The optimized library of synthetic galaxy spectra for Gaia

Karampelas A.[1], Kontizas E.[2], Bellas-Velidis I.[2], Kontizas M.[2]

[1] *Department of Astrophysics, Astronomy & Mechanics, Faculty of Physics, National and Kapodistrian University of Athens, Greece*
[2] *Institute for Astronomy and Astrophysics, National Observatory of Athens, Greece*

Abstract: We present the optimized library of synthetic galaxy spectra. This library will be used for the Gaia satellite observations of unresolved galaxies. These galaxy spectral templates are useful for the optimal performance of the unresolved galaxy classifier (UGC) software. The UGC will assign spectral classes to the observed unresolved galaxies by Gaia (classification) and estimate some of their intrinsic astrophysical parameters, which were used to create the synthetic library (parametrization). We also present the classification and parametrization results using the Gaia-simulated version of the optimized library of synthetic galaxy spectra. To optimize our synthetic library, we applied the principal component analysis (PCA) method to our synthetic spectra and studied the influence of the star-formation rate parameters on the spectra, and how these agree with some typical characteristics of the galaxy spectral types. We then used support vector machines (SVM) to classify and parametrize the optimal simulated spectra. In the final set of synthetic spectra, overlaps in spectral energy distributions and colors are highly suppressed, while the results of UGC classification are improved.

1. Introduction

ESA's cornerstone Gaia mission is going to repeatedly observe a billion astrophysical objects of the entire sky during the next few years. The faintest objects that are expected to be observed will have approximately a G = 20 Gaia magnitude (unfiltered light), which corresponds to a limit of V = 20−25 mag, depending on the spectral type (Jordi et al. 2006). The final observational data will include low-resolution spectrophotometry of millions of unresolved galaxies. This will be performed with Gaia's spectrophotometer, a slit less prism spectrograph with a blue (BP) and a red (RP) channel over the wavelength range between 330 nm and 1000 nm (Jordi et al. 2010). The classification of these galaxies into spectral classes and the prediction of some of their significant astrophysical parameters (Kontizas et al. 2011) are among the goals of the Gaia mission.
A corresponding software package (unresolved galaxy classifier, UGC) to accomplish this task is under development (Bellas-Velidis et al. 2010). The UGC uses Gaia-simulated synthetic galaxy spectra as templates, learning to successfully predict their spectral classes (classification) and the values of some significant astrophysical parameters (regression). Classification and regression is performed by support vector machines (SVM, Vapnik 1995).

An extended library of synthetic galaxy spectra has already been produced (Tsalmantza et al. 2009) with the PÈGASE.2 model (http://www.iap.fr/pegase) of galaxy spectral evolution (Fioc & Rocca-Volmerange 1997; Fioc & Rocca-Volmerange 1999; Le Borgne & Rocca-Volmerange 2002). This model uses the stellar evolutionary tracks of the Padova group, extended to the thermally pulsating asymptotic giant branch (AGB) and post-AGB phases (Groenewegen & de Jong 1993), and the BaSeL 2.2 library of stellar spectra to produce low-resolution (R~200) ultraviolet to near infrared synthetic spectra of galaxies. Each spectrum represents a specific evolutionary scenario, the latter including a star formation rate (SFR) law, an initial mass function (IMF), etc.

The spectral library produced corresponds to four spectral types (early type, spiral, irregular and QSFG – quenched star-forming galaxies) at various random redshift values. The synthetic spectra satisfactorily cover the (g - r) – (r - i) color–color diagram of the DR4 SDSS galaxies. They have also been simulated for Gaia's BP and RP photometers (description in Sordo & Vallenari 2009), with the addition of reddening, using the extinction law by Fitzpatrick (1999) and noise, for three G-band magnitude values (G = 15, G = 18.5, and G = 20). Support vector machines have been applied to the simulated spectra. Currently, an extensive comparison of the synthetic spectra with observed SDSS spectra is under development, leading to a semi-empirical library of galaxy spectra (Tsalmantza et al. 2012), combining real SED with information about galaxy parameters from galaxy evolution modelling.

All these numerous multi-parameter models have to be tested additionally for overlaps and for how well they represent realistic spectral classes. The quality of these spectra directly affects the efficiency of the UGC. The principal components analysis (PCA) method is able to compress the valuable information and reveal the significant correlations among the data entries in these huge databases. This is the reason why the PCA has become a very popular tool for determining the abundance of large-extent observations or simulations.

2. The library of synthetic galaxy spectra

The synthetic spectra have been constructed by using two different star-formation law (SFL) scenarios: exponential star-formation rate (SFR) for early-type galaxies, and SFR proportional to the mass of the gas for spiral galaxies, irregular galaxies, and QSFG. These SFL were then modeled by varying the corresponding SFR parameters (p_1, p_2 for the exponential SFR and p_1, p_2, p_3, t_{infall} for SFR proportional to the mass of the gas), at a fixed galaxy age (13 Gyr for early-type and spirals, 9 Gyr for irregulars and QSFG). The parameters p_1 and p_2 of the exponential SFR scenario do not correspond to the parameters p_1 and p_2 of the SFR scenario that is proportional to the mass of the gas. Their range was determined in a way to produce realistic synthetic spectra. For more details about the library of synthetic galaxy spectra, see Tsalmantza et al. (2009).

For the purposes of the UGC, the synthetic library should contain a considerable variety of spectra, which should be as typical as possible and, at the same time, as realistic as possible. This way, the relevant software would be able to classify and parameterize the spectra of unknown galaxies that Gaia will observe in a more efficient way.

This library contains spectra with realistic colours. However, a wide range of SFR parameters for different SFL and ages were used to produce SEDs with a huge variety of continua shapes and emission line strengths. Moreover, it is not clear how the simultaneous variation of two, three, or even four SFR parameters affects the shape of the produced SED, for example:

a) two different sets of SFR parameters of the same SFL can result in similar spectra;
b) this similarity in spectra could possibly also occur for two different sets of SFR parameters assuming two different SFL, i.e. a spiral galaxy similar to an early type one;
c) the various spectral types could contain a considerable amount of non-normal galaxy spectra.

To optimize the spectral library and use it for the purposes of a learning-based algorithm like UGC, it is important to know how the various sets of SFR parameters affect the shape of the resulting spectra. This information could be used to identify duplicated spectra of the same spectral type, spectral overlaps between different spectral types, and non-normal spectra. The corresponding suppression of these spectra could improve the classification and regression efficiency of UGC. Additionally, a better understanding of the modelling would be obtained for the relation between input parameters and output spectra. This would ensure a more productive future usage of the PÈGASE code.

Of course, in reality spectral overlaps between different spectral types as well as more complex cases such as mergers, are to be expected, which will make the classification and parameterization more difficult and challenging. However, such a puzzling task would be better addressed with a simple and realistic set of spectral templates.

Each galaxy spectrum can be considered as a point in a multidimensional space, with as many axes as the number of its wavelength bins. In each axis, this galaxy will have the flux value of its corresponding bin. Because a plot like this would be impossible to draw, other methods are required to reduce the dimensionality to just two or three principal dimensions to gain an overview of all spectra simultaneously and analyze the whole library of synthetic galaxy spectra at once. This can be done with the principal components analysis method, as we show in the next section.

3. Principal components analysis for the study of the synthetic spectra

The principal components analysis is part of a family of methods called unsuper-

vised methods. Unsupervised methods are used to visualize data, usually to indicate groups (clustering), or to classify data. They apply when no classes, such as spectral types, are defined a priori or when existing classes are to be confirmed. The PCA provides a linear orthogonal transformation of a data set (e.g. galaxy spectra) into a new base, where a particular characteristic of interest (e.g. the variance of the original data) is preferentially highlighted. The new set of axes onto which the original data are being projected is called the principal components (PCs).

Amongst the numerous relevant publications, we refer the reader to the PCA applications on stellar spectra from the Michigan Spectral Survey (Bailer-Jones et al. 1998) and the SDSS/SEGUE project (Re Fiorentin et al. 2007), and on galaxy spectra from SDSS (Yip et al. 2004), DEEP2 Redshift Survey (Madgwick et al. 2008), the 2dF Galaxy Redshift Survey (Folkes et al. 1999) and the Spitzer Infrared Spectrograph (Wang et al. 2010). The PCA has also been applied to spectroscopic imaging observations (Heyer et al. 1997; Steiner et al. 2009), where spatial information is available. Steiner and collaborators managed to discover the existence and the location of an active nucleus of very low luminosity in the NGC 4736 galaxy. Finally, Ronen et al. (1999) analyzed synthetic galaxy spectra, including PÈGASE spectra, with different ages, star-formation histories and metallic ties, while Tsalmantza et al. (2007) visualized PÈGASE spectra corresponding to Hubble types.

For a (n spectra s_i) × (m wavelength bins) data set, the PCA applies as follows:

(a) either the correlation matrix (standardized PCA) or the variance-covariance matrix (unstandardized PCA) of the data is computed;
(b) the eigenvalues λ and the eigenvectors (eigenspectra) u (principal components) of either the correlation matrix or the variance-covariance matrix are calculated. If $n \geq m$, then m eigenvalues and m corresponding eigenvectors are computed;
(c) the eigenvalues are sorted in decreasing order. The first principal component, u_1 (PC1), corresponds to the first (higher value) eigenvalue λ_1 and accounts for the maximum amount of the total variance of the data. The second principal component, u_2 (PC2), corresponds to the second eigenvalue λ_2, is orthogonal to PC1 and accounts for the second highest variance fraction. Lower-order principal components are found the same way.

Thus, each original spectrum s_i is decomposed onto the new set of axes u (eigenspectra) as

$$s_i = \sum_{k=1}^{m} a_{k,i} u_k \tag{1}$$

where $\alpha_{k,i}$ is the admixture coefficient (the projection of the ith spectrum onto the kth principal component). In many cases, the first few principal components account for practically the total variance of the original spectra. This means that the

most significant PCs can be used to reconstruct the original spectra with high accuracy, thus providing an efficient data compression. The reduced reconstruction $s(r)_i$ of s_i by using the r most significant PCs is

$$s(r)_i = \sum_{k=1}^{r} a_{k,i} u_k \tag{2}$$

If $r \leq 3$ is satisfactory, then the data set can be visualized by two or three-dimensional plots and be further analyzed. The reduced reconstruction can be also used to remove noise and identify unusual spectra. For this reason, the PCA can pre-process data before they are analyzed by a classifier (Bailer-Jones et al. 1998). The above denote the main advantages of PCA: data compression and dimensionality reduction.

We applied the PCA method to the library of 28 885 synthetic galaxy spectra with $z = 0$. Because we aim to retain the relative strengths of the spectral features of the synthetic spectra, we used the unstandardized PCA procedure (see also Steiner et al. 2009). We computed the variance-covariance matrix, where the diagonal elements represent the variances of the flux bins and the off-diagonal elements the covariances between them.

The amount of the total variance and the corresponding cumulative variance of the five most significant eigenvectors are listed in Table 1. The reconstruction error that corresponds to the use of up to a specific eigenvector is also listed in Table 1. This error is the mean absolute percentage error in the total normalized flux. Therefore it is sufficient to consider the first two principal components to accurately analyze the whole library of synthetic galaxy spectra and use them in UGC, because we have a low error of 2% (on average) in spectral reconstruction and a 99% inclusion of the total variance. The data dimensionality is vastly reduced, apparently without significant loss of information.

Table 1 Amount of the total variance of the five most significant eigenvectors, the cumulative variance, and the corresponding reconstruction error for the synthetic galaxy spectra.

Eigenvector	Variance (% of total)	Cumulative variance (% of total)	Reconstruction error (%)
u1	94.61	94.61	6.86
u2	4.25	98.85	1.87
u3	1.05	99.90	0.32
u4	0.07	99.98	0.12
u5	0.02	99.99	0.07

Figure 1 shows the projection of the synthetic galaxy spectra on the first and the second principal components. The representation of all the galaxies on the plane of the first two principal components can help us investigate a) among which spectral

types there are overlaps and b) which sets of SFR parameters cause these overlaps. This investigation can help us optimize the existing library of synthetic galaxy spectra by suitably adjusting the SFR parameters space, and define the context of a future extension through a better knowledge of the modelling. However, this approach has to be implemented in a way that the optimumset of synthetic spectra is still realistic.

In Figure 1, early-type galaxies are distributed toward the lower part of this plot, where the emission-line dominant PC2 is of less importance. The majority of them have negative PC2 values, which decreases the emission line strengths of the PC1 contribution. On the other hand, irregulars and QSFG tend to be distributed toward the upper left part of the diagram, where PC2 is more significant than in the previous case. Spirals show a broader variety of PC1-PC2 combinations.

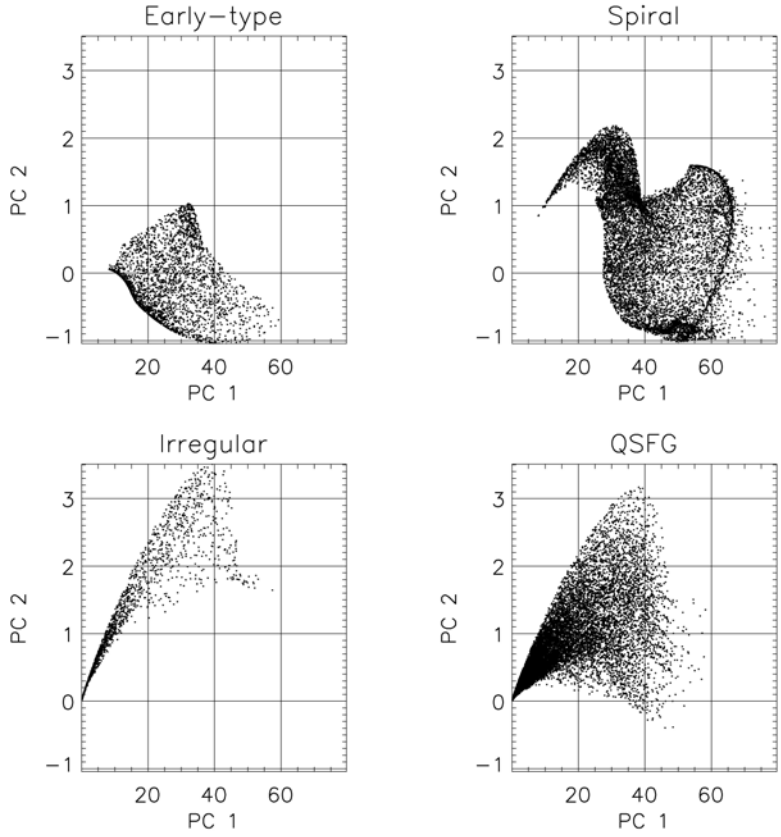

Figure 1 Projection of the synthetic galaxy spectra (Tsalmantza et al. 2009) on the first (PC1) and the second (PC2) principal components for each spectral type.

These rough distinctions together with overlaps between the various spectral types up to some reasonable level are to be expected. However, this figure shows that the overlaps are quite extended. Spiral galaxies are highly overlapped with early type

galaxies and QSFG, while QSFG are also highly overlapped with irregular galaxies. Less overlap occurs between spirals and irregulars and between QSFG and early-type galaxies. No overlap exists between early-type and irregular galaxies.

4. Optimization of the library of synthetic galaxy spectra

The findings of the previous section could explain the classification performance of the scientific tests for the UGC software, using the Gaia-simulated version of the synthetic library. For example, most of the misclassified early-type galaxies are classified as spirals, while not a single early-type galaxy is misclassified to the irregular spectral class. An optimized version of the synthetic library of galaxy spectra with less overlaps could increase the spectral type classification and the parameters regression performance of UGC.

The analysis of the results of the PCA application to the synthetic galaxy spectra implies to truncate a) emission line early-type galaxies that resemble spirals; b) non-emission line low-SFR early-type galaxies; c) spiral galaxies with high p_2 values that resemble QSFG and irregulars; d) irregulars and QSFG with extremely high p_2 and t_{infall} values, similar to each other; and e) QSFG without emission lines that resemble spirals. Table 2 lists the optimized range of the library of synthetic galaxy spectra.

The PCA method was applied to the library of the optimal spectra to investigate the changes it has undergone. The most significant components of the optimum spectra are almost identical to the corresponding PCs of the original library. Figure 2 illustrates the distribution of the optimum spectra to the two most significant PCs, which have a corresponding error of 1% in spectral reconstruction and a 99% inclusion of the total variance. Because the PCs did not change much, the distribution of the optimum spectra on them reveals similar trends like those illustrated in Figure 1. However, the optimization results are evident. Early-type galaxies form a distinct group, separately from spirals, with a desired SFR, and the spiral-irregular and spiral-QSFG overlaps have been limited to a relatively narrow region.

Additionally, the high concentration of late-type galaxies in the lower left part of this plot has been reduced. A notable overlap between irregulars and QSFG is present in the optimized data. This is not surprising, because the corresponding spectra do have many common characteristics. In any case, overlaps between the various spectral types are to be expected in real data.

Clearly, it is not necessarily correct to have galaxy spectra anywhere in a PC1-PC2 diagram. Physical constraint limit the possible spectral diversity, and a robust spectral decomposition like the one provided by the PCA method must reflect these limitations.

The optimum synthetic galaxy spectra are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via
http://cdsarc.u-strasbg.fr/viz-bin/qcat?J/A+A/538/A38.

Table 2 Optimized range of SFR parameters for each galaxy spectral type. The p_1, p_2 parameters correspond to the adopted SFL respectively.

SFR parameter	Optimized range	Units
p_1 (Early-type)	10 – 2 000	Myr
p_2 (Early-type)	0.5 – 1.5	Msolar
p_1 (Spiral)	0.3 – 2.4	---
p_2 (Spiral)	5 – 5 000	Myr/Msolar
t_{infall} (Spiral)	5 – 16 000	Myr
p_1 (Irregular)	0.6 – 3.9	---
p_2 (Irregular)	4 000 – 40 000	Myr/Msolar
t_{infall} (Irregular)	5 000 – 20 000	Myr
p_1 (QSFG)	0.6 – 3.9	---
p_2 (QSFG)	4 000 – 40 000	Myr/Msolar
p_3 (QSFG)	1 – 10	Myr
t_{infall} (QSFG)	5 000 – 20 000	Myr

5. UGC and optimum Gaia-simulated spectra

It is important for the development of the unresolved galaxy classifier to investigate the impact of optimizing the library of synthetic galaxy spectra to its performance. The UGC uses the Gaia-simulated version of the synthetic library. It is an algorithm that is based on the implementation of the supervised learning method SVM. These SVMs (Vapnik 1995) can be used for data classification through the definition of an optimum hyperplane that separates the members of the various classes that describe the data. For this purpose, a set of training data is used to train the SVMs and prepare it to classify data of unknown class. The SVMs can also be used for parameter regression. Again, a set of training data is necessary to train the SVMs and prepare them to predict the parameter values of data that lack this information.

The SVMs are trained to predict the spectral type (classification) and the SFR parameters p_1, p_2, p_3 and t_{infall} values, together with extinction and redshift values (regression).We applied this procedure to two sets of simulated data, the first containing “clean” spectra without any addition of noise, extinction, or redshift, and the second containing noisy, reddened, and redshifted (“realistic”) spectra, for G = 15 Gaia magnitude.

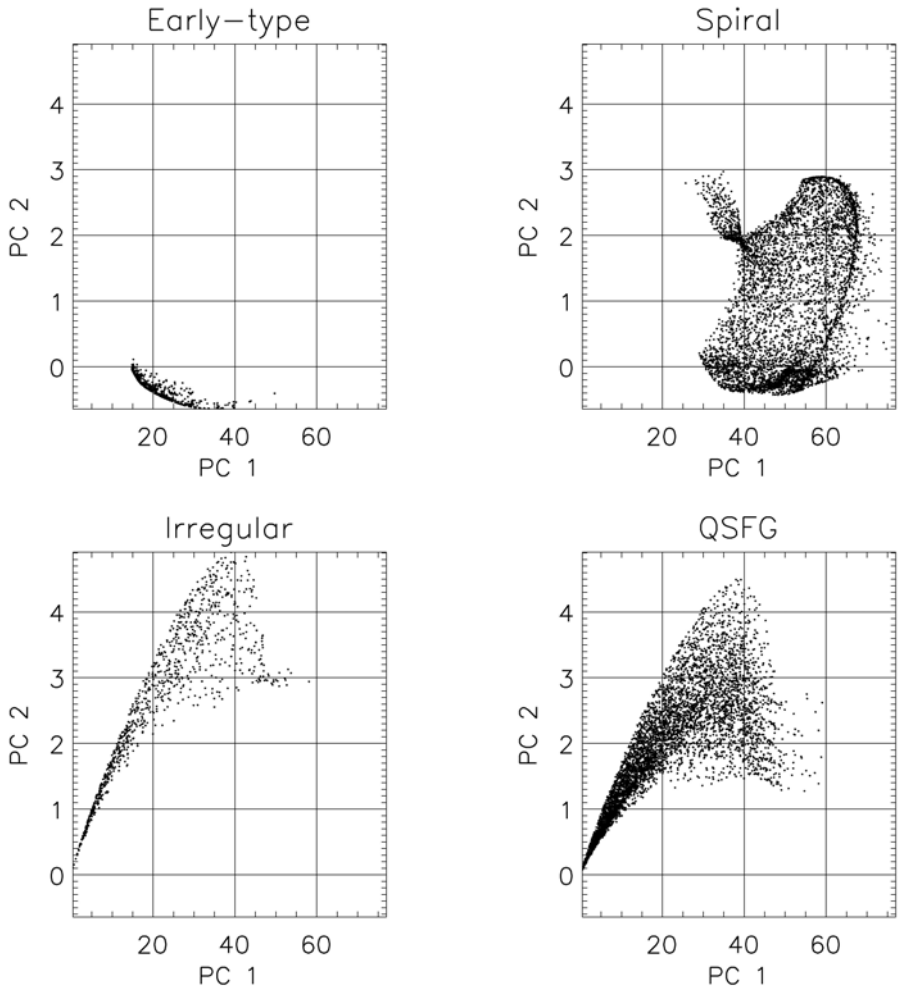

Figure 2 Projection of the optimized synthetic galaxy spectra on the first (PC1) and the second (PC2) principal components for each spectral type.

The classification efficiency (percentage of successful predictions) for the "clean" spectra before and after the optimization is ~100%. The corresponding results for the noisy reddened and red shifted spectra are shown in Figure 3. Spectral type predictions of these spectra are quite successful. Spectral optimization in general improved the classification efficiency of UGC, especially for the early-type galaxies, where the predictions where about 25% more successful. Spirals and irregulars are slightly better classified (~3%), while the optimization practically left the QSFG classification efficiency unchanged, which was already high. These results reflect the suppression of the overlaps between the various spectral types achieved through the optimization of the spectral library.

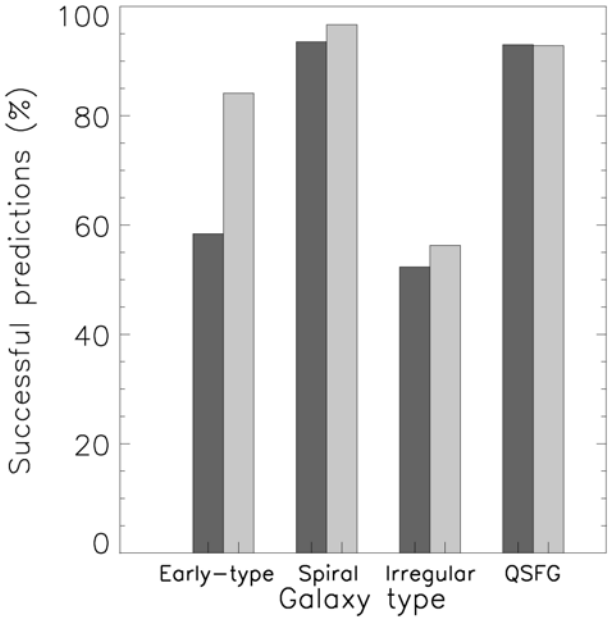

Figure 3 Classification performance of UGC for the noisy, reddened, and redshifted simulated spectra, for each spectral type. Each pair of columns demonstrates the successful predictions of the corresponding spectral type, before (left column) and after (right column) the optimization.

5. Conclusions

We optimized the library of synthetic galaxy spectra (Karampelas et al. 2012), which was produced with PÉGASE.2 code (Tsalmantza et al. 2009), setting new boundaries in the space of the galaxy parameters. The application of the principal component analysis method to this extended library vastly reduced its dimensionality without any significant loss of information and revealed spectral overlaps. It also provided ways to a better understanding of how the multi-parameter modelling affects the final shape of a synthetic spectrum. Additionally, the investigation of the various star-formation laws used in the modelling helped to trace some non-normal synthetic spectra. This investigation led to a set of more realistic synthetic spectra, where overlaps between spectra and spectral colours were highly suppressed. The findings could be used to define the context of a future extension of this spectral library, because a better understanding of the modelling was achieved.

The Gaia-simulated version of this optimum set of spectra was used for training the unresolved galaxy classifier code, which will be part of the Gaia satellite software. The training was performed by applying the support vector machines method. The classification efficiency was in general improved. Advances in the code itself, which is currently under development, could limit the errors even more.

References

Bailer-Jones, C. A. L., Irwin, M., & von Hippel, T. 1998, MNRAS, 298, 361
Fioc, M., & Rocca-Volmerange, B. 1997, A&A, 326, 950
Fioc, M., & Rocca-Volmerange, B. 1999, A&A, 351, 869
Fitzpatrick, E. L. 1999, PASP, 111, 63
Folkes, S., Ronen, S., Price, I., et al. 1999, MNRAS, 308, 459
Groenewegen, M. A. T., & de Jong, T. 1993, A&A, 267, 410
Heyer, M. H., & Schloerb, P. F. 1997, ApJ, 475, 173
Jordi, C., Høg, E., Brown, A. G. A., et al. 2006, MNRAS, 7, 290
Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, A&A, 523, A48
Karampelas, A., Kontizas, M., Rocca-Volmerange, B., et al. 2012, A&A, 538, A38
Kontizas, M., Bellas-Velidis, I., Rocca-Volmerange, B., et al. 2011, EAS, 45, 337
Le Borgne, D., & Rocca-Volmerange, B. 2002, A&A, 386, 446
Madgwick, D. S., Coil, A. L., Conselice, C. J., et al. 2003, ApJ, 599, 997
Re Fiorentin, P., Bailer-Jones, C. A. L., Lee, Y. S., et al. 2007, A&A, 467, 1373
Ronen, S., Aragon-Salamanca, A., & Lahav, O. 1999, MNRAS, 303, 284
Sordo, R., & Vallenari, A. 2009, GAIA-C8-DA-OAPD-RS-004, http://www.rssd.esa.int/llink/livelink/open/2936253
Steiner, J. E.,Menezes, R. B., Ricci, T. V., & Oliveira, A. S. 2009, MNRAS, 395, 64
Tsalmantza, P., Kontizas, M., Bailer-Jones, C. A. L., et al. 2007, A&A, 470, 761
Tsalmantza, P., Kontizas, M., Rocca-Volmerange, B., et al. 2009, A&A, 504, 1071
Tsalmantza, P., Karampelas, A., Kontizas, M., et al. 2012, A&A, 537, A42
Vapnik, V. 1995, The Nature of Statistical Learning Theory, Springer-Verlag Inc., New York
Wang, L., Farrah, D., Connolly, B., et al. 2011, MNRAS, 411, 1809
Yip, C. W., Connolly, A. J., Szalay, A. S., et al. 2004, AJ, 128, 585