SEMI-EMPIRICAL LIBRARY OF GALAXY SPECTRA FROM SDSS

prepared by: P. Tsalmantza, M. Kontizas, C. A. L. Bailer-Jones, I. Bellas-Velidis, E. Kontizas, E. Livanou, B. Rocca-Volmerange, R. Korakitis, A. Dapergolas, A. Karampelas, M. Fioc, A. Vallenari

approved by:
reference: GAIA-C8-TN-UOA-PAT-003-1
issue: 1
revision: 0
date: 2008-02-28
status: Issued

Abstract

This document describes the construction of a semi-empirical library of galaxy spectra that will be used in cycle 4 of simulations. The library consists of approximately 33600 spectra from SDSS which were compared through χ^2-fitting with the second library of synthetic galaxy spectra (GAIA-C8-TN-UOA-PAT-002-1) and found in good agreement with them. Based on this comparison the observational spectra were extended to the Gaia wavelength range and known astrophysical parameters were assigned to them.

1 Introduction

The performance of the classification and regression algorithms built for Gaia observations depends a lot on the spectral libraries used to train them. In the case of galaxies two libraries of synthetic spectra (GAIA-C8-TN-UOA-PAT-001-1, Tsalmantza et al. 2007, GAIA-C8-TN-UOA-PAT-002-1, Tsalmantza et al. 2008) have already been produced with PEGASE.2 code (Fioc & Rocca-Volmerange 1997). The first library included a small number of typical spectra of 7 Hubble types while the second library was more complete containing a large sample of spectra of 4 types. In figure 1 we present the data of those libraries on the color-color diagrams of SDSS galaxies.

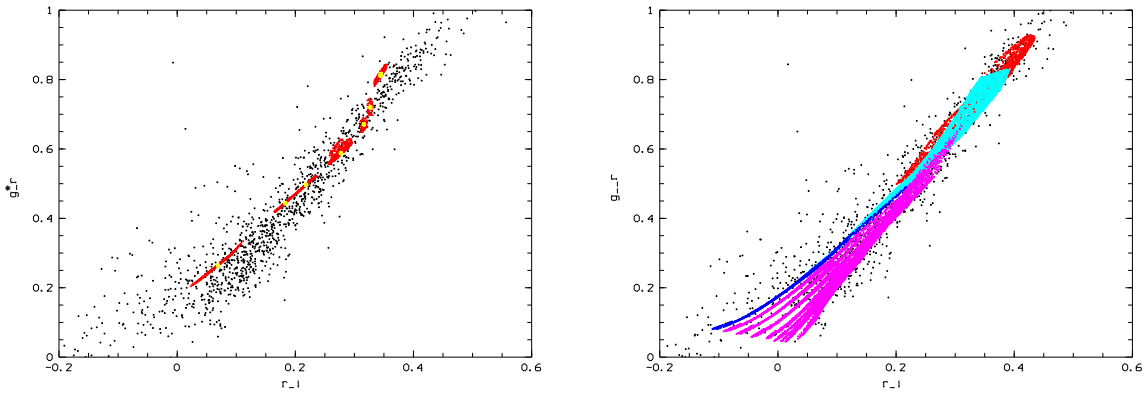

Figure 1: LEFT: The first library of synthetic galaxy spectra. The SDSS galaxies, the galaxies produced in the first library and the typical synthetic spectra of PEGASE.2 are presented with black, red and yellow dots respectively. RIGHT: Random models of irregular (blue), starburst (magenta), spirals (light blue) and early type galaxies (green). Black dots are SDSS galaxies and green the 8 typical synthetic spectra of PEGASE.2.

Even though the synthetic libraries are in very good agreement with observational data an empirical library in the procedure of classification and parametrization is very important. The advantage of an empirical library is that it provides a set of real observed spectra, with the corresponding astrophysical parameters defined from comparison of each spectrum with the synthetic library. A library of observed galaxy spectra combined with the already produced synthetic libraries will improve our results, check the reliability and completeness of our libraries and extend our research by using it in new ways such as the performance of unsupervised methods.

For the construction of an empirical library we decided to use spectra from SDSS since it is the best and the most complete galaxy survey at the moment. The only problem with the spectra of SDSS is that they cover a smaller wavelength range than the one needed by the Gaia simulator. To solve this problem we decided to extend the ends of the observational spectra with synthetic spectra of our second library. In this way we had also the opportunity to compare our library with real spectra. We describe the selection of the SDSS spectra and the procedure we followed

in order to extend them to Gaia wavelengths. Closing this report we present a comparison between the types of galaxies provided by SDSS and the types obtained for the observational spectra through their comparison with the synthetic galaxy spectra of the second library.

2 The library of observed spectra of galaxies

2.1 The selection of the observed spectra of galaxies

The empirical library is derived from the 5th Data Release (DR5) of SDSS spectra. This data release includes observations for 552156 galaxies among which not all of them are suitable for Gaia purposes. To have a better opinion about those galaxies and decide our selection criteria we built histograms for the 12 main parameters which we obtained from the SDSS catalogues. The distributions of those parameters are given in figure 2. In those diagrams we can see that SDSS observations concern galaxies of various luminosities and at different distances. This implies that many of the galaxies in this sample are extended objects and therefore may not be observed by Gaia.

One of the main problems of the observed spectra is that very few astrophysical parameters are known for the galaxies from which they are obtained. In the catalogues of SDSS only a rough classification is given. This classification is based on the paper of Yip et al. (2004). According to this paper galaxies in SDSS with eClass$<$-0.1 are considered to be early type galaxies while galaxies with eClass$>$-0.1 late types. The index eClass is provided by the SDSS catalogues and it allows us to have a rough knowledge of the galaxy types. In figure 2 we can see the distribution of this index for the SDSS galaxies.

Another criterium used in SDSS for the classification of the observed galaxies is the one given by Strateva et al. (2001). According to this paper galaxies in SDSS having a concentration index smaller or larger than 2.6 are expected to be late or early type galaxies respectively. From the distribution of the concentration index C in figure 2 we see, as in the case of the eClass, that the SDSS observations cover a wide range of galaxy types. For all the galaxies in our sample we have obtained the informations that are related with the type of the galaxy.

Unfortunately neither classifications presented above are very strict, so they can be used only for a rough classification of our sample. Additionally those two criteria are not in a very good agreement (figure 3).

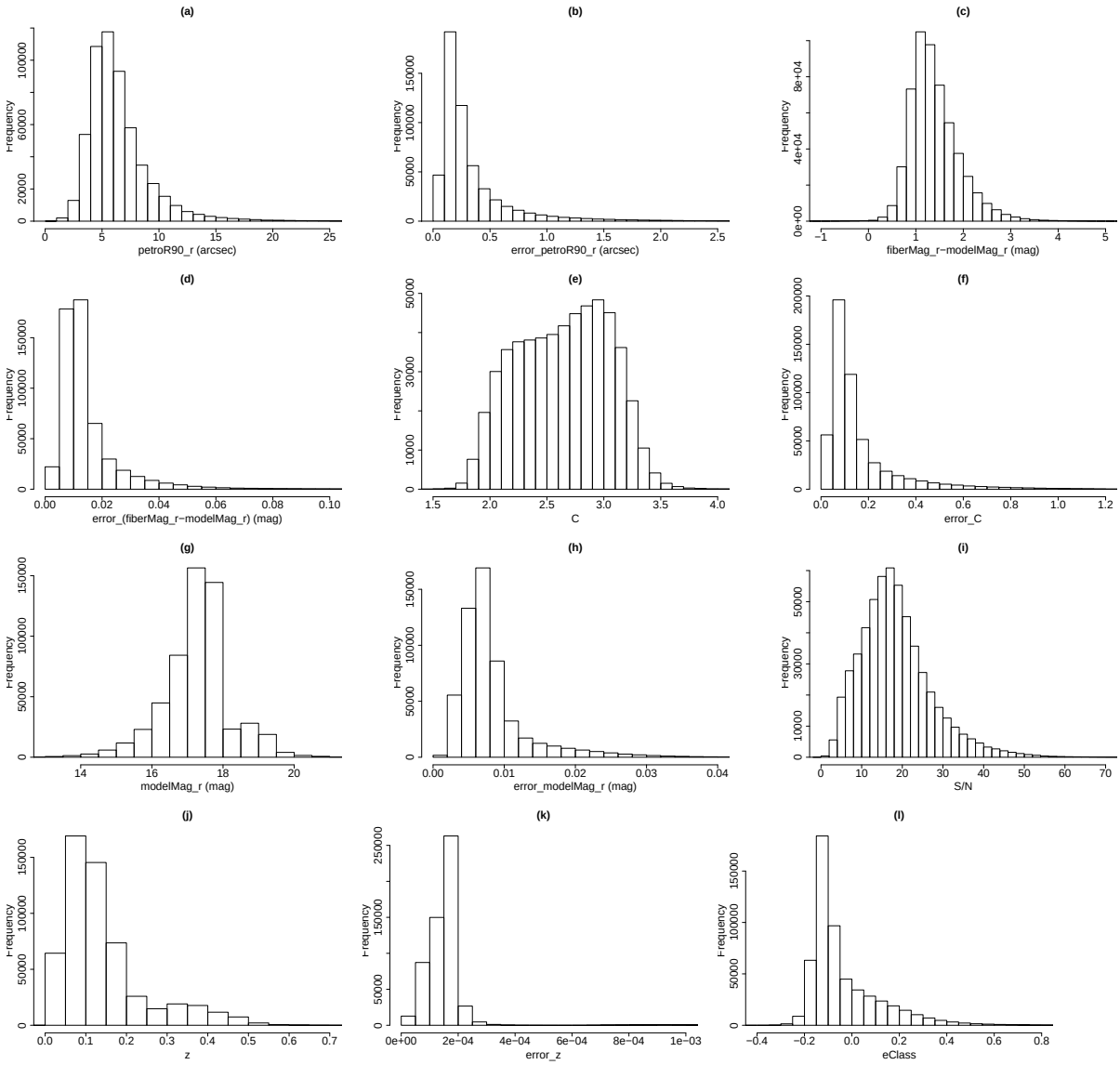

Figure 2: Distributions of the SDSS (DR5) galaxies for the parameters **(a)** radius measure (R90), **(b)** error in R90, **(c)** difference between fiber and model magnitudes in r band, **(d)** error in difference between Fiber and Model magnitudes in r band, **(e)** concentration index C and **(f)** error in concentration index C, **(g)** model magnitude in r band, **(h)** error in model magnitude in r band, **(i)** S/N, **(j)** redshift, **(k)** error in redshift and **(l)** the index eClass.

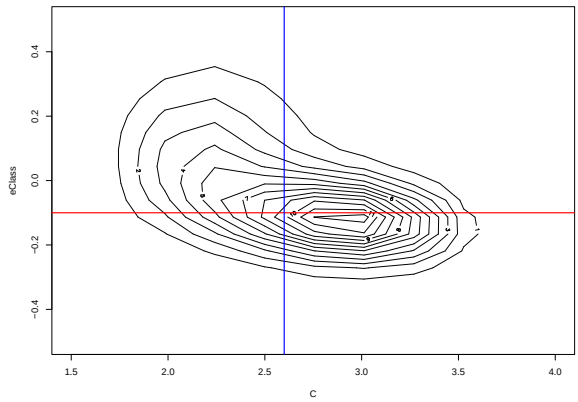

Figure 3: The concentration index C vs the index eClass for the whole sample of SDSS galaxies.

Our main concern was to select spectra with small errors that represent the whole area of the galaxies observed. To achieve this, we chose spectra with S/N>16 and we used three criteria to make sure that the spectra correspond to a large area of the galaxy. i) Knowing that the fiber of SDSS has a diameter of 3 arcsec we demanded the radius (R90) that includes 90 % of the (petrosian) flux of a galaxy in the r band, which corresponds to almost the whole area of the galaxy, to be less than 4 arcsec. In this way almost all the light of a galaxy is observed by the fiber and included in the spectrum. ii) The radius R50 is comparable with the fiber diameter only for galaxies with redshift larger than 0.04 (Strauss et al. 2002). For galaxies with smaller distances, small compact areas inside the galaxies are considered as independent galaxies and therefore the SDSS photometry as well as the radius R90 are not representatives of the whole galaxy. For that reason we excluded from our sample all galaxies with z<0.04. iii) Additionally we kept only galaxies with difference between fiber and model magnitudes in the r band less than 1 mag.

Having selected only galaxies with dimensions small enough so the spectrum comes from the whole galaxy, we apply more criteria to exclude observations with large errors. More specifically we kept only galaxies with errors in fiber and model magnitudes in r band less than 0.01 mag and error in concentration index C less than 0.15. In figure 4 we present step by step the distribution of every parameter after the application of the previous criteria. The final sample contains 33670 spectra of galaxies that cover the whole range of redshifts and types of the SDSS observations. In figure 5 we can see that the disagreement between the two classification criteria (index C and eClass) remains in our final sample.

2.2 The extension of the observed spectra of galaxies

As we have already mentioned in the introduction the extension of the observed spectra at Gaia wavelengths was made by using the 28885 synthetic spectra of the second library produced at a random grid of parameters. In order to find the synthetic spectrum that is in best agreement with

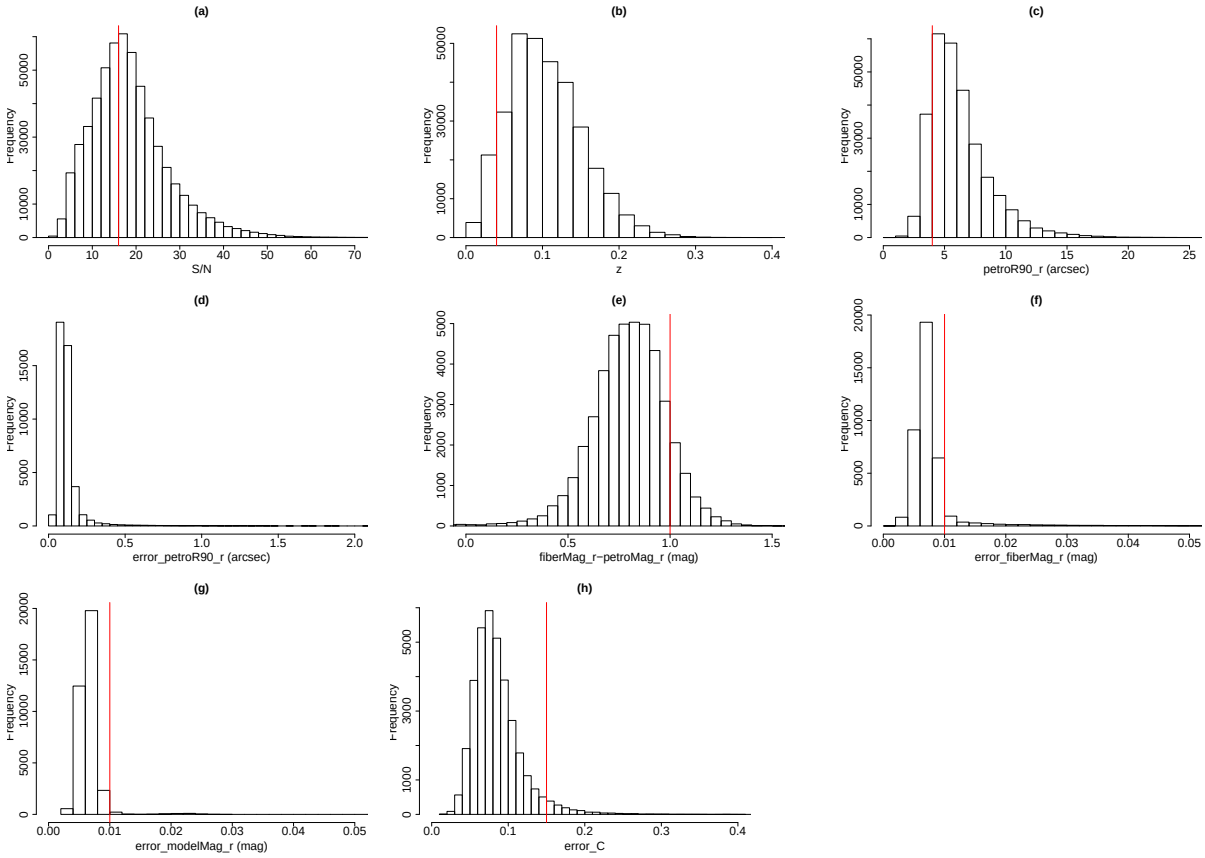

Figure 4: Distributions for the parameters **(a)** S/N, **(b)** z, **(c)** R90, **(d)** error in R90, **(e)** difference between fiber and model magnitudes in the r band, **(f)** error in model and **(g)** fiber magnitudes in the r band and **(h)** error in the concentration index C for the DR5 galaxies of SDSS after the sequential application of criteria. The red line in each plot represents the selection criterium applied for each parameter.

each observed spectrum we first had to make the two libraries compatible. The main differences between them are the effects of reddening, noise and redshift that were included in the SDSS spectra and were absent at the synthetic ones. The first two effects could not be removed from the spectra and therefore will have an impact in our results. However in the case of noise we have selected only spectra with high S/N, so noise issues are limited.

On the other hand redshift must be corrected before making the comparison between the observed and the synthetic spectra. An example of this correction which was made by keeping the energy constant in each spectral bin while relabeling the wavelength axis is given in figure 6.

The next step was to rebin the SDSS spectra in order to reduce their resolution to the one of PEGASE spectra. An example of this correction which was made by considering that the flux is constant in each spectral bin is given in figure 6.

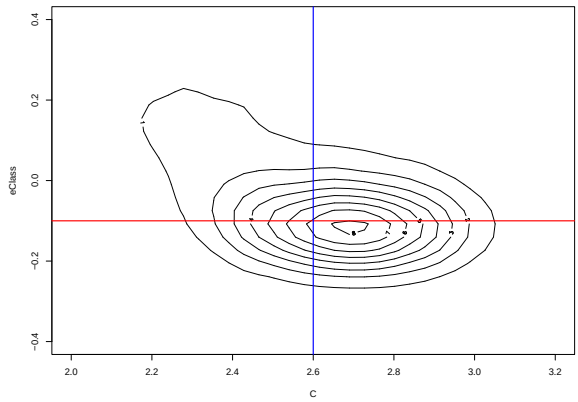

Figure 5: The concentration index C vs the index eClass for the whole sample of the 30670 SDSS galaxies of our final sample.

Finally we normalized the fluxes by dividing the whole spectrum by the mean flux (or luminosity) between the wavelengths 5490 and 5510 Å.

Having done all the above we were ready to perform a χ^2-fitting between the two libraries of galaxy spectra. Every one of the 33670 observed spectra was compared with the whole sample of the 28885 spectra of our library. The comparison was made by masking the areas where the strongest emission lines occur and the edges of the spectra. In figure 7 we present the areas that were masked during the comparison and the results of the reduced χ^2-difference between each SDSS spectrum and the synthetic spectrum with which they were found in best agreement.

From the distribution of the differences between the two libraries we see that in many cases the value of the square difference is quite high. To test if this is due to systematic errors during χ^2-fitting we have plotted the mean square difference for all the spectra of SDSS with the best fitted synthetic spectrum at each wavelength (figure 8). From figure 8 it is obvious that the differences are larger at the wavelengths where emission lines which were not excluded from our comparison occur.

This implies that even in the cases that the value of the difference is large this might be due to strong emission lines that are included in the observed spectra and missing from the synthetic ones. In figure 9 we present six examples of the comparison within the whole range of values of difference.

From those figures we can see that even in the cases with larger differences the fitting of the continuum is very good. This is not true for cases with square difference greater than 15 where the fitting mainly near the 4000 Å discontinuity is poor due to problems either in the synthetic or the observational spectra. However as it is obvious from figure 7 the large majority of our sample was fitted with square differences less than 15 and therefore the agreement between the

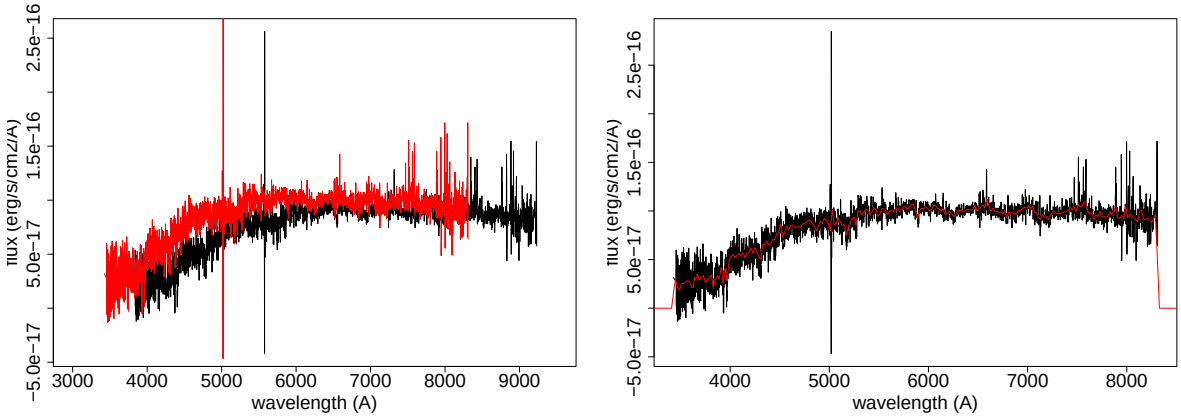

Figure 6: LEFT: Example of an SDSS galaxy spectra (51691-0342-384) before (black) and after (red) the correction of redshift.RIGHT: Example of the same spectra before (black) and after (red) the correction of the resolution.

observed and synthetic spectra is very good. We should also keep in mind that the spectral areas that are going to be extended with synthetic edges are small and Gaia will have very small efficiency at that wavelengths. For that reason the extension of the observed spectra does not need to be very accurate.

The good performance of the χ^2-fitting allowed us to proceed with the extension of the observed spectra. The extension took place at the original spectra of SDSS, i.e. as they were before the correction of redshift, change of the resolution and normalization. To do so we had to apply the observed redshift to every synthetic spectra that found in best agreement with each observational one. After this step we also had to multiply the normalized fluxes of the synthetic spectra with the flux used to normalize the observed one.

The last step before the connection was to change the resolution of PEGASE spectra at the points were the connection would occur (i.e. 3792.2765 Å and 9236.3419 Å). To succeed in this we linearly interpolated the fluxes of synthetic spectra at the wavelength range between 3000 and 11000 Å with a step of 10 Å at the synthetic edges and a step that was equal to the step in the real spectra at the middle. Between those two areas we added two points so that the pass from the parts with low to the ones with high resolution to be smoother. In figure 10 we see an example of a synthetic spectrum after the application of the above procedure.

From figure 10 we see once again that the agreement between the real and the synthetic spectra is very good (see upper left figure 9 for comparison).

The last step was to extend the real spectra with the synthetic edges and change the units of flux and wavelength to the ones needed by the Gaia simulator, ($W/m^2/nm$ and nm respectively). The header of each spectrum of this library provides the values of all the parameters obtained

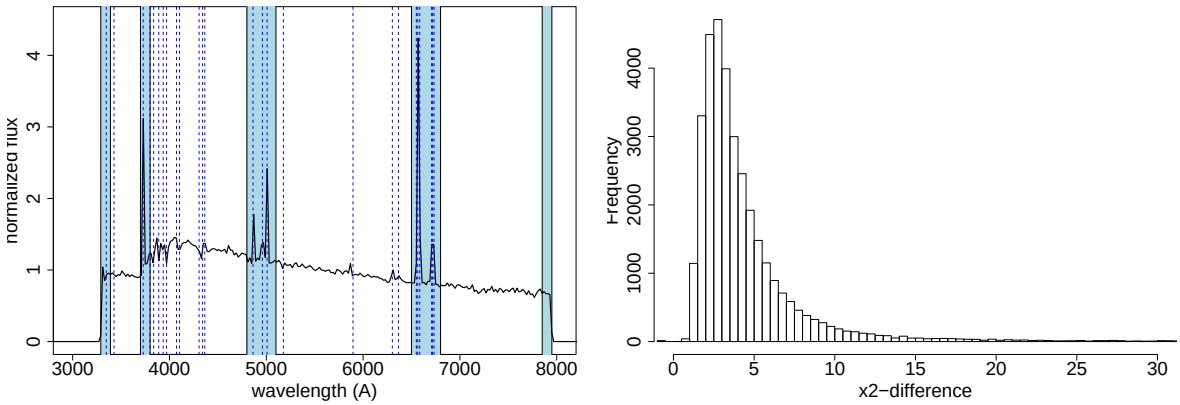

Figure 7: LEFT: The areas of the spectra that were masked during the χ^2-fitting. RIGHT: Distribution of the difference between the observed and the adopted synthetic galaxy spectra.

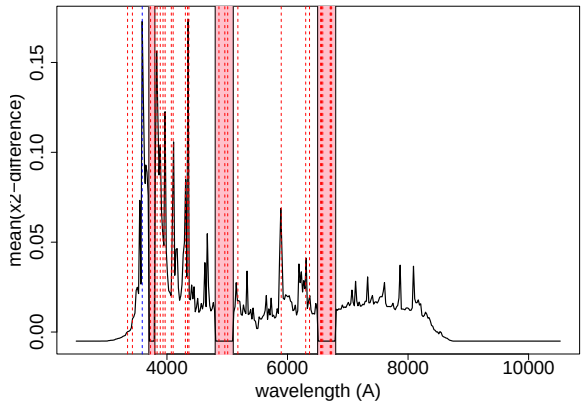

Figure 8: The mean square difference for all the spectra of SDSS with the best fitted synthetic spectrum for every wavelength.

by the SDSS catalogues for each spectrum as well as the ones extracted by PEGASE.2 code for the synthetic spectrum used to extend it. Those parameters are listed in table 1.

In this way we managed to construct an empirical library of 33670 SDSS spectra which will be a very useful tool both for the classification and regression of Gaia observations. Even though we do not have knowledge of the astrophysical parameters of the observed galaxies they are known for the synthetic spectra produced by PEGASE and therefore they can be used for parametrization purposes. However we should not forget that in the comparison of the two types of spectra we did not take under consideration the emission lines and therefore this library cannot be used for the extraction of parameters that are related to them (e. g. SNI and SNII rate). On the other hand regression methods can be used for this library for the parameters related with the continuum within some errors.

2.3 The type of galaxies in the empirical library of galaxy spectra

In the header of each observed spectrum of the empirical library we provide three numbers indicating the type of the galaxy: the eClass, the C=R90/R50 (both from the SDSS catalogue) and the T type given by our classification in the PEGASE library of synthetic spectra. In this paragraph we are going to compare the results of the SDSS classification with the ones extracted by the χ^2-fitting with the second library of synthetic spectra. In figure 11 we show the results of the comparison.

In these plots we see that our classification based on the second library of synthetic spectra gives quite good results. For the large majority of the galaxies classified by SDSS as early types (at least for 90% of them), our classification gives the same result. The few galaxies classified as spirals are not in disagreement with SDSS classification since in those catalogues Sa galaxies are considered as early types while in our library as spirals.

The results are a little poorer in the case of late type galaxies were a little more than 2237 of them (21.56% of our sample) have been classified by our method as early types. This was expected since in our classification we have excluded the emission lines which are very important features for this kind of galaxies.

These results show that the synthetic spectra of our library describe quite well the different types of galaxies that exist. On the other hand as we have already shown in figure 3 the two criteria used for the classification by SDSS are not very strict and the diagrams in figure 11 are rough estimates of the real types. However a combination of all the three classification criteria described here can give us enough safety at the knowledge of the types in order to check the performance of our classification algorithms.

References

Fioc M. & Rocca-Volmerange B. 1997, A&A, 326, 950

Strauss M. A., Weinberg D. H., Lupton R. H. et al., 2002, AJ, 124, 1810

Strateva I., Ivezic Z., Knapp G. R., Narayanan V. K., Strauss M. A., Gunn J. E., Lupton R. H., Schlegel D., Bahcall N. A., Brinkmann J. and 19 coauthors, 2001, AJ,, 122, 1861

Tsalmantza P., Kontizas M., Korakitis R., Rocca-Volmerange B., Kontizas E., Livanou E., Bellas-Velidis I., Dapergolas A., Bailer-Jones C. A. L., Vallenari A., Fioc M. 2006, GAIA-C8-TN-UOA-PAT-001-1

Tsalmantza P., Rocca-Volmerange B., Kontizas M., Bellas-Velidis I., Kontizas E., Bailer-Jones C. A. L., Korakitis R., Dapergolas A., Livanou E., Fioc M., Vallenari A., 2007,

GAIA-C8-TN-UOA-PAT-002-1

Tsalmantza P., Kontizas M., Bailer-Jones C. A. L., Rocca-Volmerange B., Korakitis R., Kontizas E., Livanou E., Dapergolas A., Bellas-Velidis I., Vallenari A., Fioc M. 2007, A&A 470, 761

Tsalmantza P., Kontizas M., Rocca-Volmerange B., Bailer-Jones C. A. L., Kontizas E., Bellas-Velidis I., Korakitis R., Livanou E., Dapergolas A., Vallenari A., Fioc M. 2008, in preparation

Yip C. W., Connolly A. J., Szalay A., Budavari T., SubbaRao M., Frieman J., Nichol R., Hopkins A., York D., Okamura S., Brinkmann J., Csabai I., Thakar A. R., Fukugita M., Ivezic Z., AJ, 2004, 128, 585

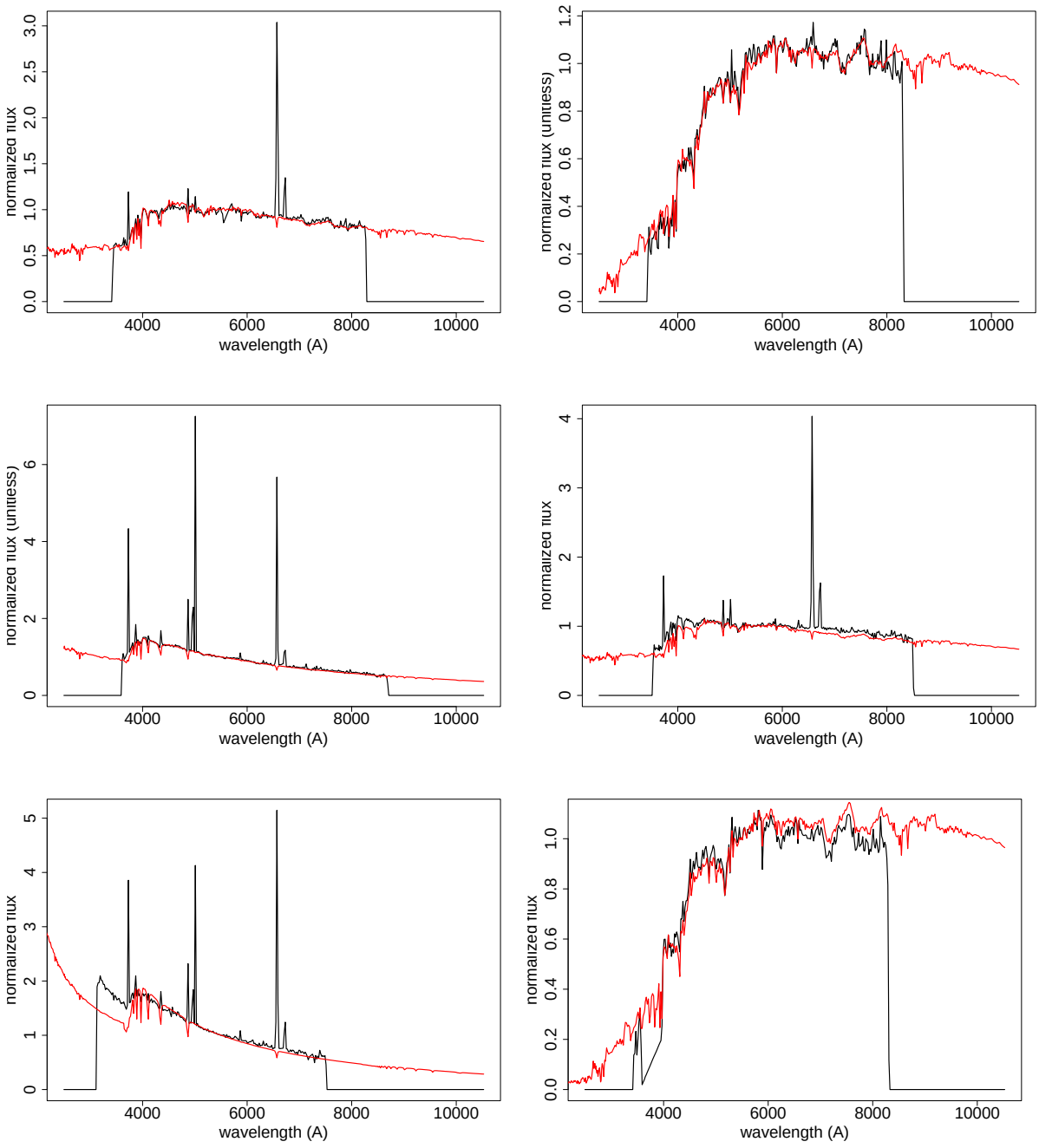

Figure 9: UPPER LEFT: SDSS spectrum 53062-1445-172 (black) and the synthetic spectrum (red) with the smallest χ^2-difference equal to 0.33. UPPER RIGHT: The same for the spectrum 51691-0342-384 with difference 2.38. MIDDLE LEFT: The same for the spectrum 51984-0279-123 with difference 7.50. MIDDLE RIGHT: The same for the spectrum 52377-0624-415 with difference 15.00. BOTTOM LEFT: The same for the spectrum 53171-1680-527 with difference 30.47. BOTTOM RIGHT: The same for the spectrum 53472-2005-279 with difference 13907.66.

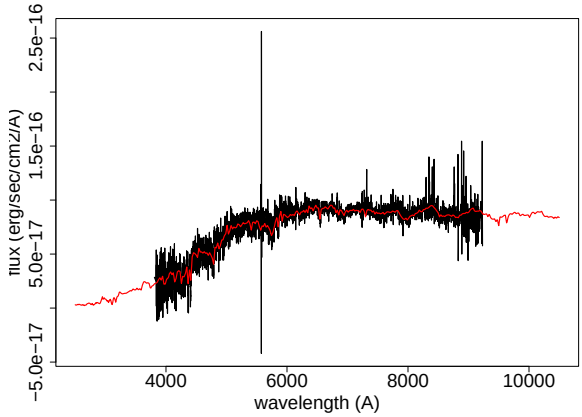

Figure 10: The initial SDSS spectrum 51691-0342-384 (black) and the synthetic spectrum (red) with the smallest difference after the application of redshift, change of the normalization and resolution.

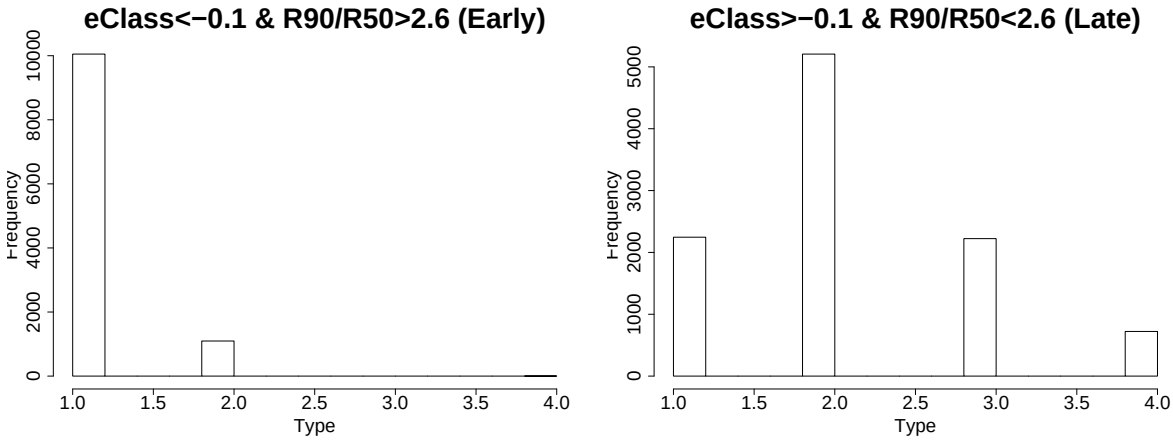

Figure 11: Comparison of the classification of SDSS galaxies based on the Yip et al. (2004) and Strateva et al. (2001) criteria (eClass and index C respectively) and the classification based on the synthetic spectra. With 1, 2, 3 and 4 we represent the early. spiral, irregular and starburst galaxies of the synthetic library respectively.

Table 1: Parameters given for each observational galaxy spectrum in the empirical library.

Parameter	Units
Index number	
morphological type (early type:1, spiral:2, irregular:3, starburst:4)	
type of star formation rate (proportional Mgas:1, prop. Mgas with stop:2, exp:3)	
p1 of the star formation rate	Myr or –
p2 of the star formation rate	Myr/$M_\odot$
time since the star formation rate stopped	Myr
infall timescale	Myr
age	Myr
normalized mass of the galaxy	$M_\odot$
normalized mass in stars	$M_\odot$
normalized mass in white dwarfs	$M_\odot$
normalized mass in neutron stars and black holes	$M_\odot$
normalized mass in substellar objects	$M_\odot$
normalized mass in the gas	$M_\odot$
metallicity of the interstellar medium (mass fraction)	
mean metallicity of stars averaged on the mass	
the mean metallicity of stars averaged on the bolometric luminosity	
normalized bolometric luminosity	erg/s
the optical depth in the V-band (5500 A) from side to side	
ratio of the luminosity emitted by the dust to the bolometric luminosity	
normalized star formation rate	$M_\odot$/Myr
normalized number of Lyman continuum photons emitted	1/s
normalized SNII rate	1/Myr
normalized SNIa rate	1/Myr
mean age of the stars averaged on the mass	Myr
mean age of stars averaged on the bolometric luminosity	Myr
objID	
specobjID	
ra	deg
dec	deg
fiberMag-r	mag
fiberMagErr-r	mag
modelMag-r	mag
modelMagErr-r	mag
petroR90-r	arcsec
petroR90Err-r	arcsec
petroR50-r	arcsec
petroR50Err-r	arcsec
z	
zErr	
sn-1	
mjd	
plate	
fiberID	
eClass	
difference in χ^2-fitting	