RAA 2016 Vol. **16** No. **3**, 42 (12pp) doi: 10.1088/1674–4527/16/3/042
http://www.raa-journal.org *http://iopscience.iop.org/raa*

***R**esearch in*
***A**stronomy and*
***A**strophysics*

Spectral Synthesis via Mean Field approach to Independent Component Analysis

Ning Hu, Shan-Shan Su and Xu Kong

CAS Key Laboratory for Research in Galaxies and Cosmology, Department of Astronomy, University of Science and Technology of China, Hefei 230026, China; *huning@mail.ustc.edu.cn, xkong@ustc.edu.cn*

Received 2015 April 29; accepted 2015 September 7

Abstract We apply a new statistical analysis technique, the Mean Field approach to Independent Component Analysis (MF-ICA) in a Bayseian framework, to galaxy spectral analysis. This algorithm can compress a stellar spectral library into a few Independent Components (ICs), and the galaxy spectrum can be reconstructed by these ICs. Compared to other algorithms which decompose a galaxy spectrum into a combination of several simple stellar populations, the MF-ICA approach offers a large improvement in efficiency. To check the reliability of this spectral analysis method, three different methods are used: (1) parameter recovery for simulated galaxies, (2) comparison with parameters estimated by other methods, and (3) consistency test of parameters derived with galaxies from the Sloan Digital Sky Survey. We find that our MF-ICA method can not only fit the observed galaxy spectra efficiently, but can also accurately recover the physical parameters of galaxies. We also apply our spectral analysis method to the DEEP2 spectroscopic data, and find it can provide excellent fitting results for low signal-to-noise spectra.

Key words: methods: data analysis — methods: statistical — galaxies: evolution — galaxies: fundamental parameters — galaxies: stellar content

1 INTRODUCTION

Spectra contain plentiful information about the properties of a galaxy (Kong et al. 2014). Finding a way to analyze the spectra of observed galaxies and determine the parameters of a large sample of galaxies would not only help us to investigate galaxy formation and evolution, but also allow us to derive cosmological information from a large number of galaxies (Conroy 2013). Many methods, based on the relevant features, have been devised to measure and understand the physical parameters of galaxies, either by using spectral indices (Worthey et al. 1994) or emission features (Kewley et al. 2001; Shi et al. 2014), or by fitting the full spectrum (Tremonti et al. 2004; Cid Fernandes et al. 2005; Ocvirk et al. 2006; Tojeiro et al. 2007; Liu et al. 2013). Due to the abundance of high-quality galaxy spectra, two different population synthesis approaches have been commonly used to study the stellar contents of a galaxy. The empirical population synthesis method (Faber 1972; Bica 1988; Cid Fernandes et al. 2001; Kong et al. 2003) is based on modeling galaxies as a mixture of several observed spectra of stars or star clusters. However, this method does not consider stellar evolution and is limited by the observed stellar/cluster spectral library. Recently, a more direct approach, called evolutionary population synthesis (Vazdekis, 1999; Girardi et al. 2000; Bruzual & Charlot 2003 hereafter BC03; Maraston 2005; Chen et al. 2015) has been widely used. In this approach, the spectra of stellar populations are modeled by combining stellar evolution tracks, stellar spectral libraries and star formation histories (SFHs). Up to now, a popular simple stellar population (SSP) library was provided by the isochrone synthesis technique (BC03). Several groups have selected a few SSPs from this library as templates to fit observed galaxy spectra (Tremonti et al. 2004; Cid Fernandes et al. 2005).

However, the advent of large-area spectroscopic surveys, such as the Third Sloan Digital Sky Survey (SDSS-III; Eisenstein et al. 2011), the Deep Extragalactic Evolutionary Probe 3 (DEEP3) Galaxy Redshift Survey (Cooper et al. 2011), and the Large sky Area Multi-Object fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012), will be providing oceans of data, thus the development of fast and automated extraction methods is required. We note that statistical analysis techniques have been commonly implemented. For example, Richards et al. (2009) utilized the diffusion k-means method to draw several prototype spectra from the SSP database as input templates of the spectral synthesis software STARLIGHT (Cid Fernandes et al. 2005). Nolan et al. (2006) applied a data-driven Bayesian approach to the spectra of early-type galaxies. Another blind source separation (BSS) technique applied to spectra is principal component analysis (PCA, Mittaz et al. 1990; Kong & Cheng 2001; Yip et al. 2004), but the

interpretation of the individual component spectra seems rarely illuminating. Here, we explore a new statistical multivariate data processing technique, independent component analysis (ICA), in our spectral analysis. This technique has been implemented in studies of the Cosmic Microwave Background (Maino et al. 2007) and the analysis of spectra (Lu et al. 2006; Allen et al. 2013); however, the Ensemble Learning ICA (EL-ICA, also known as naive mean field or NMF) method used in Lu et al. (2006) is known to fail in some circumstances (e.g. low signal-to-noise (S/N) spectra) (Højen-Sørensen et al. 2001), and Allen et al. (2013) only applied this technique to emission-line galaxies. For the sake of non-negative values in the spectral analysis of galaxies, we adopt a new ICA algorithm, the mean field approach to independent component analysis (MF-ICA), which can constrain the sources and the mixing matrix to be non-negative with a more efficient and more reliable algorithm.

The paper is structured as follows. In Section 2, we introduce the MF-ICA method, and derive a few templates from evolutionary population models of Charlot & Bruzual (2007, CB07 hereafter) which can be later used to analyze the spectra of galaxies. In Section 3, the simulated galaxy spectra are used to analyze the reliability of the MF-ICA method. In Section 4, we analyze galaxy spectra observed by the SDSS, and compare our results with those obtained from the MPA/JHU[1] catalogs, to investigate whether our synthesis results are reasonable. In Section 5, some galaxy spectra from the DEEP2 galaxy redshift survey are fitted by our method, and our conclusions are outlined in Section 6.

2 METHOD

2.1 Stellar Population Models

Stellar population models can be generated by several population synthesis codes. Here we adopt the 2007 version of Galaxev[2] (CB07), which is a new version of BC03. The CB07 models have undergone a major improvement recently with the new stellar evolution prescriptions of Marigo & Girardi (2007) for the Thermally-Pulsing Asymptotic Giant Branch (TP-AGB) evolutionary phase of low- and intermediate-mass stars. An accurate modeling of this phase is related to correctly predicted fluxes in the wavelength range of $1-2.5\ \mu\mathrm{m}$ (CB07).

The CB07 models use an empirical spectral library with a range of wavelength ($91\,\text{Å} - 36\,000\ \mu\mathrm{m}$, $N = 6917$), and spectral resolution of about 3 Å. Moreover, CB07 contains a large sample of SSPs, which covers 221 different ages from 1.0×10^5 to 2.0×10^{10} yr, and a wide range of initial chemical compositions, $Z = 0.0001$, 0.0004, 0.004, 0.008, 0.02, 0.05 and 0.1 ($Z_\odot = 0.02$). The observed spectrum of a galaxy can be expressed as a combination of these individual SSPs with weights. This SSP database will be used to derive our templates in Section 2.2.3.

2.2 MF-ICA Technique

2.2.1 Independent Component Analysis (ICA)

ICA is a new multivariate data processing method which aims at decomposing complex multivariate observations into a combination of a few hidden original sources (Hyvärinen et al. 2001). Compared to the traditional multivariate data processing methods, such as principal component analysis (PCA) or factor analysis, ICA is much more powerful at finding hidden sources, even when traditional methods fail completely. The following generative model of ICA shows that multivariate observations or mixed signals x^i, $i = 1, 2, ..., m$, are a combination of hidden sources, i.e. Independent Components (ICs), h_k, $k = 1, 2, ..., n$, with additive Gaussian noise Γ^i, weighted by the mixing weights w_k^i ($m \times n$)

$$x^i = \sum_{k=1}^{n} w_k^i h_k + \Gamma^i \quad (i = 1, 2, ..., m)\,. \tag{1}$$

In our analysis, we take multivariate observations as the spectra of stellar systems (e.g. SSP database), and adopt the assumption that each spectrum $f^i(\lambda)$ can be expressed as a sum of several ICs, $\mathrm{IC}_k(\lambda)$, so the model can be written as

$$f^i(\lambda) = \sum_{k=1}^{n} w_k^i \mathrm{IC}_k(\lambda) + \Gamma^i(\lambda)\,. \tag{2}$$

Here, we only know the spectrum $f^i(\lambda)$. The unknown mixing weights w_k^i, the ICs $\mathrm{IC}_k(\lambda)$ and the noise can be estimated from ICA algorithms, such as Joint Approximate Diagonalization of Eigenmatrices (JADE; Cardoso & Souloumiac 1993), extended InfoMax (Bell & Sejnowski 1995), FastICA (Hyvärinen et al. 2001), Ensemble Learning ICA (EL-ICA; Miskin & MacKay 2001), Mean Field ICA (MF-ICA; Højen-Sørensen et al. 2002) and many others.

2.2.2 Mean field approach ICA (MF-ICA)

The ICA algorithm we adopt in our spectral analysis is the MF-ICA method. Compared to other algorithms, MF-ICA is a Bayesian iterative algorithm which can constrain sources and the mixing matrix to be positive by offering priors for them. The main advantage of the MF-ICA algorithm is the simplicity and generality of its implementation.

In this approach, the likelihood for the parameters and sources is defined as $P(\mathbf{X}|\mathbf{W}, \mathbf{\Sigma}, \mathbf{H})$ given by

$$P(\mathbf{X}|\mathbf{W}, \mathbf{\Sigma}, \mathbf{H}) = (\det 2\pi\mathbf{\Sigma})^{-\frac{n}{2}} \times \mathrm{e}^{-\frac{1}{2}\mathbf{Tr}(\mathbf{X}-\mathbf{WH})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{WH})}, \tag{3}$$

where $\mathbf{W}$ is the mixing matrix, $\mathbf{X} = [x^1, x^2, ..., x^m]^{\mathrm{T}}$ is the mixed signals matrix, $\mathbf{\Sigma}$ is the noise covariance matrix, n

[1] *http://www.mpa-garching.mpg.de/SDSS/*

[2] *http://www.bruzual.org*

is the number of input source signals, and $\det$ is the determinant of the matrix. The likelihood of the parameters is defined as $P(\mathbf{X}|\mathbf{W},\mathbf{\Sigma})$ obtained from

$$P(\mathbf{X}|\mathbf{W},\mathbf{\Sigma}) = \int d\mathbf{H} P(\mathbf{X}|\mathbf{W},\mathbf{\Sigma},\mathbf{H}) P(\mathbf{H}). \quad (4)$$

If priors on the mixing weight $P(\mathbf{W})$ and the sources $P(\mathbf{H})$ are taken into account, then the posteriors of sources and the mixing matrix are obtained from $P(\mathbf{H}|\mathbf{X},\mathbf{W},\mathbf{\Sigma}) \propto P(\mathbf{X}|\mathbf{W},\mathbf{\Sigma},\mathbf{H})P(\mathbf{H})$ and $P(\mathbf{W}|\mathbf{X},\mathbf{\Sigma}) \propto P(\mathbf{X}|\mathbf{W},\mathbf{\Sigma})P(\mathbf{W})$, respectively. In the MF-ICA method, the noise covariance $\mathbf{\Sigma}$ and mixing matrix $\mathbf{W}$ can be obtained from maximum a posteriori estimation, while sources $\mathbf{H}$ can be obtained from their posterior mean. The mean field approach can be solved by:

$$\hat{\mathbf{H}} = \langle \mathbf{H} \rangle , \quad (5)$$

$$\hat{\mathbf{W}} = \mathbf{X}\langle \mathbf{H}^{\mathrm{T}} \rangle \langle \mathbf{H}\mathbf{H}^{\mathrm{T}} \rangle^{-1} , \quad (6)$$

$$\mathbf{\Sigma} = \frac{1}{n} \langle (\mathbf{X} - \hat{\mathbf{W}}\hat{\mathbf{H}})(\mathbf{X} - \hat{\mathbf{W}}\hat{\mathbf{H}})^{\mathrm{T}} \rangle , \quad (7)$$

where $\langle \cdot \rangle = \langle \cdot \rangle_{\mathbf{H}|\mathbf{W},\mathbf{\Sigma},\mathbf{X}}$ denotes the posterior average with respect to the sources given the mixing matrix and noise covariance. The solution of the MF-ICA algorithm equals the updated noise covariance (Eq. (7)) and mixing matrix (Eq. (6)), and estimating sources (Eq. (5)). Thus the optimized matrices of mixing matrix $\hat{\mathbf{W}}$, noise covariance $\mathbf{\Sigma}$, and sources $\hat{\mathbf{H}}$ can be derived from this iterative method. More details about the MF-ICA method can be found in Højen-Sørensen et al. (2002) and the available MATLAB toolbox (*http://mole.imm.dtu.dk/toolbox/ica*). Through Bayesian inference about the mixing matrix and sources, their priors can be constrained to be non-negative, which will be useful in processing observed galaxy spectra, since the spectral parameters should not be negative. Although the EL-ICA method has been implemented in galaxy spectral analysis (Lu et al. 2006), here we adopt the MF-ICA method, which relies on advanced mean field approaches: linear response theory and an adaptive version of the mean-field approach. Højen-Sørensen et al. (2001, 2002) have investigated both the MF-ICA and EL-ICA methods. They concluded that compared to the EL-ICA method, the advanced mean field approaches can recover the correct sources even when ensemble learning theory fails, and the convergence rate of the MF-ICA method is found to be faster. A comparison of these two ICA methods will be described in Section 3.3.

2.2.3 Analysis SSPs using MF-ICA

Through the multivariate data processing technique, we expect to derive a minimal number of non-negative templates, which can represent the spectra of a galaxy with minimal loss of information. Here, we adopt the MF-ICA algorithm to compress the spectral library of SSP from CB07 models (Sect. 2.1).

The SSP database of CB07 contains 1547 spectra (Sect. 2.1). Each spectrum was first truncated to the high resolution wavelength range of $3322 - 9200$ Å, to match that of the SDSS spectrograph. In the EL-ICA method, the number of sources (i.e. ICs) should be the same as the number of mixed signals. Therefore, Lu et al. (2006) picked up a subsample out of the BC03 SSP database as the mixed signals matrix $\mathbf{X}$ in the EL-ICA method, and estimated 74 hidden spectra. Finally they choose several ICs from these hidden spectra by the average fractional contribution to the BC03 SSP database. However, the MF-ICA method that we applied can perform dimensionality reduction. Here the whole CB07 SSP database was set as the input mixed signals matrix $\mathbf{X}$, then the MF-ICA method will be applied to them, and the output ICs will be more precise. To avoid negative values appearing in spectral analysis, we set the priors of the mixing matrix and sources to be positive.

As has been mentioned above, the number of ICs can be less than the number of mixed signals in the MF-ICA method, thus it should be predefined. The correct number can be determined as follows. We apply the Root Sum Square (RSS) method to select the proper number of ICs. The value of RSS between the original mixed signals (i.e. whole SSP database) and the recovered mixed matrix can be calculated by

$$\mathrm{RSS} = \left(\sum_{j=1}^{n} \sum_{i=1}^{m} (x_j^i - \hat{x_j^i})^2 \right)^{1/2} , \quad (8)$$

where the recovered mixed matrix $\hat{\mathbf{X}}$ is calculated from the estimated mixing matrix and sources: $\hat{\mathbf{X}} = \hat{\mathbf{W}}\hat{\mathbf{H}}$. We preset the initial number of sources as one, then increase the number and the value of RSS will be reduced. We repeat this process until the reduction is no longer significant. Finally, the number of ICs can be set as 12.

Using the number of ICs we determined, the sources can be obtained from the MF-ICA calculation. Therefore, the SSP database can be compressed into 12 ICs. We present these 12 ICs in figure 4 of Su et al. (2013).

To confirm the reliability and quality of the ICs, we used the 12 estimated ICs to recover the 1547 SSPs in the CB07 database as follows

$$f_{\mathrm{SSP}}^i(\lambda) = \sum_{k=1}^{12} w_k^i \mathrm{IC}_k(\lambda) \quad (i = 1, 2, ..., 1547), \quad (9)$$

and we found that the spectra reconstructed by these 12 ICs excellently match those in the SSP database.

2.3 Fitting Galaxy Spectra

The aim of this study is to use these estimated 12 ICs to fit galaxy spectra from large surveys. The SFHs of a galaxy can be approximated as a combination of discrete bursts, thus the population of a galaxy can be decomposed into a combination of SSPs. As shown in Section 2.2.3, the SSP database can be recovered with 12 ICs, so the model of

observed galaxy spectra, $f_g(\lambda)$, can be fitted by these 12 ICs as

$$f_g(\lambda) = r(\lambda) \sum_{k=1}^{12} a_k \mathrm{IC}_k(\lambda, \sigma), \qquad (10)$$

where $r(\lambda)$ is the reddening term, which describes the intrinsic starlight reddening and can be modeled by the extinction law of Charlot & Fall (2000). $\mathrm{IC}_k(\lambda, \sigma)$ is the $k-$th IC convolved with a Gaussian function. The Gaussian width σ corresponds to the stellar velocity dispersion of a galaxy. During the fitting process, we mask points around prominent lines, such as Balmer lines (Hϵ, Hγ, Hδ, Hβ, Hα) and strong forbidden lines ([O II]λ3727, [Ne III]λ3869, [O III]$\lambda\lambda$4959, 5007, [He I]λ5876, [O I]λ6300, [N II]$\lambda\lambda$6548, 6584 and [S II]$\lambda\lambda$6717, 6721).

After subtracting the modeled stellar population spectrum, emission lines can be fitted with Gaussians simultaneously, similar to Tremonti et al. (2004): the forbidden lines ([O II], [O III], [O I], [N II] and [S II]) are set to have the same line width and velocity offset; the same treatment is applied to Balmer lines (Hγ, Hδ, Hβ and Hα). By using the procedures above, the observed galaxy spectra can be quickly recovered.

3 RELIABILITY OF THE FITTING METHOD

3.1 Simulations

In this section, we analyze the simulated galaxy spectra to examine the reliability of the MF-ICA method. All simulated spectra are generated from the 2007 version of BC03 stellar population synthesis code. For the sake of simplicity, we parameterize each SFH of the simulated galaxy in terms of an underlying continuous model superimposed with random bursts on it (Kauffmann et al. 2003). The spectral energy distribution (SED) at time t of a stellar population characterized by an exponentially declining star formation law $\psi(t) \propto e^{-\gamma t}$ is given by

$$F_\lambda(t) = \int_0^t \psi(t - t') S_\lambda(t', Z) dt', \qquad (11)$$

where $S_\lambda(t', Z)$ is the power radiated by an SSP of age t' and metallicity Z per unit wavelength per unit initial mass.

The added SFHs are described below:

(1) The time when a galaxy begins forming stars t is distributed uniformly between 0.1 and 13.5 Gyr. Star formation timescale γ is uniformly distributed between 0 and 1 Gyr^{-1}.
(2) Random bursts occur at any time with the same probability. Bursts are parameterized in terms of the fraction of stellar mass produced, which is logarithmically distributed between 0.03 and 4, and their duration can vary between 0.03 and 0.3 Gyr.
(3) The metallicities Z are uniformly distributed between $0.02\,Z_\odot$ and $2\,Z_\odot$, which represent the range of stellar metallicities inferred from the spectra of $\sim 2 \times 10^5$ SDSS galaxies.

We apply our spectral analysis method to 500 simulated spectra over the range $3322 - 9200$ Å. We also use the extinction law of Charlot & Fall (2000) to attenuate each spectrum, where the absorption optical depth τ_V is uniformly distributed between 0 and 5. The velocity dispersion σ is uniformly distributed between 50 $\mathrm{km\,s^{-1}}$ and 450 $\mathrm{km\,s^{-1}}$. Finally we added Gaussian noise with (S/N) $= 10$, 20 and 30.

3.2 Results

From fitting simulated spectra, we expect to examine the reliability of our spectral analysis approach which is based on the MF-ICA algorithm. Our main parameters of interest are $\mathrm{A_V}$, σ, t and Z. The following steps are used to estimate age and metallicity:

(1) The pure spectrum of a stellar system in a galaxy, $f_g(\lambda)$, can be recovered by ICs, and it can also be represented by a combination of N SSPs. Thus we can solve the equation

$$f_g(\lambda) = \sum_{k=1}^{12} a_k \mathrm{IC_k}(\lambda) = \sum_{j=1}^{N} b_j f^j_{\mathrm{SSP}}(\lambda). \qquad (12)$$

(2) We adopt 60 SSPs from the CB07 database including models of 15 different ages ($t = 0.001$, 0.003, 0.005, 0.01, 0.025, 0.04, 0.1, 0.2, 0.6, 0.9, 1.4, 2.5, 5, 11, 13 Gyr) and four different metallicities ($Z = 0.004$, 0.008, 0.02, 0.05).
(3) After solving Equation (12), the age and metallicity can be computed by

$$\langle \log t \rangle_L = \sum_{j=1}^{60} b_j \log(t_j)\,, \qquad (13)$$

$$\log\langle Z \rangle_L = \log \sum_{j=1}^{60} b_j Z_j\,. \qquad (14)$$

Figure 1 shows the input parameters versus estimated values from simulated spectra with S/N $= 10$, 20 and 30. Clearly, the values of starlight reddening A_V and stellar velocity dispersion σ are relatively well recovered. The mean square errors (MSEs) between recovered and input values are less than 0.20 and 7.45, respectively, as shown in Table 1.

Table 1 Summary of parameter error estimates for simulated spectra. The different rows list the MSE between output and input values of the corresponding quantity, as obtained from simulations with different values of S/N.

S/N	$\mathrm{MSE_{A_V}}$ (mag)	MSE_σ ($\mathrm{km\,s^{-1}}$)	$\mathrm{MSE}_{\langle \log\ t \rangle_L}$	$\mathrm{MSE}_{\log\langle Z \rangle_L}$
10	0.191	7.449	0.201	0.201
20	0.169	6.301	0.189	0.202
30	0.119	6.017	0.169	0.196

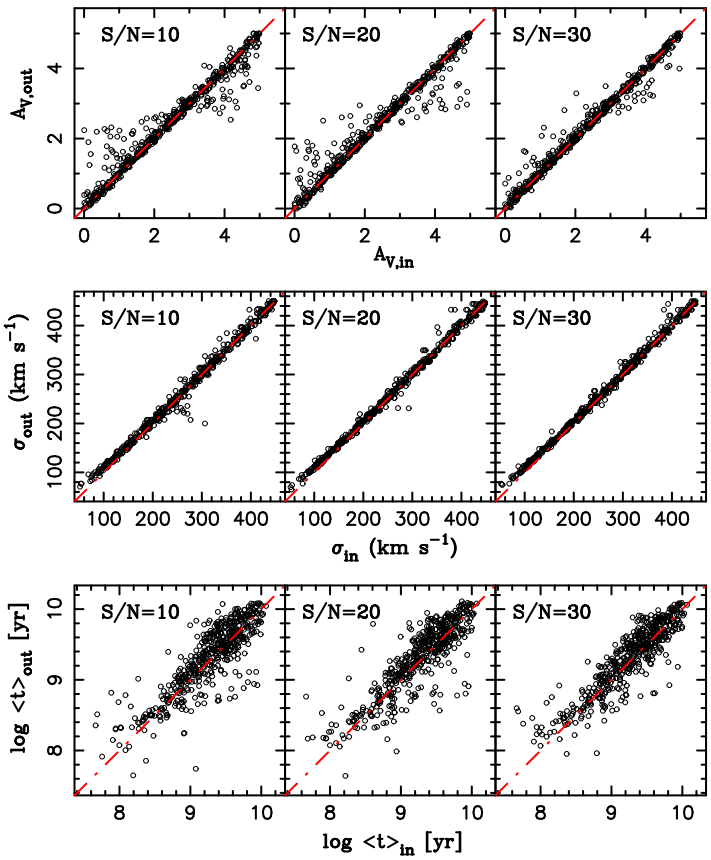

Fig. 1 Comparison of the input $A_{\rm V}$ (magnitude), σ (km s^{-1}) and stellar ages (yr) with output values that are estimated from simulations, with S/N=10, 20 and 30, using our MF-ICA method. The red dot-dashed line is the identity line ($y = x$).

From the above method, a galaxy spectrum can be decomposed into 60 SSPs with weights. The estimated weights b_j can reflect the fractional contributions of the $j-$th SSP with age t_j and metallicity Z_j. Therefore, the light-weighted age and metallicity can be estimated. As shown in Figure 1 (bottom panel), the recovered and input values of $\langle \log t \rangle_L$ have no significant difference with MSE less than 0.20. According to the age-metallicity degeneracy problem (Bressan et al. 1996), the values of $\log\langle Z \rangle_L$ recovered by our method must have some differences. However, we can estimate the parameters with reasonable accuracy, and the Spearman's rank correlation coefficient $r_{\rm s}$ between output and input $\log\langle Z \rangle_L$ is about 0.70 for S/N $= 20$. Finally, the summary of MSE for parameters from simulated spectra can be found in Table 1.

3.3 Comparison with the EL-ICA Method

To carefully test the performance of ICA algorithms, we re-estimated the ICs by the EL-ICA algorithm (Miskin & MacKay 2001). The EL-ICA method, which is also known as the naive mean field ICA method, has been applied in galaxy spectral analysis by Lu et al. (2006). Here we used the same steps as Lu et al. (2006) and also derived six ICs, which we present in Figure 2.

We use these ICs to refit the simulated spectra. The input parameters versus the estimated values that are output and the MSE between them for simulations with S/N=20 are shown in Figure 3. The dispersions of parameters derived by the EL-ICA method are larger than those by our method, which are plotted in Figure 1. The MSEs of starlight reddening, velocity dispersion, stellar age and metallicity ($A_{\rm V}$, σ, $\langle \log\ t \rangle_L$, $\log\langle Z \rangle_L$) are 0.421, 33.841, 0.405 and 0.299 for S/N=20, respectively. These values are much larger than those of our method, as shown in Table 1. Finally, we also fit the SSP database using these ICs. The spectra recovered are not as good as those by our method. We conclude that our method which is based on the MF-ICA algorithm is more precise and reliable.

4 APPLICATION TO SDSS SPECTRA

Using our MF-ICA spectral analysis method, we fit the SDSS galaxy spectra, analyze their stellar population properties, and measure their emission-line properties from the starlight-subtracted spectra. In this section, we compare the physical parameters obtained from stellar population analysis of the continuum and measurements of emission lines. We also compare parameters estimated from our fitting technique with those derived by the MPA/JHU group. Because the aim of this section is only to test whether

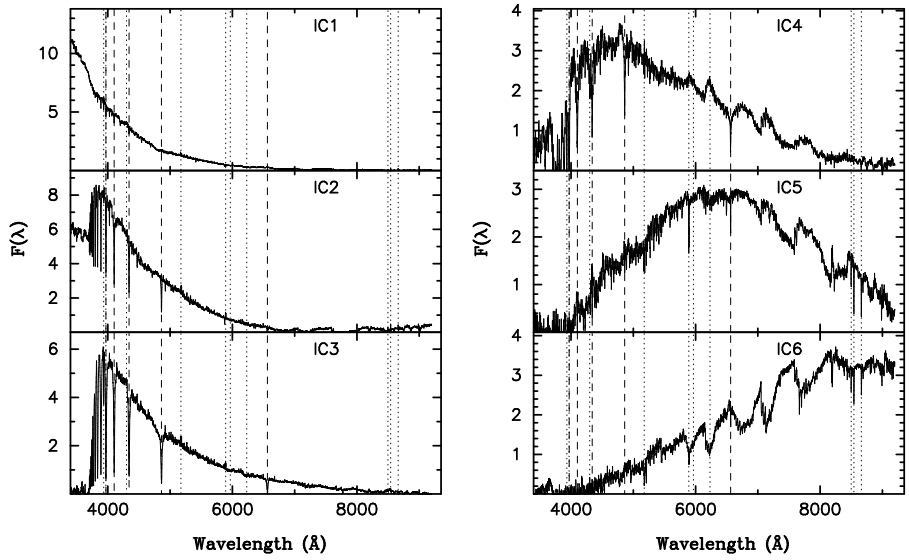

Fig. 2 The spectra of six ICs estimated by the EL-ICA method. Some prominent spectral features are labeled the same as those in figure 4 in Su et al. (2013).

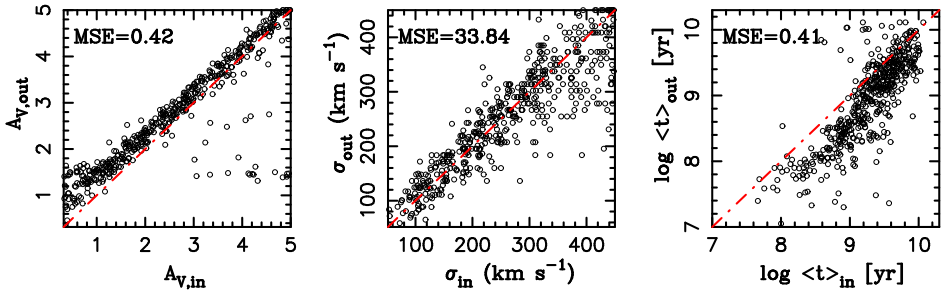

Fig. 3 Comparison of the input $A_{\rm V}$ (magnitude), σ (km s^{-1}) and stellar ages (yr) with the estimated output values from simulations, with S/N=20, using the EL-ICA method. The red dot-dashed line is the identity line ($y = x$).

the results from our spectral analysis method are reasonable and meaningful, we will not investigate their physical properties, such as the formation and evolution of galaxies.

4.1 Data Preparation

The Sloan Digital Sky Survey (SDSS; York et al. 2000) has released huge amounts of high-quality observed spectra of objects. In this work, our sample of spectra was extracted from spectroscopic plates of SDSS Data Release 8 (DR8; Aihara et al. 2011). Moreover, we choose the objects which have been spectroscopically classified as galaxies. The spectra obtained from SDSS span a wavelength range from 3800 Å to 9200 Å with mean spectral resolution $R = \lambda/\Delta\lambda \sim 1800$, and are taken with three arcsecond diameter fibers. We finally fit about one million spectral samples of galaxies with redshift less than 1, which are obtained from the SDSS spectroscopic pipeline.

We use the MF-ICA method, which was described in Section 2.3, to fit the spectral sample of galaxies from SDSS. First, the spectra of galaxies were corrected for foreground Galactic extinction, using the maps of Schlegel et al. (1998). Then, they were transformed into the rest frame, with spectroscopic redshifts. The spectral fitting results give a median χ^2/d.o.f (degree of freedom) of 1.13, nearly the excellent value of 1 that we expect.

Figure 4 shows some examples of the fitting. The spectra can be well recovered by visual inspection, which suggests that our MF-ICA spectral analysis approach works well.

4.2 Comparisons with the MPA/JHU Database

The MPA/JHU group has provided catalogs of estimated physical parameters of SDSS galaxies publicly available on the website.[3] They inferred the SFHs of DR8 galaxies on the basis of CB07 models, which were similar to results from our research. Here we compare our own estimated parameters, such as the emission line measurements and stellar population properties, with those obtained from the MPA/JHU catalogs. Although we do not expect our estimated parameters to be perfectly consistent with theirs,

[3] See *http://www.sdss3.org/dr8/spectro/spectro_access.php*

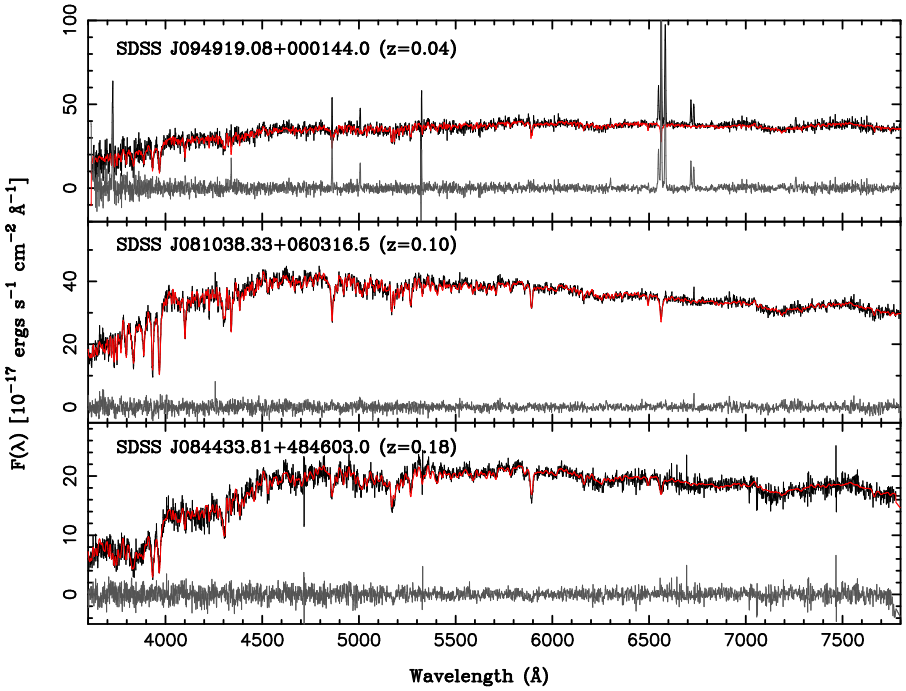

Fig. 4 The spectral fitting results of some galaxies in our SDSS DR8 sample at a range of redshifts. The black lines show the observed spectrum, red lines show the modeled stellar spectrum, grey lines show the residual spectrum, and the redshift is labeled in the top left corner of each panel.

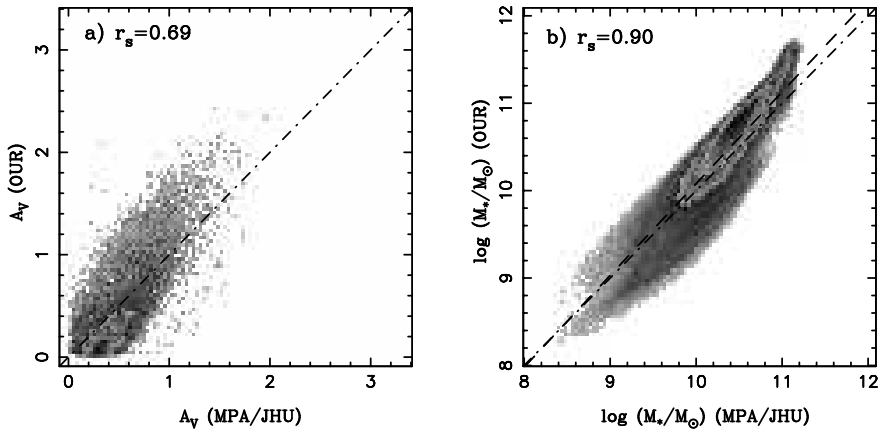

Fig. 5 The values of stellar extinction $A_{\rm V}$ and stellar mass M_* computed from the MPA/JHU database versus those values computed by our code. The dot-dashed line is the identity line ($y = x$). The dashed line in the right panel is a robust fit for the relation. The number in the top left corner is the Spearman rank correlation coefficient.

we analyze the relationships between these parameters to examine the accuracy and reliability of the MF-ICA algorithm.

4.2.1 Stellar extinction

In our fitting technique, the extinction of optical galaxy spectra is modeled as one free parameter $A_{\rm V}$. In Figure 5(a), we plot the values of $A_{\rm V}$ estimated by our method versus those estimated by the MPA/JHU group, which adopt the same attenuation curve by Charlot & Fall (2000). Since only a few galaxies with $A_{\rm V} < 0$ are found in previous research works, we constrain $A_{\rm V}$ to be positive in our analysis. Therefore, this constraint will not have a significant impact on the results of our analysis.

We adopt the value of Spearman's rank correlation coefficient $r_{\rm s}$ to describe the relationship between two variables. As shown in Figure 5(a), our results are well and linearly correlated to those extracted from the MPA/JHU catalogs, with $r_{\rm s} = 0.69$. However, the extinction values $A_{\rm V}$ obtained from the MPA/JHU database are systematically lower than our values, similar to findings in Chen et al. (2012, in fig. 3f). One possible reason for this dis-

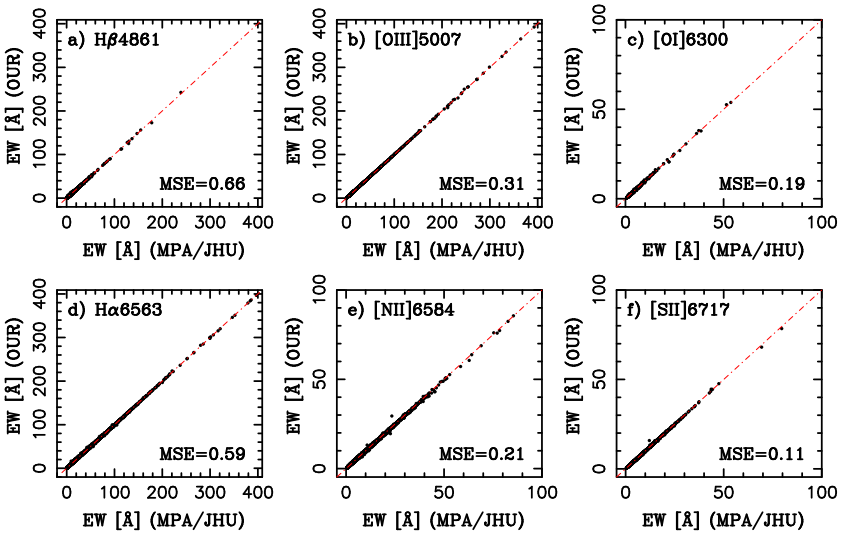

Fig. 6 The comparison of EWs of Hβ, [O III]λ5007, [O I]λ6300, Hα, [N II]λ6584 and [S II]λ6717 measured by the MPA/JHU group with those by our code. The red dot-dashed line is the identity line ($y = x$), while the number in the bottom-right corner of each panel indicates the MSE.

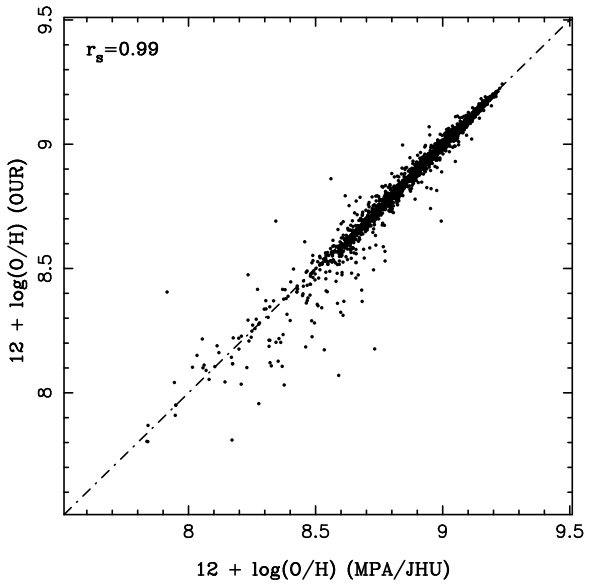

Fig. 7 Plot of our estimated nebular oxygen abundances versus those obtained by the MPA/JHU group. The dot-dashed line is the identity line ($y = x$), while the number in the top left corner is the Spearman rank correlation coefficient.

crepancy is that we only use the optical-band spectra to estimate the stellar extinction $A_{\rm V}$.

4.2.2 Stellar mass

By using our stellar population analysis method, the light-weighted stellar mass $\log\langle M\rangle_L$ of an SDSS galaxy also can be estimated. We calculate the mass to light (M/L) ratio by adding the weighed M/L ratios of each SSP component, and then derive the stellar mass by multiplying it by luminosity.

In Figure 5(b), we plot our estimated stellar mass versus the MPA/JHU extinction-corrected stellar mass. The results from the two methods are very consistent, with $r_{\rm s} = 0.90$. The small discrepancy is caused due to the different estimation methods. In our method, the stellar masses are obtained from the M/L ratio, which is estimated through the best χ^2 model. However, the MPA/JHU group estimated their M/L ratio through a Bayesian inference method, which connects two indices, H$\delta_{\rm A}$ and $D_{\rm n}(4000)$, with a model obtained from a large library of Monte Carlo realizations of galaxies with different SFHs (Kauffmann et al. 2003).

4.2.3 Emission lines and nebular metallicities

In our case the emission lines were measured from a starlight-subtracted spectrum. The MPA/JHU group

adopted a similar method, however, they only used a single metallicity CB07 model to fit the observed continuum. We plot our estimated equivalent widths (EWs) of emission lines, such as Hβ, [O III]λ5007, [O I]λ6300, Hα, [N II]λ6584 and [S II]λ6717 versus those measured by the MPA/JHU group in Figure 6. As shown in Figure 6, our values are consistent with those measured by the MPA/JHU group, with a small discrepancy. We adopt the MSE to quantify the discrepancy between them. On the whole, the MSEs of all these values are less than 1, suggesting that there are no significant differences. The small discrepancy appearing is due to the different measurement of the synthesized spectrum, which is related to different subtracted stellar absorption.

We also compared the value of nebular oxygen abundance $12 + \log(\mathrm{O/H})$, which can be obtained from the equation described in Tremoni et al. (2004). As shown in Figure 7, our estimated values of nebular oxygen abundance show a high degree of correspondence with those drawn from the MPA/JHU catalog. The value of the Spearman rank coefficient is 0.99, nearly a perfect Spearman correlation of $r_{\mathrm{s}} = 1$, which indicates an ideal linear relationship.

This part can be summarized as follows. We have compared estimated parameters such as stellar extinction, stellar mass and emission line measurements with those obtained from the MPA/JHU catalogs. According to the analysis of relationships between these parameters, we conclude that the MF-ICA method is reasonable and reliable.

4.3 Empirical Relations

In this subsection, the accuracy of our method is tested in another way. The parameters estimated from the analysis of the continuum were compared with those estimated by using measured emission lines. We analyze the relationships between these parameters to investigate whether results derived from our method are reasonable.

4.3.1 Relations between Balmer features and stellar age

The value of 4000 Å break index (Balogh et al. 1999) can reflect the age of a galaxy. Higher $D_{\mathrm{n}}(4000)$ values are related to older, metal–rich galaxies, while lower values are related to younger stellar subpopulations in galaxies. The strength of Hδ_{A} (Worthey & Ottaviani 1997) is another age indicator. Strong Hδ_{A} absorption of a galaxy reflects a burst of star formation that occurred in the past $0 - 1$ Gyr. Therefore, our estimated ages of galaxies should increase with $D_{\mathrm{n}}(4000)$ values and decrease with Hδ_{A} values.

In Figure 8(a) and (b), correlations between ages, $D_{\mathrm{n}}(4000)$ and Hδ_{A} values are shown as expected relationships; $D_{\mathrm{n}}(4000) - \langle \log\ t \rangle_L$ trends with $r_{\mathrm{s}} = 0.85$ are strongly positive, and H$\delta_{\mathrm{A}} - \langle \log\ t \rangle_L$ trends with $r_{\mathrm{s}} = -0.77$ are strongly negative.

For a galaxy with emission lines, the Hα emission line corresponds to the instantaneous star formation rate (SFR) of a galaxy (Kennicutt & Evans 2012). Therefore the EW of Hα is also an indictor of age, which would be larger for younger galaxies.

Figure 8(c) shows that the light-weighted stellar age $\langle \log\ t \rangle_L$ correlates negatively with EW(Hα) ($r_{\mathrm{s}} = -0.79$), as we expect. These relations reflect that the stellar ages $\langle \log\ t \rangle_L$ we obtained by our spectral synthesis are reasonable.

4.3.2 Stellar mass and velocity dispersion

According to the viral theorem, for constant mass surface density, the stellar mass (log M_*) is expected to be positively correlated with the stellar velocity dispersion (log σ). In a sample of old galaxies, σ is related to galaxy mass through the Faber-Jackson relation. Moreover, the stellar velocity dispersion of young, star forming galaxies is contributed from the bulge and disk, thus it is related to galaxy mass through the Tully-Fisher relation. In Figure 8(d), we plot our estimates of stellar mass (log M_*) with velocity dispersion (log σ). The $M_* - \sigma$ relation shows a strong positive trend with $r_{\mathrm{s}} = 0.82$ as we expect, which suggests our synthesis results are physically meaningful.

We have analyzed correlations between physical parameters obtained from stellar populations, such as stellar ages and stellar masses, with independent quantities. The strong correlations between $\langle \log\ t \rangle_L - D_{\mathrm{n}}(4000)$, H$\delta_{\mathrm{A}}$, EW(H$\alpha$) and $M_* - \sigma$ suggest that the parameters derived by our spectral synthesis approach through the MF-ICA method are reasonable and meaningful.

5 APPLICATION TO SPECTRA OF GALAXIES WITH HIGHER REDSHIFT

Optical galaxy redshift surveys are not only vitally important in cosmology, but also critical to understanding physical processes related to galaxy formation and/or evolution (Fang et al. 2015). In the last few years, redshift surveys, such as the 2dF Galaxy Redshift Survey (2dFGRS; Colless et al. 2001) and SDSS, have measured redshifts of millions of low redshift galaxies (with a median redshift of $z = 0.1$). With larger aperture telescopes, a new generation of redshift surveys, such as DEEP2, BigBOSS and LAMOST, will measure galaxies with higher redshifts (Davis et al. 2007; Schlegel et al. 2009; Kong & Su 2010; Zou et al. 2011). The motivation for this work is that we want to provide an easy-to-use full-spectrum fitting package and determine parameters for spectra collected by the LAMOST extragalactic surveys (Kong & Su 2010). Since the regular spectroscopic survey of LAMOST is just beginning, we apply our synthesis approach to the spectra of galaxies from the DEEP2 survey, which has a similar S/N as spectra from LAMOST (Luo et al. 2015).

In the Extended Groth Strip (EGS) field, utilizing the Deep Imaging Muti-object spectrograph (DEIMOS) mounted on the Keck 10 m telescope, the DEEP2 galaxy redshift survey provides spectral data from galaxies with redshifts from 0 to 1.4. DEIMOS has a high-revolution grating of 1200 line mm^{-1}, covering the range $6500 -$

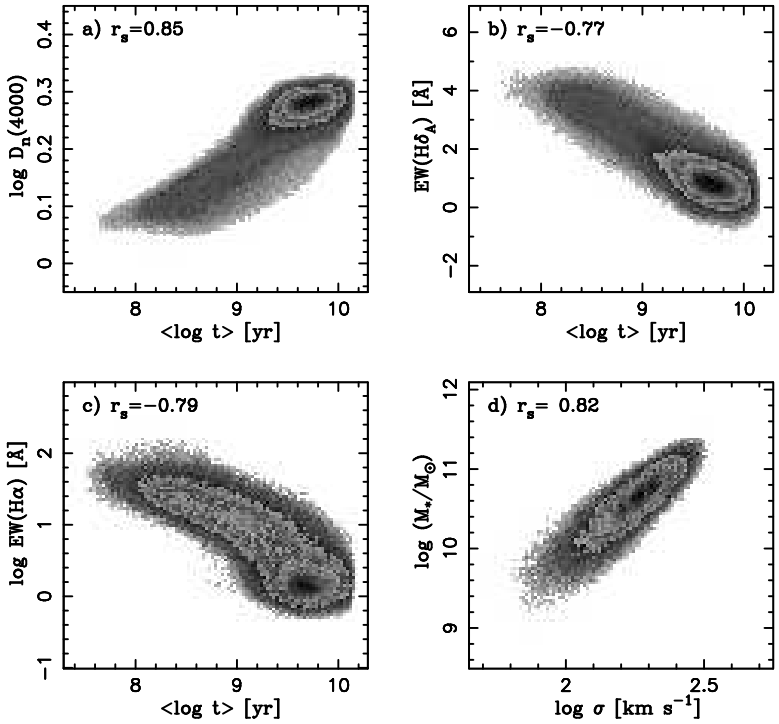

Fig. 8 Relations of the 4000 Å break index versus the light-weighted mean stellar age (a); the H$\delta_{\rm A}$ index versus the light-weighted mean stellar age (b); the EW of Hα versus the light-weighted mean stellar age (c); the comparison of our estimated stellar mass ($\log M_*$) with velocity dispersion ($\log \sigma$) (d). The number in the top left corner of each panel is the Spearman rank correlation coefficient $r_{\rm s}$.

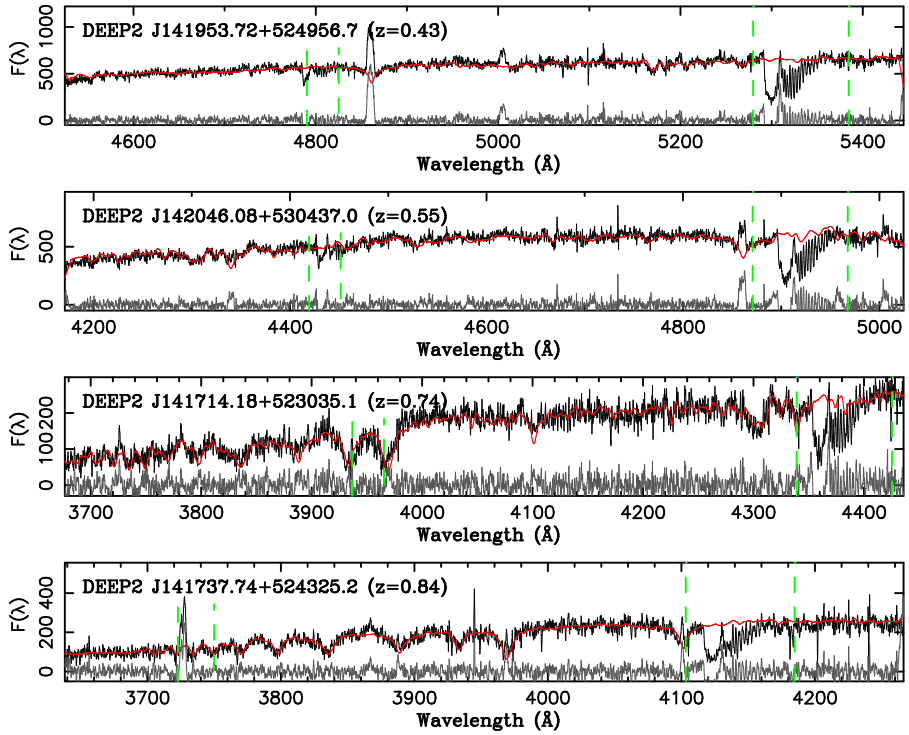

Fig. 9 The spectral fitting results of some galaxies in our DEEP2 sample covering a range of redshifts, which are labeled in the top left corner of each panel. The black lines show the observed spectrum, red lines show the modeled stellar spectrum, and grey lines show the residual spectrum. We also mask the "telluric absorption" regions between dashed lines (observed frame: $7750-7700$Å and $6850-6900$Å).

9100 Å, with a spectral resolution of $R = \lambda/\Delta\lambda \sim 6000$ (Faber et al. 2003). In our study, we only analyze galaxies with redshift quality $Q \geq 3$. Thus, we obtained 9501 galaxies with $Q \geq 3$ in the EGS, corresponding to a median redshift of 0.74. Details about extraction of spectra for these galaxies can be found in Davis et al. (2007). Finally, we obtained about 1400 sources with $\mathrm{S/N} > 3$, and show some examples of the fitting in Figure 9. It can be seen that our MF-ICA fitting method works well, and we will analyze their physical properties in future work.

6 SUMMARY

In this work, we have presented the MF-ICA method to compress the CB07 SSP library into a few ICs that can act as templates to fit observed galaxy spectra. Although many statistical multivariate data processing techniques are available, MF-ICA seems to be among the most useful, since it has the capability of providing good estimates of the results by selecting proper parameters. The goal of our project is to estimate physical properties quickly and accurately for a large sample of galaxies. By using the MF-ICA algorithm, we can fit an observed spectrum of a galaxy in only a few seconds, which is efficient in terms of time for analysis of galaxy spectra observed by large-area surveys, such as LAMOST and BigBOSS.

We have tested our method to fit simulated and SDSS DR8 galaxy spectra. Simulations show that important parameters of galaxies can be accurately recovered by our method, such as stellar contents, SFHs, starlight reddening and stellar velocity dispersion.

We have compared parameters estimated from our fitting technique to those obtained from the MPA/JHU group for DR8 galaxies. These physical parameters and measurements are in good agreement. We also analyze the correlations between physical parameters obtained from stellar populations with independent quantities. We find strong correlations between $M_* - \sigma$, $\langle \log\, t \rangle_L - D_{\mathrm{n}}(4000)$, $\mathrm{H}\delta_{\mathrm{A}}$ and EW(Hα).

In future studies, we intend to apply our fitting technique to other large databases, such as the LAMOST ExtraGAlactic Surveys (LEGAS) and the DEEP2 galaxy redshift survey. We have fitted more than 1400 DEEP2 galaxy spectra. Our next step will be to analyze their physical properties.

Acknowledgements We are grateful to the anonymous referee for making constructive suggestions to improve the paper. We thank Stephane Charlot for providing the unpublished CB07 stellar population synthesis models and helpful discussions. This work is supported by the Strategic Priority Research Program "The Emergence of Cosmological Structures" of the Chinese Academy of Sciences (No. XDB09000000), the National Basic Research Program of China (973 Program) (2015CB857004), and the National Natural Science Foundation of China (NSFC, Nos. 11225315, 1320101002, 11433005 and 11421303).

References

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29
Allen, J. T., Hewett, P. C., Richardson, C. T., Ferland, G. J., & Baldwin, J. A. 2013, MNRAS, 430, 3510
Balogh, M. L., Morris, S. L., Yee, H. K. C., Carlberg, R. G., & Ellingson, E. 1999, ApJ, 527, 54
Bell, A. J., & Sejnowski, T. J. 1995, Neural Computation, 7, 1129
Bica, E. 1988, A&A, 195, 76
Bressan, A., Chiosi, C., & Tantalo, R. 1996, A&A, 311, 425
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Cardoso, J.-F., & Souloumiac, A. 1993, IEE Proceedings F (Radar and Signal Processing, 140, 362
Charlot, S. & Bruzual, A. G. 2007, in preparation
Charlot, S., & Fall, S. M. 2000, ApJ, 539, 718
Chen, Y., Bressan, A., Girardi, L., et al. 2015, MNRAS, 452, 1068
Chen, Y.-M., Kauffmann, G., Tremonti, C. A., et al. 2012, MNRAS, 421, 314
Cid Fernandes, R., Mateus, A., Sodré, L., Stasińska, G., & Gomes, J. M. 2005, MNRAS, 358, 363
Cid Fernandes, R., Sodré, L., Schmitt, H. R., & Leão, J. R. S. 2001, MNRAS, 325, 60
Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039
Conroy, C. 2013, ARA&A, 51, 393
Cooper, M. C., Aird, J. A., Coil, A. L., et al. 2011, ApJS, 193, 14
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 1197
Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, ApJ, 660, L1
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Faber, S. M. 1972, A&A, 20, 361
Faber, S. M., Phillips, A. C., Kibrick, R. I., et al. 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes, ed. M. Iye & A. F. M. Moorwood, 1657
Fang, G.-W., Ma, Z.-Y., Chen, Y., & Kong, X. 2015, RAA (Research in Astronomy and Astrophysics), 15, 819
Girardi, L., Bressan, A., Bertelli, G., & Chiosi, C. 2000, A&AS, 141, 371
Højen-Sørensen, P. A., Winther, O., & Hansen, L. K. 2001, In Advances in Neural Information Processing Systems 13 (NIPS2000), 542
Højen-Sørensen, P. A., Winther, O., & Hansen, L. K. 2002, Neurocomputing, 49, 213
Hyvärinen, A., Karhunen, J., & Oja, E. 2001, Independent Component Analysis (New York: John Wiley & Sons)
Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 33
Kennicutt, R. C., & Evans, N. J. 2012, ARA&A, 50, 531

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, ApJ, 556, 121
Kong, X., & Cheng, F. Z. 2001, MNRAS, 323, 1035
Kong, X., Charlot, S., Weiss, A., & Cheng, F. Z. 2003, A&A, 403, 877
Kong, X., & Su, S. 2010, in IAU Symposium, 262, ed. G. R. Bruzual & S. Charlot, 295
Kong, X., Lin, L., Li, J.-R., et al. 2014, Chinese Astronomy and Astrophysics, 38, 427
Liu, G.-C., Lu, Y.-J., Chen, X.-L., Du, W., & Zhao, Y.-H. 2013, RAA (Research in Astronomy and Astrophysics), 13, 1025
Lu, H., Zhou, H., Wang, J., et al. 2006, AJ, 131, 790
Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA (Research in Astronomy and Astrophysics), 15, 1095
Maino, D., Donzelli, S., Banday, A. J., Stivoli, F., & Baccigalupi, C. 2007, MNRAS, 374, 1207
Maraston, C. 2005, MNRAS, 362, 799
Marigo, P., & Girardi, L. 2007, A&A, 469, 239
Miskin, J., & MacKay, D. 2001, Ensemble Learning For Blind Source Separation in Independent Component Analysis: Principles and Practice, eds. Roberts, S., & Everson R. (Cambridge: Cambridge Univ. Press), 209
Mittaz, J. P. D., Penston, M. V., & Snijders, M. A. J. 1990, MNRAS, 242, 370
Nolan, L. A., Harva, M. O., Kabán, A., & Raychaudhury, S. 2006, MNRAS, 366, 321
Ocvirk, P., Pichon, C., Lançon, A., & Thiébaut, E. 2006, MNRAS, 365, 46
Richards, J. W., Freeman, P. E., Lee, A. B., & Schafer, C. M. 2009, MNRAS, 399, 1044
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schlegel, D. J., Bebek, C., Heetderks, H., et al. 2009, arXiv:0904.0468
Shi, F., Liu, Y.-Y., Kong, X., et al. 2014, MNRAS, 444, L49
Su, S., Kong, X., Li, J., & Fang, G. 2013, ApJ, 778, 10
Tojeiro, R., Heavens, A. F., Jimenez, R., & Panter, B. 2007, MNRAS, 381, 1252
Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898
Vazdekis, A. 1999, ApJ, 513, 224
Worthey, G., Faber, S. M., Gonzalez, J. J., & Burstein, D. 1994, ApJS, 94, 687
Worthey, G., & Ottaviani, D. L. 1997, ApJS, 111, 377
Yip, C. W., Connolly, A. J., Szalay, A. S., et al. 2004, AJ, 128, 585
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579
Zou, H., Yang, Y.-B., Zhang, T.-M., et al. 2011, RAA (Research in Astronomy and Astrophysics), 11, 1093