lz}). Then the image is convolved with a 2D gaussian function of $FWHM = \sqrt{(f_{lz}^2 - f_i^2)}$, where f_i is the local galaxy's initial resolution. Third, the image is binned to reach the expected angular size at high redshift with the 0.15" pixel scale. In this step, the image is also scaled to its new magnitude. Finally, we drop the galaxy in a real background image.

Once the simulated galaxies are dropped in a real background, we measure 5 morphological parameters (C, A, S, M20, Gini), 2 shape parameters (ellipticity and CLASS_STAR), 2 size parameters (isophotal area, petrosian radius), a luminosity parameter (magnitude) and the photometric redshift as a distance parameter to build a 12-D space that we use to train the SVM.

3.2.2 12-D versus 2-D SVM

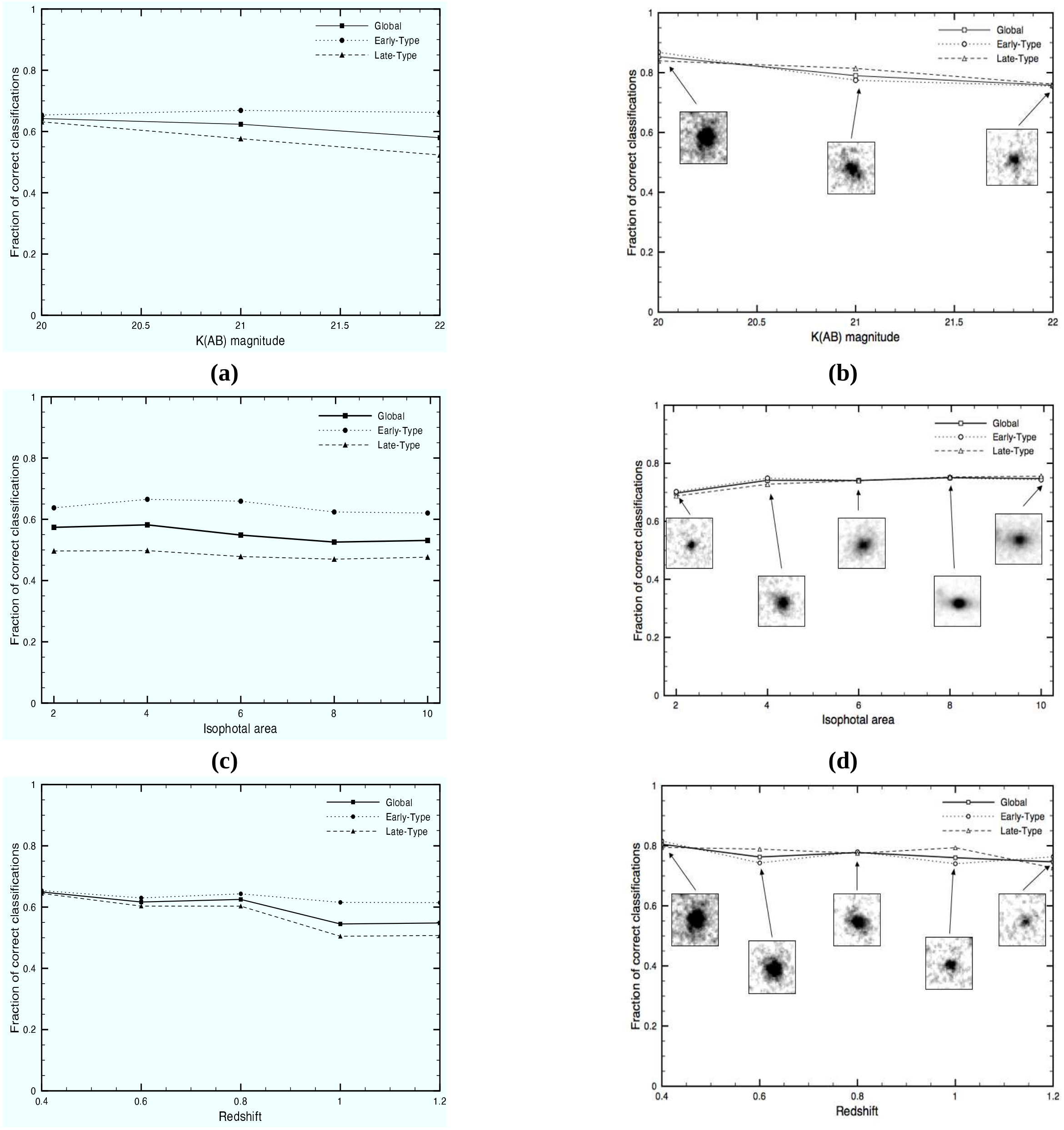
We trained then 2 machines: the first one, with only 2 parameters (C and A), which should globally give the same results as a classical C/A classification (Huertas-Company et al. 2007) and the second one with 12 parameters described above. We then tested both machines by looking at the fraction of galaxies that are correctly classified. Results for the whole sample are summarized in table 1. We observe that including more than two parameters in the classification results in a significant gain for this sample where C/A cannot do much better than a random choice. To establish the robustness of this effect we look at the accuracy of the classification as a function of 3 main properties of the galaxies: luminosity, distance and area (Fig. 1) by progressively adding objects and measuring each time : a) the global accuracy, i.e. the fraction of galaxies that are classified correctly by the machine, and b) the accuracy per morphological type, i.e. the fraction of predicted early (late) type galaxies that are visually classified as early (late) type respectively ($N_{E\rightarrow E}$ and $N_{S\rightarrow S}$). We observe that the 12-D machine results in a more robust and symmetric response in all the magnitude, redshift and size bins.

	Classical C/A		SVM C/A		SVM 12-D	
	Early-Type	Late-Type	Early-Type	Late-Type	Early-Type	Late-Type
Visual Early-Type	0.59 (96)	0.51 (321)	0.57 (304)	0.45 (113)	0.75 (365)	0.18 (52)
Visual Late-Type	0.41 (65)	0.49 (309)	0.43 (236)	0.55 (138)	0.25 (149)	0.82 (225)

Table 1. Comparison of the accuracy of three classifications of the WIRCam sample: Classical C/A, SVM C/A and 12-D SVM. The table shows for each method the relations between the visual and the predicted morphological classes. The number of objects are enclosed in parentheses. (see text for details)

4 Conclusions

We have presented a new method to perform morphological classification of cosmological samples based on support vector machines. It can be seen as a generalization of the classical non-parametrical C/A classification method but with an unlimited number of dimensions and non linear boundaries between the decision regions. The method is specially adapted to be used on large cosmological surveys since it is fully automated and errors are estimated objectively allowing an easy comparison between surveys with different properties. As a test, we use our method to classify a near-infrared seeing-limited sample observed with WIRCam at CFHT with a training set of ∼ 1500 objects from the SDSS. We show that increasing the number of parameters in the analysis reduces errors by more than a factor 2; leading to a mean accuracy of ∼ 80% of correct classification up to the sample completeness limit ($K_{AB} \sim 22$). The presented method is intended as a framework for future studies. The library is available for download at `http://www.lesia.obspm.fr/~huertas/galsvm.html`



(a) **(b)**

(c) **(d)**

(e) **(f)**

Fig. 1. Cumulative accuracy of classifications for a 2D machine (left column) and a 12D one (right column) as a function of magnitude (a and b), area (c and d) and redshift (e and f). Solid line shows the global accuracy, i.e. the number of galaxies correctly identified, dotted and dashed lines show respectively the fraction of early type and late type galaxies classified correctly. Stamps in the right column show a typical galaxy for every magnitude, area and redshift range.

References

Abraham, R. G., Valdes, F., Yee, H. K. C., &van den Bergh 1994, ApJ, 432, 75

Abraham, R., van den Bergh, S., Glazebrook, K., Ellis, R., Santiago, B., Surma, P., & Griffiths, R. 1996, ApJ Supplement, 107, 1

Abraham, R. G., van den Bergh, S., & Nair, P. 2003, ApJ, 588, 218

Arnouts, S. , Walcher, C. J. , Le Fevre, O. , Zamorani, G. , Ilbert, O. , Pozzetti, L. , Bardelli, S. , Tresse, L. , Zucca, E. , Le Brun, V. , Charlot, S. , Lamareille, F. , McCracken, H. J. , Bolzonella, M. , Iovino, A. , Lonsdale, C. , Polletta, M. , Surace, J. , Bottini, D. , Garilli, B. , Maccagni, D. , Picat, J. P. , Scaramella, R. , Scodeggio, M. , Vettolani, G. , Zanichelli, A. , Adami, C. , Cappi, A. , Ciliegi, P. , Contini, T. , de la Torre, S. , Foucaud, S. , Franzetti, P. , Gavignaud, I. , Guzzo, L. , Marano, B. , Marinoni, C. , Mazure, A. , Meneux, B. , Merighi, R. , Paltani, S. , Pello, R. , Pollo, A. , Radovich, M. , Temporin, S. , Vergani, D. 2007, ArXiv e-prints, astro-ph/0705.2438

Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645

Brinchmann, J., Abraham, R., Schade, D., Tresse, L., Ellis, R. S., Lilly, S., Le Fevre, O., Glazebrook, K., Hammer, F., Colless, M., Crampton, D., &Broadhurst, T. 1998, ApJ, 499, 112

Conselice, C. J., Bershady, M. A., & Jangren, 2000, ApJ, 529, 886

Conselice, C. J., Bershady, M. A., & Dickinson, M., Papovich, C. 2003, AJ, 126, 1183

Hubble, N.P. 1936, ApJ, 415

Huertas-Company, M., Rouan, D., Soucail, G., Le Fèvre, O. , Tasca, L., Contini, T. 2007, A&A, 468, 937

Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163

Scarlata, C., Carollo, C. M., Lilly, S. J., Sargent, M. T., Feldmann, R., Kampczyk, P., Porciani, C., Koekemoer, A., Scoville, N., Kneib, J., Leauthaud, A., Massey, R., Rhodes, J., Tasca, L., Capak, P., Maier, C., McCracken, H. J., Mobasher, B., Renzini, A., Taniguchi, Y., Thompson, D., Sheth, K., Ajiki, M., Aussel, H., Murayama, T. , Sanders, D. B., Sasaki, S., Shioya, Y., &Takahashi, M. 2006, ArXiv Astrophysics e-prints, astro-ph/0611644

Scoville, N. Z., &COSMOS Team 2005, Bulletin of the American Astronomical Society, 2005, 1309

Simard, L., Willmer, C. N. A., Vogt, N. P., Sarajedini, V. L., Phillips, A. C., Weiner, B. J., Koo, D. C., Im, M., Illingworth, G. D., & Faber, S. M. 2002, ApJ Supplement, 142, 1

Tasca, L. & White, S. 2006, ArXiv Astrophysics e-prints, astro-ph/0507249

Vapnik, V. 1995, Springer-Verlag, 536

Zucca, E. , Ilbert, O. , Bardelli, S. , Tresse, L. , Zamorani, G. , Arnouts, S. , Pozzetti, L., Bolzonella, M. , McCracken, H. J. , Bottini, D. , Garilli, B. , Le Brun, V. , Le Fèvre, O. , Maccagni, D. , Picat, J. P. , Scaramella, R. , Scodeggio, M. , Vettolani, G. , Zanichelli, A. , Adami, C. , Arnaboldi, M. , Cappi, A. , Charlot, S. , Ciliegi, P. , Contini, T. , Foucaud, S. , Franzetti, P. , Gavignaud, I. , Guzzo, L. , Iovino, A. , Marano, B. , Marinoni, C. , Mazure, A. , Meneux, B. , Merighi, R. , Paltani, S. , Pellò, R. , Pollo, A. , Radovich, M. , Bondi, M. , Bongiorno, A. , Busarello, G. , Cucciati, O. , Gregorini, L. , Lamareille, F. , Mathez, G. , Mellier, Y. , Merluzzi, P. , Ripepi, V. , & Rizzo, D. 2006, AAP, 455, 879