A&A 478, 971–980 (2008)
DOI: 10.1051/0004-6361:20078625

Astronomy & Astrophysics

A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images

I. Method description⋆

M. Huertas-Company[1,4], D. Rouan[1], L. Tasca[3], G. Soucail[2], and O. Le Fèvre[3]

[1] LESIA-Paris Observatory, 5 place Jules Janssen, 92195 Meudon, France
e-mail: marc.huertas@obspm.fr
[2] Laboratoire d'Astrophysique de Toulouse-Tarbes, CNRS-UMR 5572 and Université Paul Sabatier Toulouse III, 14 avenue Belin, 31400 Toulouse, France
[3] LAM, Marseille Observatory, Traverse du Siphon, Les trois Lucs, BP 8, 13376 Marseille Cedex 12, France
[4] IAA-C/ Camino Bajo de Huétor, 50, 18008 Granada, Spain

Received 6 September 2007 / Accepted 7 November 2007

ABSTRACT

Context. Morphology is the most accessible tracer of the physical structure of galaxies, but its interpretation in the framework of galaxy evolution still remains a problem. Its dependence on wavelength renders the comparison between local and high redshift populations difficult. Furthermore, the quality of the measured morphology being strongly dependent on the image resolution, the comparison between different surveys is also a problem.
Aims. 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, which can be seen as a generalization of the classical C/A 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
Methods. To test the method, we use a seeing limited near-infrared (K_s band, 2, 16 μm) 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 μm) and artificially redshifted to match the observing conditions. We use a 12-dimensional volume, including 5 morphological parameters, and other characteristics of galaxies such as luminosity and redshift. A fraction of the simulated sample is used to test the machine and assess its accuracy.
Results. 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.

Key words. galaxies: fundamental parameters – galaxies: high-redshift – methods: data analysis

1. Introduction

The process of galaxy formation and the way galaxies evolve is still one of the key unresolved problems in modern astrophysics. In the currently accepted hierarchical picture of structure formation, galaxies are thought to be embedded in massive dark halos that grow from density fluctuations in the early universe (Fall & Efstathiou 1980) and initially contain baryons in a hot gaseous phase. This gas subsequently cools, and some fraction eventually condenses into stars (Lilly et al. 1996; Madau et al. 1998). However, many of the physical details remain uncertain, in particular the process and history of mass assembly. One classical observational way to test those models is to classify galaxies according to morphological criteria, i.e., the organization of their brightness as projected on the sky's plane and observed at a particular wavelength, defined in the nearby Universe (Hubble 1936; de Vaucouleurs 1948; Sandage 1961), and to follow this classification across time (Abraham et al. 1996; Simard et al. 2002; Abraham et al. 2003). Comparison of distant populations with those found in the nearby Universe might help to clarify the formation history of the galaxy (Cole et al. 2000; Baugh et al. 1996). Progress in this field has come from observing deeper and larger samples, but also from obtaining higher spatial resolution at a given flux and at a given redshift. In the visible, progress has been simultaneous on those two fronts, thanks in particular to the ultra-deep HDF fields observed with the Hubble Space Telescope (HST). HST imaging has brought observational evidence that galaxy evolution is differentiated with respect to morphological type and that a large fraction of distant galaxies

⋆ Based on observations obtained at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique of France, and the University of Hawaii.

have peculiar morphologies that do not fit into the elliptical-spiral Hubble sequence (Brinchmann et al. 1998; Wolf et al. 2003; Ilbert et al. 2006b).

However, the difficulty in quantifying morphology of high redshift objects with a few simple, reliable measurements is still a major obstacle. 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 fitting 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 (Simard et al. 2002; Peng et al. 2002). The main advantage of such a method is that the fitting output provides a quantitative morphology, i.e. a complete set of parameters that describe the galaxy's shape (e.g. disk scale length, bulge effective radius). Results are, unfortunately, often degenerated because of the high number of parameters to be adjusted (Huertas-Company et al. 2007), even when the residuals are almost null, and the obtained accuracy strongly depends on the observing conditions (e.g. angular resolution, S/N). Moreover, this approach assumes that the galaxy is well described by a simple, symmetric profile, which is not true for irregular or well resolved objects.

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. The resulting morphology however will be more qualitative. Abraham et al. (1996) 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 then modified the original definitions to make C and A more robust to surface-brightness selection, centering errors or redshift dependence (Brinchmann et al. 1998; Wu 1999; Bershady et al. 2000; 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 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 these parameters brings a different amount of information concerning the galaxy shape. However, because of the so-called "curse of dimensionality", adding additional parameters does not always result in a significant gain. The information added should indeed outweigh the penalty of complicating the classification procedure, and this depends on the analyzed sample. Consequently, one interesting question is under which conditions it becomes useful to use all these parameters simultaneously. 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 correlations between those parameters and defined six 2D planes resulting from the combinations of those parameters. The classification however, was done independently in each plane without considering all the information simultaneously. In the framework of the COSMOS consortia (Scoville & COSMOS Team 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.

Another key point in this kind of analysis is to correctly calibrate the volume, i.e. to correctly estimate the decision regions. One approach is to use boundaries defined in the nearby universe from a visually classified sample and assume that they will remain unchanged for a sample at high redshift, observed at a different wavelength and with another instrument (Abraham et al. 1996). However, it is well known that the galaxy morphology depends on wavelength (K-correction) and on the observing conditions, which is why some corrections should be applied to take these effects into account (Brinchmann et al. 1998). Another approach consists in visually classifying a fraction of the sample and plotting the boundaries according to the positions of galaxies in the space (Menanteau et al. 2006; Scarlata et al. 2006). This of course takes into account the observing conditions of the sample but requires enough resolution and S/N to be able to decide the galaxy morphology visually. This is possible for space observations but becomes more difficult for ground-based observations, where the low resolution does not allow a reliable visual classification. In all these approaches, boundaries are forced to be linear (2D lines or hyper-planes) and are generally plotted manually, which introduces a subjective element that makes a correct estimate of errors more difficult.

In this paper, therefore, we propose a generalization of the non-parametric classification that uses an unlimited number of dimensions and non-linear separators, enabling us to use all the information brought by the different morphological parameters simultaneously. 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 one 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 therefore allows 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 Sect. 3 we make sure that SVM work properly when applied to a nearby sample, and investigate the effect of adding dimensions to a well-resolved sample. In Sect. 4, 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 (4.1), we define the parameters measured for the morphological classification (4.3) and we finally describe several tests performed to probe the accuracy of the method (4.4).

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 we take a set of observations of a given phenomenon, in which each observation consists of a vector $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1, ..., l$

and an associated "truth" y_i. For instance, in a classical concentration and asymmetry classification plane, $\boldsymbol{x}_i$ 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 $\boldsymbol{x}_i \mapsto y_i$ defined by a set of possible mappings $\boldsymbol{x} \mapsto f(\boldsymbol{x}, \alpha)$. 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 (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: $\{\boldsymbol{x}_i, y_i\}, i = 1, ..., l\ y_i \in \{-1, 1\}, \boldsymbol{x}_i \in \mathbb{R}^d$ without loss of generalization. The basic idea is to find a hyperplane that separates the positive from the negative examples. If this plane exists, the points $\boldsymbol{x}$ that lie on the hyperplane satisfy $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, with $\boldsymbol{w}$ normal to the hyperplane; $|b|/\|\boldsymbol{w}\|$, 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: $d_+ + d_-$. The algorithm will then simply look for the separating hyperplane with the largest margin. This can be formulated as follows:

1. $\boldsymbol{x}_i \cdot \boldsymbol{w} + b \geq +1$ for $y_i = +1$
2. $\boldsymbol{x}_i \cdot \boldsymbol{w} + b \leq -1$ for $y_i = -1$.

The training points for which the equalities hold and whose removal would change the solution are called support vectors (Fig. 1).

It is possible (and it is the most common case) that there is no linear hyperplane that perfectly separates the two data sets. In this case we can relax constraints by introducing a positive stack variable ξ_i, $i = 1, ..., l$ and the equations become then:

1. $\boldsymbol{x}_i \cdot \boldsymbol{w} + b \geq +1 - \xi_i$ for $y_i = +1$
2. $\boldsymbol{x}_i \cdot \boldsymbol{w} + b \leq -1 + \xi_i$ for $y_i = -1$.

The global effect is to change the objective function to be minimized to $\|w\|^2/2 + T(\sum \xi_i)$, where T is a parameter to be chosen by the user, a larger T corresponding to assigning a higher penalty to errors.

Another feature that can be added to solve more complex problems are non linear decision functions. To do this, 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 $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$ then the training algorithm would only depend on the data through dot products in H, i.e. on functions of the form $\Phi(x_i) \cdot \Phi(x_j)$. If there is a "kernel function" K such that $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, we would never need to explicitly know what Φ is. Examples of kernels are: $K(x, y) = (x.y + 1)^p$ (Polynomial), $K(x, y) = e^{-g\|x-y\|^2}$ (Gaussian RBF).

In summary, SVM are a particular family of learning machines that:

- for linearly separable data, simply look for the optimal separating hyperplane between distributions by maximizing the margin;
- for non separable data, a "tolerance" parameter C must be added that controls the tolerance to errors;

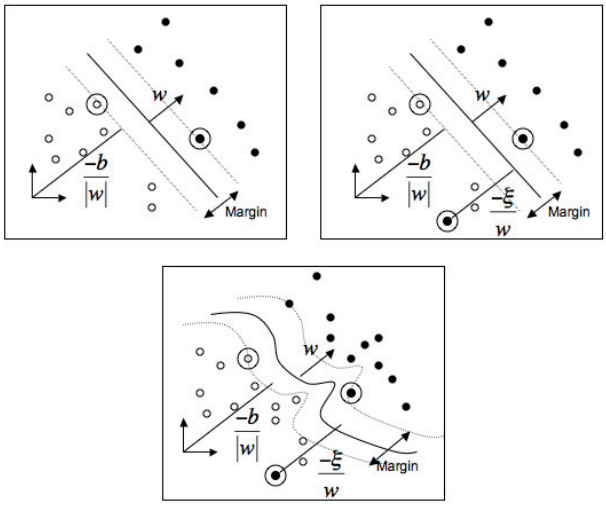

Fig. 1. 2D illustration of the three cases of SVM classification. *Top left*: linearly separable data with linear boundaries. *Top right*: non-linearly non-separable data with linear boundaries. *Bottom*: non-linearly non-separable data with non-linear borders.

- 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. The kernel parameters must be then adjusted too.

2.2. Application to galaxies

Abraham et al. (1996) 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 the logarithms of those values in a 2D plane results in quite a 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 and a logarithmic kernel. 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.

Previous works (Abraham et al. 1996; Brinchmann et al. 1998) have shown that morphological classification is far from being a linearly separable problem, since the contamination in the C/A plane is quite high. We have chosen therefore to use the most general SVM, i.e. a non linear machine with an RBF kernel. A machine is thus parameterized with two parameters: the tolerance (T) and the kernel exponential factor (g). Each possible combination of those two parameters generates a family of functions $f_{T,g}(\alpha, x_i)$. T and g must be fixed before performing the training and α is the result of the training procedure. There exist several techniques for finding the best T and g values for a given problem; here we will use a cross-validation method described in Chang & Lin (2001). This simply consists in performing a

Table 1. Comparison of the accuracy of three classifications of the SDSS sample: classical C/A, SVM C/A and 4D 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.)

	Classical C/A		SVM C/A		SVM 4D	
	Early-type	Late-type	Early-type	Late-type	Early-type	Late-type
Visual early-type	0.80 (254)	0.09 (17)	0.79 (256)	0.08 (15)	0.79 (251)	0.10 (20)
Visual late-type	0.20 (65)	0.91 (172)	0.21 (72)	0.92 (166)	0.21 (67)	0.90 (171)

systematic search over a grid of possible values and selecting the pair that gives the best results.

Our goal is therefore to train a support vector based machine to estimate the morphology of a high redshift sample. Throughout the paper we use the free available package *libSVM* (Chang & Lin 2001). The procedure is basically the same as in a classical C/A classification but using a trained SVM to plot the optimal boundaries. As we show below, this allows us to use more than two morphological parameters simultaneously, and also to measure errors in an automated and objective way, which is essential for comparing different classifications.

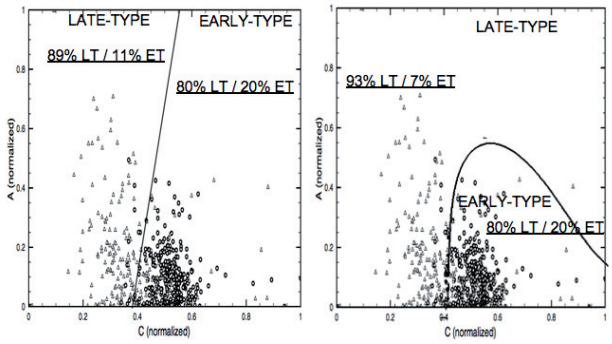

Fig. 2. C/A classical classification (*left*) versus C/A SVM classification (*right*) of 500 nearby objects. Triangles are galaxies visually classified as late-type and circles are galaxies visually classified as early-type. Numbers show the probability that the predicted morphological type is the same as the visual one.

3. Test on a well-resolved nearby sample

3.1. Classical C/A classification versus 2D SVM

In order to verify that SVM work properly when applied to morphological classification of galaxies, we start with a simple test, i.e. classifying a local sample from the Sloan Digital Sky Survey (SDSS) in the *i* band, that has been visually classified (Tasca & White 2005). Galaxies are nearby and consequently well-resolved and with a high S/N. Classical C/A classifications have been proved to give good results in such cases (e.g. Abraham et al. 1996; Menanteau et al. 2006), therefore the idea is to verify that we get at least the same results using SVM and that no extra-biases are introduced.

We therefore measure concentration and asymmetry parameters (see Sect. 4.3 for details). On the one hand we try to plot the best linear boundary by eye as usually done to separate galaxies into two classes (late type and early type), while on the other hand we train a SVM, finally, we compare the outputs. Figure 2 shows the two resulting boundaries. The shape of the boundaries are quite different since SVM does not produce a linear boundary, but when looking at the global accuracy we see that both methods are fully consistent. Indeed, the completeness (the fraction of visual classified galaxies that are correctly recovered) and the contaminations (the fraction of visual classified galaxies that are misclassified) are practically the same for the two methods (see Table 1).

To confirm this consistency and to verify that no extra biases are introduced, we also made a one-to-one comparison of all the galaxies classified with the two methods. We find that 98% (94%) of the early-type (late-type) galaxies classified with the classical C/A method are also classified as early-type (late-type) using the trained SVM.

We conclude that for a high S/N well-resolved sample, such as the SDSS sample, the use of SVM to plot the boundaries is equivalent to using classical procedures. The major advantage, however, is that the boundary is plotted automatically to minimize the errors.

3.2. Classical C/A classification versus 4D SVM

Since one of the main advantages of using SVM is that it can work with an unlimited number of parameters, we investigate the effect of adding dimensions to the SVM classification. We thus classify the same sample as above but with four morphological parameters instead of two: concentration, asymmetry, smoothness and gini (see Sect. 4.3 for details on how they are calculated) and compare the outputs. Results are shown in Table 1. We see that there is no significant gain for this particular case. This suggests that, as proven in previous works (e.g. Abraham et al. 1996; Menanteau et al. 2006), when dealing with a well-resolved and high S/N sample, concentration and asymmetry are enough to obtain an accurate morphological classification.

4. Going to higher redshift

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. For this reason, 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, especially at high redshift where the red sequence is building up. This 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 whether the possibilities of using a large number

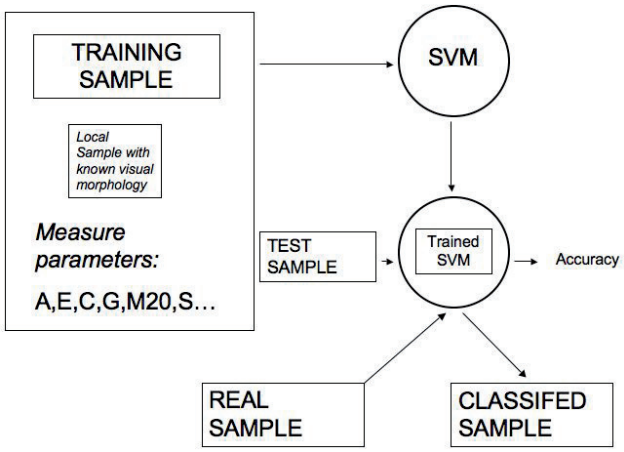

Fig. 3. Steps for morphological classification (see text for details).

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.

4.1. Description of the employed method

The proposed procedure can be summarized in 4 main steps (Fig. 3):

1. 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. This is fully described in Sect. 4.2.
2. Measure a set of morphological parameters on the sample.
3. Train a support vector based learning machine with a fraction of the simulated sample and use the other fraction to test and estimate errors.
4. Classify real data with the trained machine and correct for possible systematic errors detected in the testing step.

In the following sections, we describe each of the steps enumerated above.

4.2. The training set

The most important step in obtaining the morphology with a non-parametric method is to correctly calibrate the volume filled by the data in the multi-dimentional space. This is a critical step since it will determine the decision regions that will be used to perform the classification. Indeed, galaxy morphology depends on the physical properties of the galaxy (luminosity, redshift, wavelength) and on the observing conditions (background level, resolution). A suitable calibration set should consequently reproduce closely all the properties of the sample to be analyzed. One classical approach consists in visually classifying a fraction of the sample and using it as a training set to optimize boundaries (Menanteau et al. 1999, 2006). However, this is not possible for seeing-limited data where the resolution is too low to enable a reliable visual classification. Here, we decide to simulate the high redshift sample from a visually classified local catalog, selected in the rest-frame color of the high redshift sample. This has three main advantages: first, it is free from K-correction effects, second it does not introduce any modeling effect, since the used galaxies are real and finally, the training set is built to reproduce the observing and physical properties of the sample to be analyzed, but it is classified locally, so it does not require a specially high resolution.

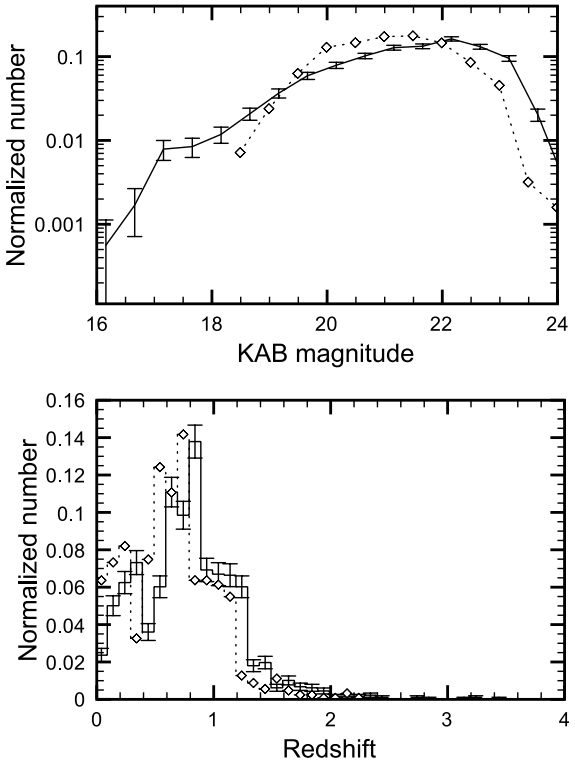

Fig. 4. Magnitude and redshift distributions of the real and simulated sample. Solid line: real sample. Dotted line: simulated sample. Error bars show poissonian errors for the real sample. See text for explanations concerning the differences between the simulated and real distributions.

4.2.1. Real sample

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 & COSMOS Team 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 (Fig. 4). Images are reduced wit the Terapix pipeline[1] and have a pixel scale of $0.15''$ with a mean *FWHM* of $0.7''$. These data are particularly interesting because K-band data have the advantage of probing old stellar populations in the rest-frame, enabling a determination of galaxy morphological types unaffected by recent star formation. Moreover, no space data in this wavelength range are available today. Photometric redshifts come from the publicly available catalogue from Ilbert et al. (2006a), computed with the *LePhare*[2] code on the CFHTLS Deep survey Terapix release T0003 and its multi-color photometric catalogs.

[1] http://terapix.iap.fr
[2] http://cencos.oamp.fr/cencos/CFHTLS/

4.2.2. Building the sample

We therefore used 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 has been visually classified (Tasca & White 2005).

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 (see Fig. 4).

Then, for every galaxy stamp, we proceed in four steps:

1. First, we remove all the foreground stars and all other sources that do not belong to the galaxy itself. For that purpose we use the SExtractor segmentation map (Bertin & Arnouts 1996) and replace all the surrounding sources with a random noise with the same statistics (mean value and variance) as the real background noise. The foreground stars that fall within the galaxy are replaced with the mean value in the galaxy area.
2. Second, we degrade the resolution to reach that 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 that 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.
3. 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. In the scaling procedure we force the final mean background level of the simulated stamp to be at least 3 times lower than the real background. This is to avoiding the local noise dominates over the high-redshift noise when dropping the galaxy in a real background. This implies that objects that are too bright (typically $K_s < 17$) cannot be simulated since the necessary scaling factor is too small, and explains the difference between the real and the simulated magnitude distribution in Fig. 4. The difference in the faint end is due to the fact that some simulated objects are not detected by SExtractor.
4. Finally, we drop the galaxy in a real background image.

Figure 5 illustrates the entire procedure for a spiral galaxy.

In summary, we simulate a high redshift sample from a local sample, selected in the high redshift sample's rest-frame to avoid *K*-correction effects. The sample reproduces the observing conditions (background level, noise, resolution) and physical properties (redshift and magnitude distribution) of the sample to be analyzed.

4.3. Measuring morphological parameters

Once the simulated galaxies are dropped in a real background, we measure the following 5 morphological parameters:

– Concentration: basically, this measures the ratio of light within a circular or elliptical inner aperture to the light within a circular or elliptical outer aperture. Generally, it is defined in slightly different ways by different authors. Here we adopt the Bershady et al. (2000) definition as for the ratio of the circular radii containing 20% and 80% of the "total flux":

$$C = 5\log(r_{80}/r_{20}). \quad (1)$$

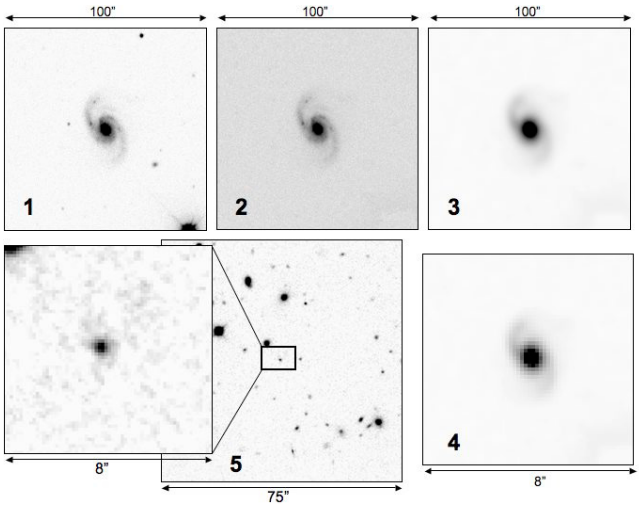

Fig. 5. Example of simulation for a galaxy. 1: SDSS *i* band image; 2: image after subtraction of foreground stars; 3: image after convolution; 4: image after binning; 5: final simulated field with real and simulated galaxies.

We use Conselice's (2003) definition of the total flux as the flux contained within $1.5r_p$ (Petrosian radius). For the concentration measurement, the galaxy's center is that determined by the asymmetry minimization (see below).

– Asymmetry: this quantifies the degree to which the light of a galaxy is rotationally symmetric. It is measured by subtracting the galaxy image rotated by 180° from the original image:

$$A = \frac{1}{2}\left(\frac{\sum |I(i,j) - I_{180}(i,j)|}{\sum I(i,j)} - \frac{\sum |B(i,j) - B_{180}(i,j)|}{\sum I(i,j)}\right), \quad (2)$$

where I is the galaxy image and I°_{180} is the galaxy image rotated by 180° about the galaxy's central pixel, and B is the average asymmetry of the background. The central pixel is determined by minimizing A.

– Smoothness: developed by Conselice et al. (2000), this quantifies the degree of small-scale structure. The galaxy image is smoothed by a boxcar of given width and then subtracted from the original image:

$$S = \frac{1}{2}\left(\frac{\sum |I(i,j) - I_S(i,j)|}{\sum I(i,j)} - \frac{\sum |B(i,j) - B_S(i,j)|}{\sum I(i,j)}\right), \quad (3)$$

where I_S is the galaxy's image smoothed by a boxcar of width $0.25r_p$.

– Moment of Light: introduced by Lotz et al. (2004), the total second-order moment M_{tot} is the flux in each pixel f_i multiplied by the squared distance to the center of the galaxy, summed over all the galaxy pixels assigned by the SExtractor segmentation map: $M_{tot} = \sum f_i[(x_i - x_c)^2 + (y_i - y_c)^2]$, where x_c and y_c is the galaxy's center. The second-order moment of the brightest regions of the galaxy traces the spatial distribution of any bright nuclei, bars, spiral arms and off-center star clusters. We define M_{20} as the normalized second-order moment of the 20% brightest pixels of the galaxy.

– Gini Coefficient: this is a statistic based on the Lorentz curve, i.e. the rank-ordered cumulative distribution function of a population's wealth, or in this case a galaxy's pixel values (Abraham et al. 2003). For the majority of local galaxies, the Gini coefficient is correlated with the concentration index and increases with the fraction of light in a central

Table 2. Comparison of the accuracy of three classifications of the WIRCam sample: classical C/A, SVM C/A and 12D 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.)

	Classical C/A		SVM C/A		SVM 12D	
	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)

component. However, unlike C, G is independent of the large-scale spatial distribution of galaxy's light. Therefore, G differs from C in that it can distinguish between galaxies with shallow light profiles (which have both low C and G) and galaxies where much of the flux is located in a few pixels but not at the center (having low C but high G).

Each of the above parameters measures different properties of a galaxy and therefore give a different amount of information concerning the galaxy's morphological type. For instance, Lotz et al. (2004) used the M_{20}/Gini plane to identify merger candidates, whereas the C/A plane is tipically used to separate late from early type galaxies. A multi-dimensional analysis allows one to use all the information brought by each of the morphological parameters simultaneously to increase the accuracy of the classification. Moreover, previous works have shown that the measured parameters might also depend on the size, the luminosity or the redshift of the galaxy (Brinchmann et al. 1998; Bershady et al. 2000). Therefore, including non-morphologcial parameters should help the machine to take into account systematic trends in the morphological parameters due to luminosity or size variations. We thus measure 7 more parameters that we distribute in 4 classes: shape, size, luminosity and distance, according to the kind of information they measure:

- Shape: we include 2 shape parameters, the galaxy ellipticity as measured by SExtractor, i.e. the ratio of the minor and major axes of the isophotal ellipses describing the galaxy, and the CLASS_STAR parameter, also from SExtractor. This parameter is intended to separate galaxies from stars and results from a neural network classification. Since it spans a continuum range between 0 and 1, it can be interpreted as a measure of the galaxy's compactness.
- Size: the size parameters include the isophotal galaxy area and the petrosian radius.
- Luminosity: we use the apparent magnitude of the galaxy and the mean surface brightness.
- Distance: we adopt the photometric redshift as a measure of the distance.

4.4. Training and testing

We perform several tests to probe the accuracy of the proposed method. For all the tests we adopt the same procedure: we use a fraction of the simulated catalogue (typically 500 galaxies) to train the machine, and the remaining 1000 objects to test it by looking at the fraction of galaxies that are correctly classified. We limit the analysis to only 2 broad morphological classes (late type and early type). The main reason for this choice is that there are too few irregular galaxies in the employed local sample to define a class. There is, however, no loss of generalization and the same analysis can be performed with an unlimited number of classes, provided of course that they are correlated with measured parameters.

4.4.1. Classical C/A classification versus 2D SVM

The first point we try to answer is how good would the classification of this sample be using a classical linear C/A classification. We therefore take the brightest objects of the sample (with known visual morphology) ($K_s < 20$) and try to plot a linear boundary between the two distributions. As expected, the distributions are now poorly separated and plotting a linear boundary becomes extremely difficult. This is confirmed when trying to classify the whole sample (Table 2): the completeness and contaminations are basically the same as we would have obtained with a random choice. We conclude that concentration and asymmetry alone cannot be used on this sample to obtain a reliable morphological classification.

In a second step, we classify this sample with a SVM machine with the same two parameters. Results are shown in Table 2. We observe a slight gain due to the fact that SVM can adapt boundaries in a non-linear way, but the accuracy is still comparable with a random choice.

4.4.2. *n*D versus 2D SVM

Global effect. We then trained 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 as shown in Sect. 3, 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 2. We observe that including more than two parameters in the classification results in a significant gain for this sample where C/A alone has accuracy comparable to a random choice. Indeed we almost recover the same accuracy that was obtained on the nearby sample (Table 1).

Robustness. We now try to establish the robustness of this effect. For this purpose, we look at the accuracy of the classification as a function of 3 main properties of the galaxies: luminosity, distance and area (Fig. 6). We progressively add objects and measure 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: $N_{E\rightarrow E}$ ($N_{S\rightarrow S}$).

Several conclusions can be made from this comparison:

- First, using more than two parameters simultaneously clearly increases the global accuracy of the classification in all the redshift, area or luminosity ranges. Indeed, the mean fraction of correct classifications in the 2 dimensional machine

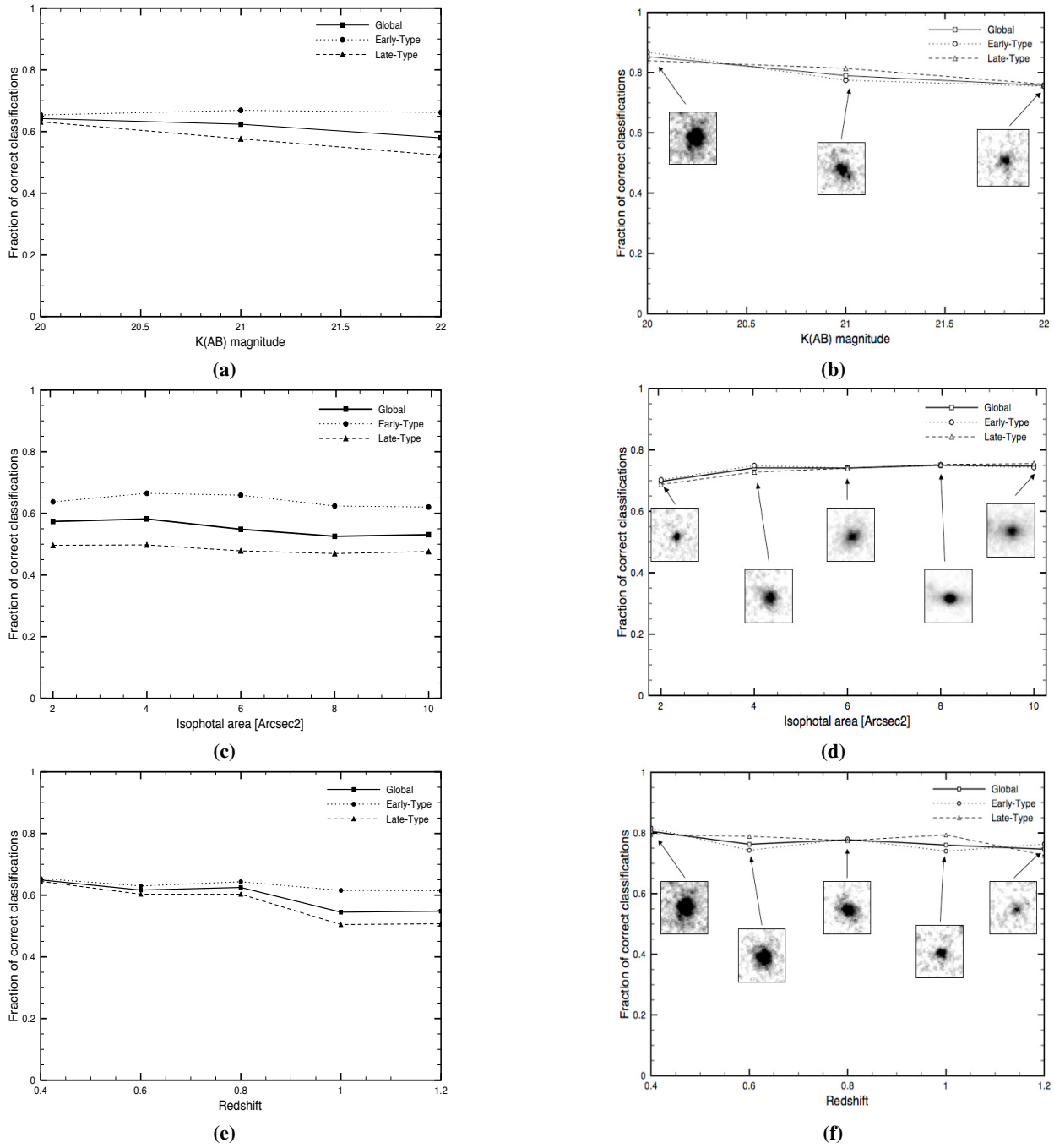

Fig. 6. Cumulative accuracy of classifications for a 2D machine (*left column*) and a 12D one (*right column*) as a function of magnitude (**a)** and **b)**), area (**c**) and **d)**) and redshift (**e**) and **f)**). A sample of ∼1000 galaxies simulated from real SDSS galaxies is used (see text for details). Solid lines show the global accuracy, i.e. the number of galaxies correctly identified, dotted and dashed lines show the fraction of early type and late type galaxies classified correctly, respectively. Stamps in the right column show a typical galaxy for every magnitude, area and redshift range.

is around ∼60% and decreases to ∼50%, which means that there is a high contamination in the C/A plane, whereas it rises to more than ∼80% when using a 12 dimensional machine, which is comparable to what is obtained in space observations (Brinchmann et al. 1998; Menanteau et al. 2006).

– Second, the gain is even higher when looking at the $N_{E\rightarrow E}$ and $N_{S\rightarrow S}$ coefficients. For the C/A classification, there is indeed an asymmetric response of the machine: early type galaxies are better identified (∼65%) whereas the fraction of late type is significantly lower (∼50%), which means that an important fraction of late type galaxies are classified as early type. This will result, when doing the classification, in an important bias towards too a high fraction of elliptical galaxies. However, in the 12 parameter classification, the accuracies are almost perfectly symmetric for the two morphological types.
– Third, when looking at the evolution as a function of distance, size and luminosity, the 12 dimensions machine results in a more stable response, in particular as a function of magnitude and redshift.

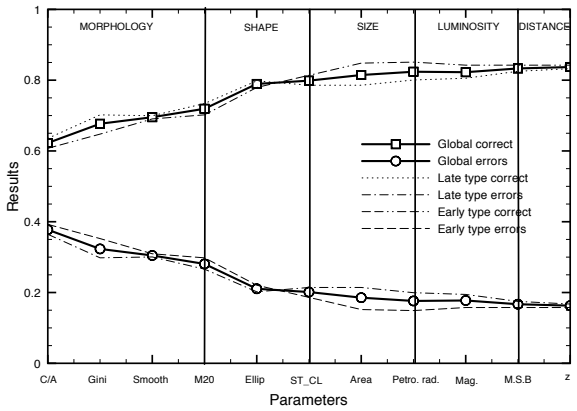

Fig. 7. Accuracy of the classification as a function of the number of parameters. The first point corresponds to a classical C/A classification and each new point adds a dimension. Parameters are classified in 5 classes: morphology, shape, size, luminosity, and distance.

4.4.3. How to fix the number of parameters?

In Sect. 4.4.2 it is shown that the use of more than 2 dimensions to obtain morphology clearly increases the accuracy of the global classification. However, this raises two questions. Are all these parameters necessary? Might some parameters introduce a degeneracy and consequently reduce the machine's accuracy?

To try to answer these questions we make a single test that consists in training several machines with an increasing number of parameters. We thus start with a classical 2 parameter machine (C and A) and we progressively add dimensions until we reach the 12 dimensions described in the previous section. Results are plotted in Fig. 7. As above, we plotted the global accuracy and the accuracy per morphological type. The dimensions are separated into 5 categories (morphology, shape, size, luminosity and distance).

Two important points arise at first sight. First, not all the parameters bring the same amount of useful information. The morphology and the shape carry practically the amount of information necessary to reach 80% accuracy. Second, the accuracy is a monotonic function of the number of parameters; adding a parameter can result in an almost unchanged accuracy (for instance, the magnitude) but never reduces it. This is particularly important, since it means that including more parameters than necessary does not result in a degeneracy. In addition, adding dimensions does not result in a significant increase of the computing time.

4.4.4. Influence of the training set

The method we adopted here for building the training set aims to reproduce the observing conditions and physical properties of the sample in order to reduce errors due to the difference between the training and the science samples. The machine is therefore trained to solve a specific problem and should be trained differently for every science sample. We now measure the importance of this effect by simulating the same sample as if it was observed by the adaptive optics system NACO installed on the VLT. We use NACO data that have been observed in the K_s band with 2 to 3 h exposure time for each pointing (Huertas-Company et al. 2007). The total area covered by these data reaches 7 arcmin^2 and the mean resolution is $0.1''$. We therefore repeated the same procedure but dropped the simulated catalogue in a real NACO background. We then trained the machine with this sample and tried to classify the WIRCam simulated sample with the trained machine.

Table 3. Accuracy of the classification when using a machine trained with a sample with different properties than the science sample – see text for details.

	WIRCam model	NACO model
Global	0.83	0.62
$N_{E\rightarrow E}$	0.81	0.96
$N_{S\rightarrow S}$	0.84	0.24

Results are shown in Table 3. The global accuracy of the classification falls to 62%, i.e. 40% of contaminations when using the NACO model to classify WIRCam galaxies. In particular, there is a systematic drift from late to early type galaxies. The training set must therefore be carefully built to take into account all the observing conditions.

5. Summary and conclusions

We have presented a new method to perform morphological classification of cosmological samples based on support vector machines. This method 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. Furthermore, since the calibration sample is built from a nearby sample, visually classified and adapted to reproduce the physical and instrumental properties, the method can even be employed on seeing-limited data. Selecting the calibration sample in the high redshift sample's rest-frame makes the results robust towards wavelength dependent effects and makes it easier to interpret them in terms of evolution.

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 ($K_{\rm AB} \sim 22$). Furthermore, the accuracy is a monotonic function of the number of parameters.

The presented method is intended as a framework for future studies. In particular, it can be used to look for luminosity and color evolution as a function of the morphology. However this method is far more general and can be appplied to many other samples of galaxies observed with ground-based data, with or without AO correction. Several applications are be intended in order to study the effects of local environment and galaxy density on the morphological evolution of galaxies both in the field and in rich clusters of galaxies. The library is available for download at `http://www.lesia.obspm.fr/~huertas/galsvm.html`

Acknowledgements. The authors want to thank the referee Dr. R. Abraham for his useful suggestions that helped to improve this paper.

References

Abraham, R., van den Bergh, S., Glazebrook, K., et al. 1996, ApJS, 107, 1
Abraham, R. G., van den Bergh, S., & Nair, P. 2003, ApJ, 588, 218
Arnouts, S., Walcher, C. J., Le Fevre, O., et al. 2007, ArXiv e-prints, 705

Baugh, C. M., Cole, S., & Frenk, C. S. 1996, MNRAS, 283, 1361
Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Brinchmann, J., Abraham, R., Schade, D., et al. 1998, ApJ, 499, 112
Chang, C.-C., & Lin, C.-J. 2001, LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm
Cole, S., Lacey, C. G., Baugh, C. M., & Frenk, C. S. 2000, MNRAS, 319, 168
Conselice, C. J., Bershady, M. A., Dickinson, M., & Papovich, C. 2003, AJ, 126, 1183
Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886
de Vaucouleurs, G. 1948, Annales d'Astrophysique, 11, 247
Fall, S. M., & Efstathiou, G. 1980, MNRAS, 193, 189
Hubble, E. 1936, ApJ, 415
Huertas-Company, M., Rouan, D., Soucail, G., et al. 2007, A&A, 468, 937
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006a, A&A, 457, 841
Ilbert, O., Lauger, S., Tresse, L., et al. 2006b, A&A, 453, 809
Lilly, S. J., Le Fevre, O., Hammer, F., & Crampton, D. 1996, ApJ, 460, L1
Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163
Madau, P., Pozzetti, L., & Dickinson, M. 1998, ApJ, 498, 106
Menanteau, F., Ellis, R. S., Abraham, R. G., Barger, A. J., & Cowie, L. L. 1999, MNRAS, 309, 208
Menanteau, F., Ford, H. C., Motta, V., et al. 2006, AJ, 131, 208
Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266
Sandage, A. 1961, The Hubble atlas of galaxies (Washington: Carnegie Institution)
Scarlata, C., Carollo, C. M., Lilly, S. J., et al. 2006, ArXiv Astrophysics e-prints
Scoville, N. Z., & COSMOS Team. 2005, BAAS, 1309
Simard, L., Willmer, C. N. A., Vogt, N. P., et al. 2002, ApJS, 142, 1
Tasca, L., & White, S. 2005 [arXiv:astro-ph/0507249]
Vapnik, V. 1995, The nature of statistical learning theory, 536 (Springer-Verlag)
Wolf, C., Meisenheimer, K., Rix, H.-W., et al. 2003, A&A, 401, 73
Wu, K. L.-K. 1999, Ph.D. Thesis, AA (University of California, Santa Cruz)
Zucca, E., Ilbert, O., Bardelli, S., et al. 2006, A&A, 455, 879