

Multiband Segmentation of a Spectroscopic Line Data Cube: Application to the HI Data Cube of the Spiral Galaxy NGC 4254

Farid Flitti

*Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (LSIIT) UMR 7005,
Université Louis Pasteur (Strasbourg 1), Boulevard Sébastien Brant, B.P. 10413, 67412 Illkirch, Cedex, France
Email: flitti@lsiit.u-strasbg.fr*

Christophe Collet

*Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (LSIIT) UMR 7005,
Université Louis Pasteur (Strasbourg 1), Boulevard Sébastien Brant, B.P. 10413, 67412 Illkirch, Cedex, France
Email: collet@lsiit.u-strasbg.fr*

Bernd Vollmer

*Centre de Données Astronomiques de Strasbourg (CDS), UMR 7550, Observatoire Astronomique de Strasbourg,
11 rue de l'Université, 67000 Strasbourg, France
Email: bvollmer@astro.u-strasbg.fr*

François Bonnarel

*Centre de Données Astronomiques de Strasbourg (CDS), UMR 7550, Observatoire Astronomique de Strasbourg,
11 rue de l'Université, 67000 Strasbourg, France
Email: bonnarel@astro.u-strasbg.fr*

Received 28 May 2004; Revised 26 February 2005

A new method for the multiband segmentation of a spectroscopic line data cube is presented. This method is intended to help astronomers to handle complex spectroscopic line data cubes where the inspection of the channel and moment maps is difficult. Due to the Hughes phenomenon, the number of input images for the segmentation process is limited. Therefore, the spectrum of each pixel is fitted with a mixture of 6 Gaussians with fixed mean values and variances. The maps of the Gaussian weights are the input for a Markovian segmentation algorithm. The final segmentation map contains classes of pixels with similar spectral line profiles. The application of our method to the HI data cube of the Virgo spiral galaxy NGC 4254 shows that kinematically interesting regions can be detected and masked by our method.

Keywords and phrases: spectroscopic data cube reduction, Gaussian mixture model, hierarchical hidden Markov model, multi-component image, Bayesian segmentation, HI 21 cm line spiral galaxy NGC 4254.

1. INTRODUCTION

The natural output of line observations with radio interferometers like the VLA, Westerbork, ATCA, or Plateau de Bure are 3D data cubes, with the astronomical coordinates as x - and y -axes and the frequency or velocity channels as third axis. Within these data cubes, each image pixel contains an atomic or molecular line spectrum. Single-dish-line observations, of comparable sensitivity, of an equidistant grid of sky positions can also be treated as 3D data cubes (e.g., the Parkes survey). The complexity of the 3D structure contained in these data cubes increases with the sensitivity of the observations. In the radio domain, telescope sensitivities are in-

creasing tremendously with the upgrade of existing instruments and the building of new telescopes (EVLA, ALMA, SKA). This will lead to an enormous increase of available 3D data, which will be more and more complex.

Data cubes from radio line observations are well-suited test cases for new image processing techniques, because these data contain only one single line (e.g., a CO or HI line) which is shifted according to the radial velocity of the observed gas (Doppler effect). The standard method for the study of data cubes is the inspection of the channel maps by eye and the creation of moment maps after clipping the spectra at a level of 3σ , where σ is the rms noise in one channel:

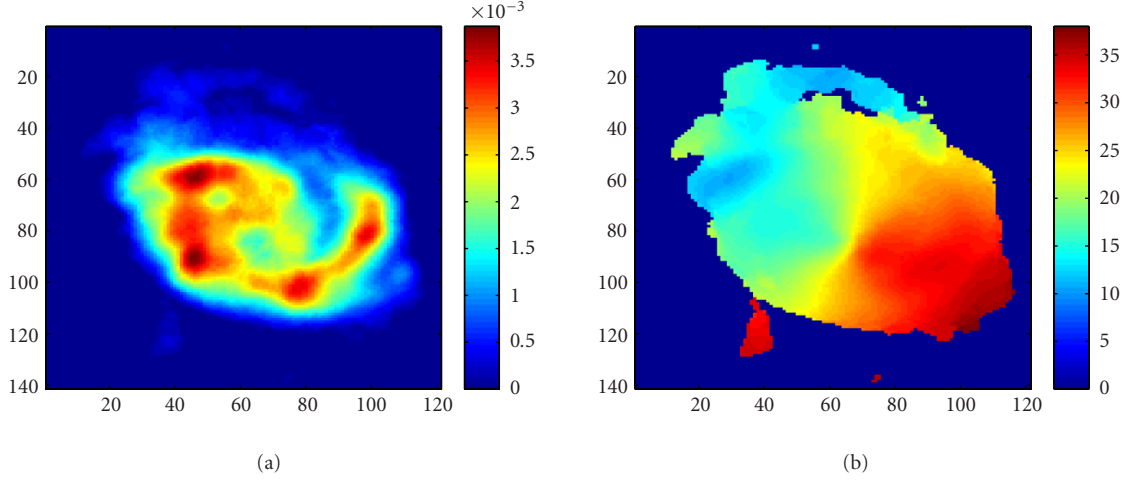


FIGURE 1: NGC 4254 [1]. (a) HI emission distribution (moment 0). The x -axis corresponds to right ascension, the y -axis to declination, the intensity is given in mJy. (b) HI velocity field (moment 1), the velocities are given in channel maps.

the zero-order moment is the integrated intensity, the first-order moment is the velocity field, and the second-order moment is the distribution of the velocity dispersion. As long as the 3D intensity distribution is not too complex, these maps give a fair impression of the 3D information contained in the cube. However, when the 3D structure becomes complex, the inspection of the velocity channels by eye becomes difficult and important information is lost in the moment maps, because they are produced by integrating the spectra, and thus do not reflect the individual line profiles. The method proposed in this paper is an attempt to provide an additional 2D segmentation map, which contains additional information on line profiles. This method aims at helping the astronomer in handling complex 3D data. With the results of our analysis, it is possible to focus the inspection of the channel maps on kinematically interesting regions.

For this purpose, we have chosen a target which is not too complex; NGC 4254, a spiral galaxy located in the Virgo cluster. The HI 21cm observations were made with the VLA [1]. The central velocity of the data cube is $v = 2408 \text{ km s}^{-1}$ at channel number 22 and the velocity resolution is 10 km s^{-1} per channel. For simplicity, we keep in this paper pixel numbers for the coordinate axis and channel numbers for the velocity axis. The rms noise of one channel map is $\sigma = 0.43 \text{ mJy}$.¹ The cube was clipped at an intensity level of $6\sigma = 2.58 \text{ mJy}$. Figure 1 shows the maps of the first two moments integrated over the whole data cube. The map of the HI emission distribution (moment 0, Figure 1a) shows an inclined gas disk with a prominent one-armed spiral to the west. The velocity field (moment 1, Figure 1b) is that of a rotating disk with perturbations to the north-east and to the north. The distribution of the velocity dispersion (moment 2), which is

not shown, is almost uniform with a small maximum in the galaxy center. The subsequent work was done on the same data cube clipped at 3σ .

The segmentation process on such a data cube requires a nontrivial modeling step based on Bayesian inference. Due to the curse of dimensionality (Hughes phenomenon), Bayesian segmentation can only be carried out on reduced data (principal or independent component analysis [2], projection pursuit [3], etc.). We therefore fit all spectra with a mixture of Gaussians (Section 2.1) and select the 6 most representative mean values (channel number of the maximum) and variances (widths) (Section 2.3). In this way the data cube, which had initially 42 channels, is reduced to 6 effective bands. In a second step, the weights of the Gaussians with fixed mean values and variances are determined by fitting again the observed spectra. The segmentation task, which is carried out on the weights, then consists of clustering the pixels into different classes (or labels) according to similar behaviors defined by a chosen criterion (Section 3). In this way, we obtain a segmentation map containing spatially homogeneous classes of pixels with a similar spectrum. We present the application of this method to the data cube of NGC 4254 in Section 4 including a comparison with other algorithms used for 3D image processing (Section 4.2). A summary and our conclusions are given in Section 5.

This method is intended to be additional and complementary to the traditional methods for the study of 3D data cubes, that is, the inspection of the channel maps by eye, the creation of moment maps, and the creation of position-velocity plots. Our method is intended to help astronomers to handle complex spectroscopic line data. The obtained segmentation map together with the moment maps can be used for a first inspection of the cube. In addition, masks of the different classes of the segmentation map can be produced to isolate kinematically interesting regions.

¹The jansky (abbreviated Jy) is a unit of radio flux density (the rate of flow of radio waves); $1\text{Jy} = 10^{-26} \text{ watts/m}^2/\text{Hz}$.

2. MULTIBAND ASTRONOMICAL IMAGE PROCESSING

When no prior knowledge is available for astronomical images, the problem of the physical meaning of the reduced data remains difficult. Apart from inspection by eye, radio astronomers usually use some parametrizations of the spectrum on each pixel to analyze huge data cubes [4], especially the intensity map (moment 0), velocity field (moment 1), and the distribution of the velocity dispersion (moment 2). In this paper, a Gaussian mixture with P components models the spectrum on each pixel as an alternative to the classical parametrizations. Our proposed modeling leads to multiband image processing, since only a reduced parameter set will be required to characterize the spectrum at each pixel ($3 \times P$ parameters instead of N values, N standing for the number of spectral bands; $P \ll N$).

2.1. Gaussian mixture model

We consider a multispectral image with N channels defined on a regular lattice S of size $H \times W$ pixels. This image can be viewed as $D = H \times W$ spectra \mathbf{y}_s , $s \in S$, each one of size N . Each spectrum is modeled by a P -component Gaussian mixture

$$\hat{\mathbf{y}}_s(j) = \sum_{k=1}^P \frac{a_k(s)}{\sqrt{2\pi\sigma_k^2(s)}} \exp\left(\frac{-1}{2\sigma_k^2(s)}(j - \mu_k(s))^2\right), \quad (1)$$

where $a_k(s)$ is the weight, $\sigma_k^2(s)$ the variance, $\mu_k(s)$ the mean associated with the k th Gaussian component of the s th spectrum, and j is the number of the frequency channel of the data cube.

2.2. Model estimation of parameters

The observed spectra of each pixel are fitted by a mixture of Gaussians using an expectation-maximization (EM) algorithm, which is an iterative method for maximum likelihood estimation. It has different applications [5, 6], but amongst all, the parameter estimation of the Gaussian density mixture model is probably one of the most widely encountered in the statistical pattern recognition community. It consists to approximate the probability density function (pdf) of an observed data set by a P -component Gaussian density mixture model [5, 6, 7]. Usually, the EM algorithm works on realizations of an unknown pdf. In our case, we assume that a spectrum represents already the pdf. Thus, we adapted the algorithm to fit directly the spectrum. To satisfy the assumptions required by the EM adaptation, each spectrum \mathbf{y}_s must be normalized to look like a pdf, that is, $\mathbf{y}_s(j) \geq 0$, for all j and $\sum_{j=1}^N \mathbf{y}_s(j) = 1$. Within the iterative process, the contribution of the k th component of the Gaussian mixture $\mathcal{R}^{[q]}(k, s, j)$ is calculated as follows:

$$\mathcal{R}^{[q]}(k, s, j) = \frac{\left(a_k^{[q]}(s)/\sqrt{2\pi\sigma_k^{2[q]}(s)}\right) \exp\left(\left(-1/2\sigma_k^{2[q]}(s)\right)(j - \mu_k^{[q]}(s))^2\right)}{\sum_{k=1}^P \left(a_k^{[q]}(s)/\sqrt{2\pi\sigma_k^{2[q]}(s)}\right) \exp\left(\left(-1/2\sigma_k^{2[q]}(s)\right)(j - \mu_k^{[q]}(s))^2\right)}, \quad (2)$$

where q is the number of the iteration step. These contributions are then inserted into the parameter calculation of the next iterative step:

$$\begin{aligned} a_k^{[q+1]}(s) &= \frac{\sum_{j=1}^N \mathbf{y}_s(j) \mathcal{R}^{[q]}(k, s, j)}{\sum_{j=1}^N \mathbf{y}_s(j)}, \\ \mu_k^{[q+1]}(s) &= \frac{\sum_{j=1}^N j \mathbf{y}_s(j) \mathcal{R}^{[q]}(k, s, j)}{\sum_{j=1}^N \mathbf{y}_s(j) \mathcal{R}^{[q]}(k, s, j)}, \\ \sigma_k^{2[q+1]}(s) &= \frac{\sum_{j=1}^N (j - \mu_k^{[q+1]}(s))^2 \mathbf{y}_s(j) \mathcal{R}^{[q]}(k, s, j)}{\sum_{j=1}^N \mathbf{y}_s(j) \mathcal{R}^{[q]}(k, s, j)}. \end{aligned} \quad (3)$$

This approach assumes that the weights $a_k(s)$ are all positive. After convergence, one obtains the parameters $a_k(s)$, $\sigma_k(s)$, and $\mu_k(s)$ to estimate $\hat{\mathbf{y}}_s(j)$ of (1).

2.3. Basis selection

The modeling using (1) requires 3 parameters for each Gaussian component and leads to $3 \times P$ parameters instead of N channels for each pixel. The Markovian classifier allows a maximum number of 9 input images due to the curse of dimensionality (Hughes phenomenon). With $P = 6$, we already obtain 18 reduced images (three parameters set $\{a_k(s), \sigma_k(s), \mu_k(s)\}_{k=1,\dots,6}$ for each $s \in S$). We therefore select the most representative $\sigma_k(s)$ and $\mu_k(s)$ among all Gaussian mixtures described in Section 2.2, which are then assumed to be constant. For this parameter selection, we used a clustering algorithm (K -means; [8]) on the set of vectors $\{(\mu_k(s), \sigma_k(s))\}_{s \in S}$. In this way, only the weights of the Gaussians have to be determined. Equation (1) becomes

$$\hat{\mathbf{y}}_s(j) = \sum_{k=1}^P \frac{\tilde{a}_k(s)}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-1}{2\sigma_k^2}(j - \mu_k)^2\right), \quad (4)$$

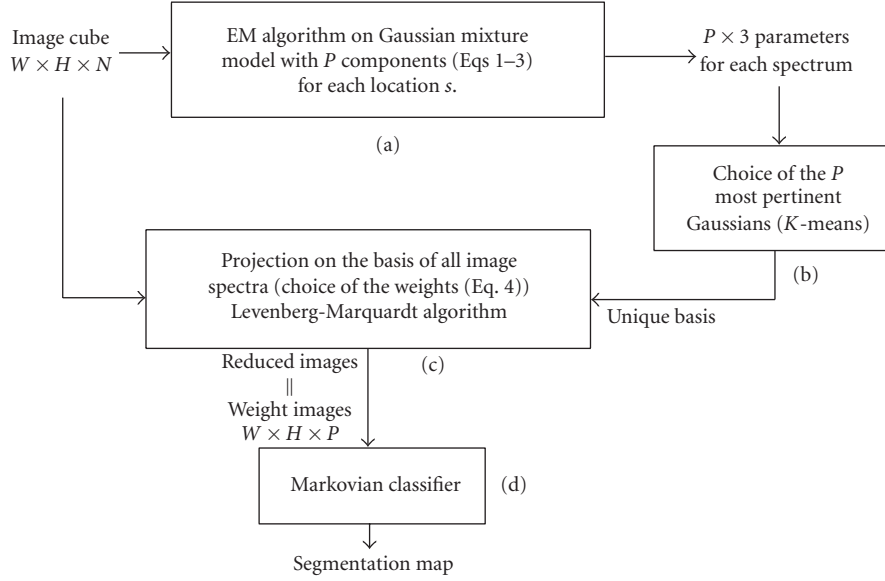


FIGURE 3: Data cube processing method summary. (a) Adapted EM algorithm to fit all spectra using (1). (b) Selection of the P most relevant Gaussians obtained in (a) using a K -means algorithm. (c) Projection on the basis chosen in (b) using a Levenberg-Marquardt algorithm. (d) The reduced images feed a Markovian classifier leading to the final segmentation map.

t^- designates the parent of site s , as illustrated in Figure 2. The likelihood of the observations z being in a state x at the bottom of the quadtree ($T^0 = S$) is expressed as the following product (assuming conditional independence): $P(z | x) = \prod_{s \in S} P(z_s | x_s)$, where for all $s \in S$, $P(z_s | x_s = \omega_i) \triangleq P_i^n(z_s)$ represents the likelihood of the data z_s . A multidimensional Gaussian pdf model is used to derive the latter expression.

Under these assumptions, the joint distribution $P(x, z)$ can be factorized as follows: $P(x, z) = P(x_r) \prod_{t \neq r} P(x_t | x_{t^-}) \prod_{t \in T^0} P(z_t | x_t)$ [16].

3.2. Bayesian segmentation

The expression of $P(x, z)$ allows us to calculate exactly and efficiently $P(x_t = \omega_i | z)$ for all nodes $t \in T$. The segmentation label map at the bottom of the quadtree is finally given by $\hat{x}_s = \arg \max_{\omega_i \in \Omega} P(x_s = \omega_i | z)$. This expression assumes that the model parameters (parameters of the Gaussian pdfs and interscale transition probabilities) are known. Such an estimation is obtained in an unsupervised way using the iterative conditional estimation (ICE) algorithm [18]. In practice, one alternates model parameter estimation and segmentation until convergence.

To summarize, one feeds a Markovian classifier with a multicomponent image $\{z_s = (\bar{a}_k(s))_{k=1, \dots, P}\}_{s \in S}$ (Figure 3). The output of this classifier is a segmentation map $\{\hat{x}_s\}_{s \in S}$ which contains spatially homogeneous classes of pixels with similar spectrum behavior.

4. RESULTS ON THE NGC 4254 CUBE

In order to test our data processing method explained in Sections 2, 3, and illustrated in Figure 3, we have applied it to reduce the dimensionality of the NGC 4254 cube, which is composed of $N = 42$ bands of size $H = W = 512$

pixels. As explained in Section 2.3, we fixed the number of Gaussian components to $P = 6$. It turned out that intensities in all velocities corresponding to channel numbers 10 to 35 are equally present in the data cube and that the widths of the corresponding lines do not vary significantly. Therefore the selected variances, which are fixed for the further processing of the data cube, have similar values: $\sigma_k = 1.5, 1.5, 1.6, 1.5, 1.4, 1.2$ and the fixed mean values (velocity of the maximum) are almost equidistant: $\mu_k = 12.3, 16.0, 20.3, 25.0, 29.2, 33.0$ (in channel number). In this way, the initial 42 velocity channels are reduced to 6 effective channels. The basis selection algorithm (Section 2.3) ensures that we do not lose important information.

Figure 4 shows the multivariate image (6 images in inverse video of the weights of the 6 Gaussians with fixed variances and mean values) obtained with our reduction technique. Due to the fact that the mean values of the basis are approximately equidistant, the multivariate images resemble binned channel maps. These reduced images feed the Markovian classifier, allowing us to isolate regions of similar spectra.

The most adapted number of classes depends on the complexity of the data cube. Selecting an overly small number of classes results in loss of information, selecting an overly large number leads to classes without physical meaning, which are sometimes made up of dispersed patches on the image. It is beyond the scope of this paper to discuss how to determine in general the most appropriate number of classes. However, the expert may have an expectation for the number of classes of interest. One then may explore different classification solutions with a variable number of classes around this expected value. In our case, we found that 7 classes are sufficient to describe the main features of the NGC 4254 data cube.

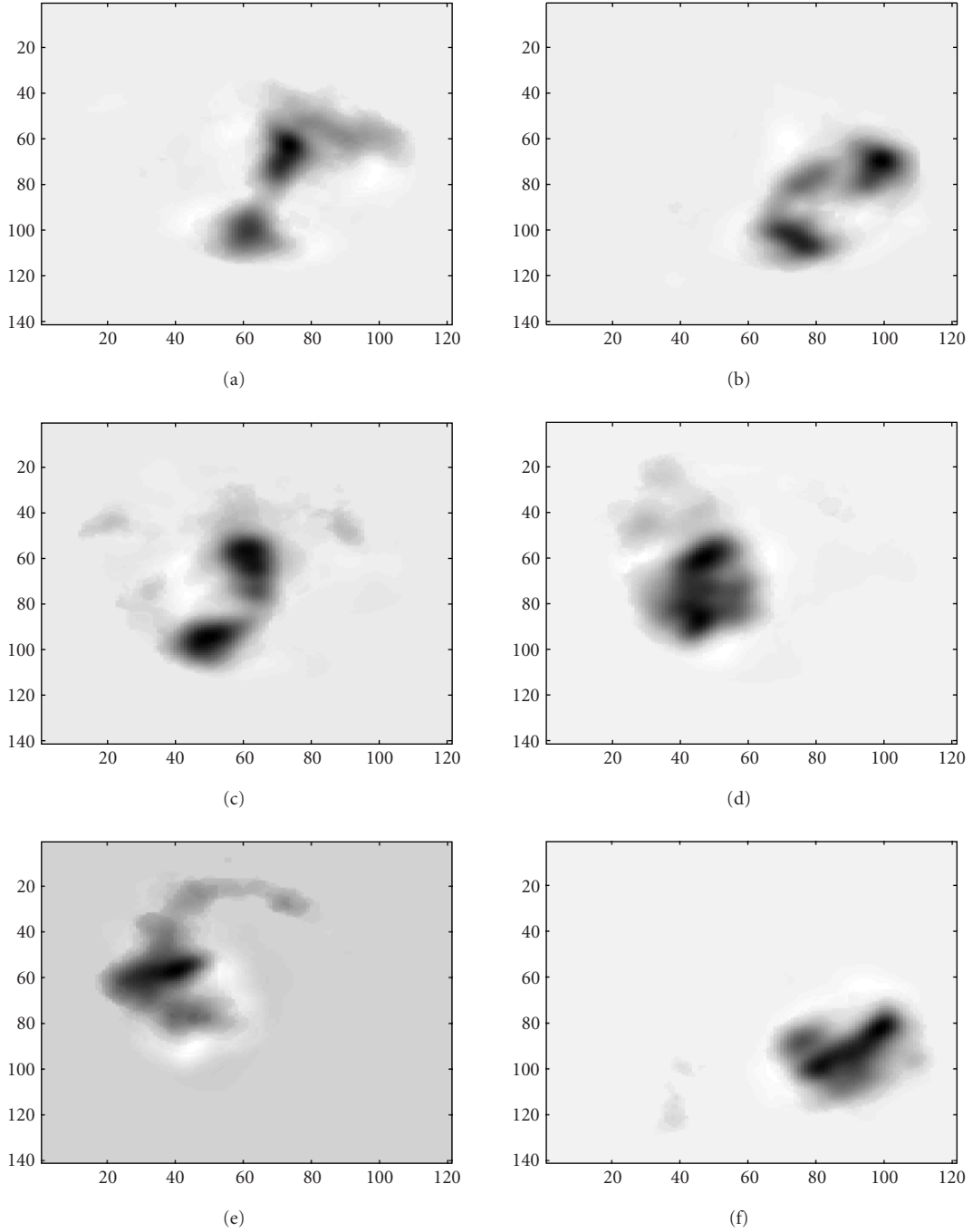


FIGURE 4: Maps of the weights of the 6 Gaussians with fixed mean values and variances. In the case of NGC 4254, this corresponds to a reduction of the original 42 channels to 6 effective channels. The x - and y -axes are the astronomical coordinates.

4.1. Physical interpretation of the segmentation results

The final segmentation map of 7 classes, together with the average observed and model spectra for each class, is shown in Figure 5. In general, the observed spectra are well fitted by a weighted combination of our Gaussian basis functions. The fact that the average observed spectra for all classes are single peaked shows that we actually obtain

classes of distinct spectral line profiles with our method. The fine structure of the model line profiles is due to the limited number of Gaussians of the basis. Since the widths of the fine structure of the average model spectra are always smaller than the widths of the average observed spectrum, we consider the fits acceptable. In the end, the important information is the average observed line profile.

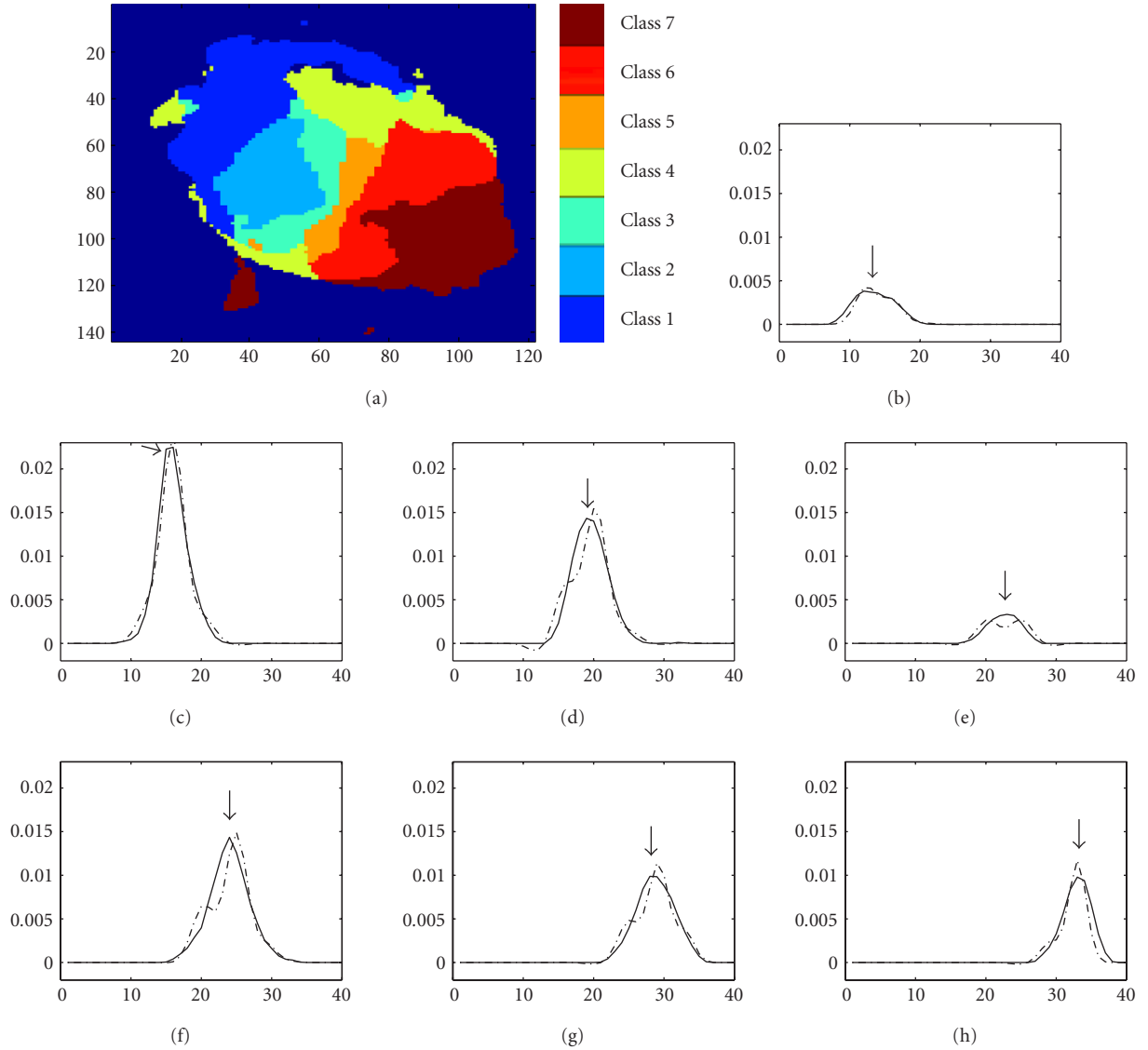


FIGURE 5: Results of segmentation of the maps shown in Figure 4. (a) Segmentation map. The x - and y -axes are the astronomical coordinates. (5b)–(5h) Average spectra for each class (solid: observed spectrum, dashed: model spectrum). The velocity channels are represented on the x -axis, the intensity on the y -axis (in Jy). The peaks of the line profiles are indicated by an arrow.

The comparison between the segmentation map and the velocity field (Figure 1b) shows that the segmentation is mainly done according to velocity, that is, the position of the peak in the 1D spectrum. This is due to the nature of the basis (almost equidistant in velocity; see Section 4). Only the Peak of the Averaged line profile of class 4 lies within the full-width half-maximum (FWHM) of the averaged line profile of class 5. Here the segmentation is also based on the peak value of the line profile which is much smaller for the line profiles of class 4 than for those of class 5. Thus for a sufficiently large number of classes (equal to or larger than the number of Gaussians of the basis), the segmentation is also based on the peak intensity of the line profiles. Two perturbations of the velocity field in the northern part of the galaxy can be identi-

fied (Figure 5a): (i) a departure from the symmetric velocity gradient near pixel (70,50), where the red/yellow region extends into the green region, and (ii) the blue region in the north of the galaxy. Both regions are recognized as distinct classes (1 and 4) in Figure 5.

We conclude that our data processing method is operating successfully. For this relatively simple data cube, the segmentation map (Figure 5) does not contain additional information to the velocity field (Figure 1b). However, our segmentation map might be used to produce masks to isolate the kinematically interesting regions in the north of the galaxy (classes 1 and 4). Phookun et al. [1] derived a rotation curve from the data cube, made position-velocity diagrams of the northern region, and subtracted the emission provided from

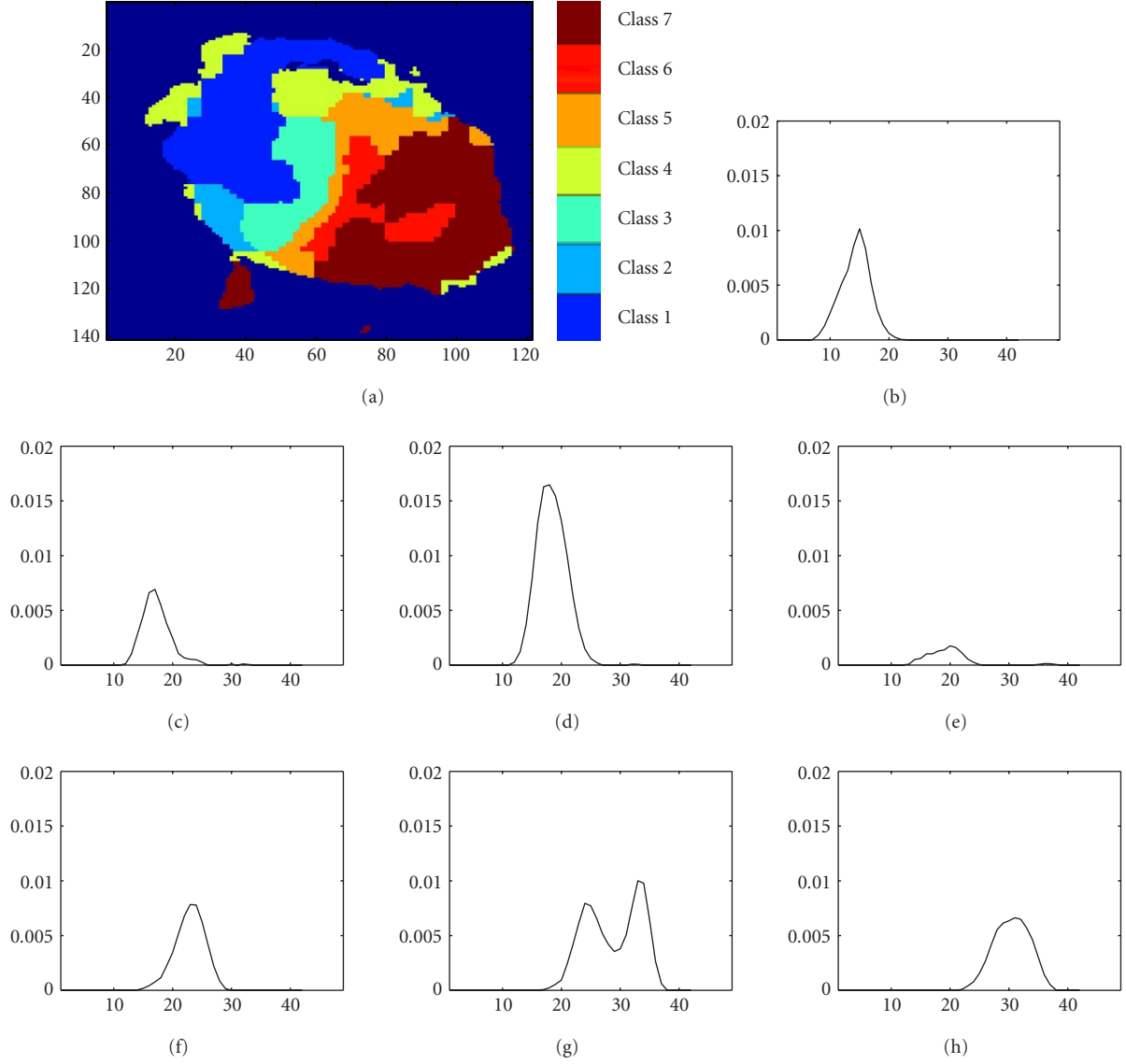


FIGURE 6: Principal component analysis (PCA) of NGC 4254. Only the images corresponding to the six largest eigenvalues are kept. (a) segmentation map. The x - and y -axes are the astronomical coordinates. (6b)–(6h) Average spectra for each class (solid: observed spectrum, dashed: model spectrum). The velocity channels are represented on the x -axis, the intensity on the y -axis (in Jy).

the rotating gas disk to isolate kinematically perturbed regions. Our approach is not intended to replace or compete with these sophisticated methods.

4.2. Comparison with different approaches

These results can be compared to different multiband image processing methods. The oldest and most popular method is principal component analysis (PCA) [8]. It calculates a linear mapping which maximizes the data scatter in the projection subspace. The transformation matrix can be easily computed by taking the eigenvector decomposition based on the data covariance matrix. In keeping only the eigenvectors corresponding to the largest eigenvalues, we project the data cube on this subspace in order to obtain a reduced repre-

sentation of the data. This technique is an unsupervised one. Since in our case the channels are almost uncorrelated, the PCA approximately ranks the channels according to their energy contained in the signal. Thus, the six largest eigenvalues corresponding to the six channels of largest power feed the Markovian classifier. The classes of the final segmentation map (Figure 6) simply reflect the flux contained in only 6 of the original 42 channels. Since line information is contained in 25 of the 42 channels, this method loses the information contained in 19 channels. In addition, the peaks of the line profiles within one class can be different by more than the width of the averaged line profiles. This leads to very broad or even double-peaked averaged line profiles (classes 2, 4, 6 in Figure 6) without physical meaning.

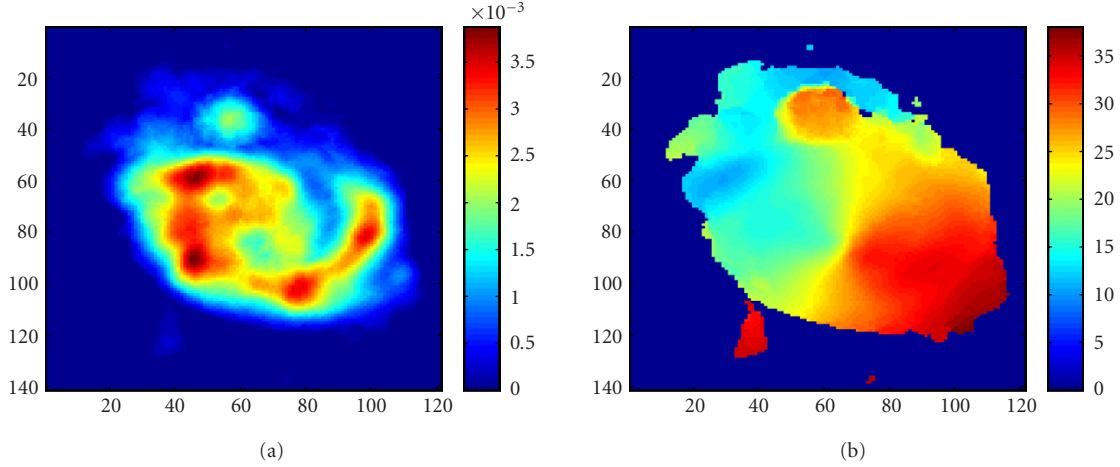


FIGURE 7: NGC 4254 data cube after addition of a region with an artificial line creating a double-line profile (center at (60, 35)). (a) HI emission distribution (moment 0). The x -axis corresponds to right ascension, the y -axis to declination, the intensity is given in mJy. (b) HI velocity field (moment 1), the velocities are given in channel maps.

Another technique based on independent component analysis (ICA) [2] calculates a linear mapping such that the components of the reduced vectors are as independent as possible. ICA finds subspaces in which the new subimages are independent instead of just uncorrelated as for the PCA. There is a large set of algorithms performing ICA based on tensorial techniques, mutual information minimization, and non-Gaussianity maximization, which are well presented in [2].

The advantage of our approach compared to PCA and ICA is twofold: first, the spectral information is not mixed, and second, the segmentation classes are physically meaningful, that is, their pixels have similar spectra.

Recently, the separation of astrophysical source maps from multichannel observations has become of great interest [19, 20]. A summary of different techniques for source separation can be found in [19]. Nevertheless, in our opinion, these techniques suffer from the lack of physical meaning of the resulting images since observations are not always the result of mixed independent images.

4.3. Double-line profiles

We have seen that our segmentation method produces a discrete velocity field and masks the regions of asymmetries in the velocity field. The principal result is thus the masking of different regions of interest. All these regions can already be approximately separated by eye on the moment maps (Figure 7). On the other hand, a feature that cannot be detected easily by the moment maps is double-line profiles. This information is lost by the velocity averaging process when the moment maps are produced. In order to investigate if our segmentation method is able to detect regions of double-line profiles, we added an artificial line to the spectra in a circular region north of the galaxy center (with its center position (60, 35), a radius of 10 pixels, and a center velocity at channel number 28). The intensity of the arti-

cial line is maximum at the center and falls off radially (we used a Gaussian profile with a width of 5 pixels). Figure 7 shows the moment maps of this new data cube. The additional lines produce a local maximum in the HI emission map (Figure 7a) and a pronounced asymmetry in the velocity field (Figure 7b).

The basis for this modified data cube is $\sigma_k = 1.5, 1.5, 1.6, 1.6, 1.5, 1.2$ and $\mu_k = 12.3, 16.0, 20.2, 24.8, 29.0, 33.0$. It is almost identical to that of the original data cube. Figure 8 shows the multivariate image (6 images in inverse video of the weights of the 6 Gaussians with fixed variances and mean values) obtained with our reduction technique on the modified data cube. The final segmentation map of this modified data cube, together with the average observed and model spectra, is presented in Figure 9. Again, in the inner part of the galaxy, a discrete velocity field is produced. The northern perturbed part of the velocity field is also recognized (class 1). The region of double-line profiles appears as a new separate class (class 5). The averaged observed and fitted line profiles clearly show a double peak with a component of class 1 (central channel number 17) and the new, artificially added component (central channel number 28). This double-line profile is clearly distinct from the fine structure of the modeled lines (see also Section 4.1), which is always contained within the width of the observed line profile. Thus, our method is able to detect regions of double-line profiles. The 42-channel maps of the original cube can then be inspected in detail using the mask of the double-line region.

The double-line profile could have been detected by the inspection of the channel maps by eye and/or by making multiple position-velocity plots. However, without a prior knowledge of the location of the double-line profile, this can be quite long and painful, especially if the cube is more complex than our test cube. Despite the loss of spectral resolution (the cube is reduced from 42 channels to 6 effective channels), our method is still able to detect and mask

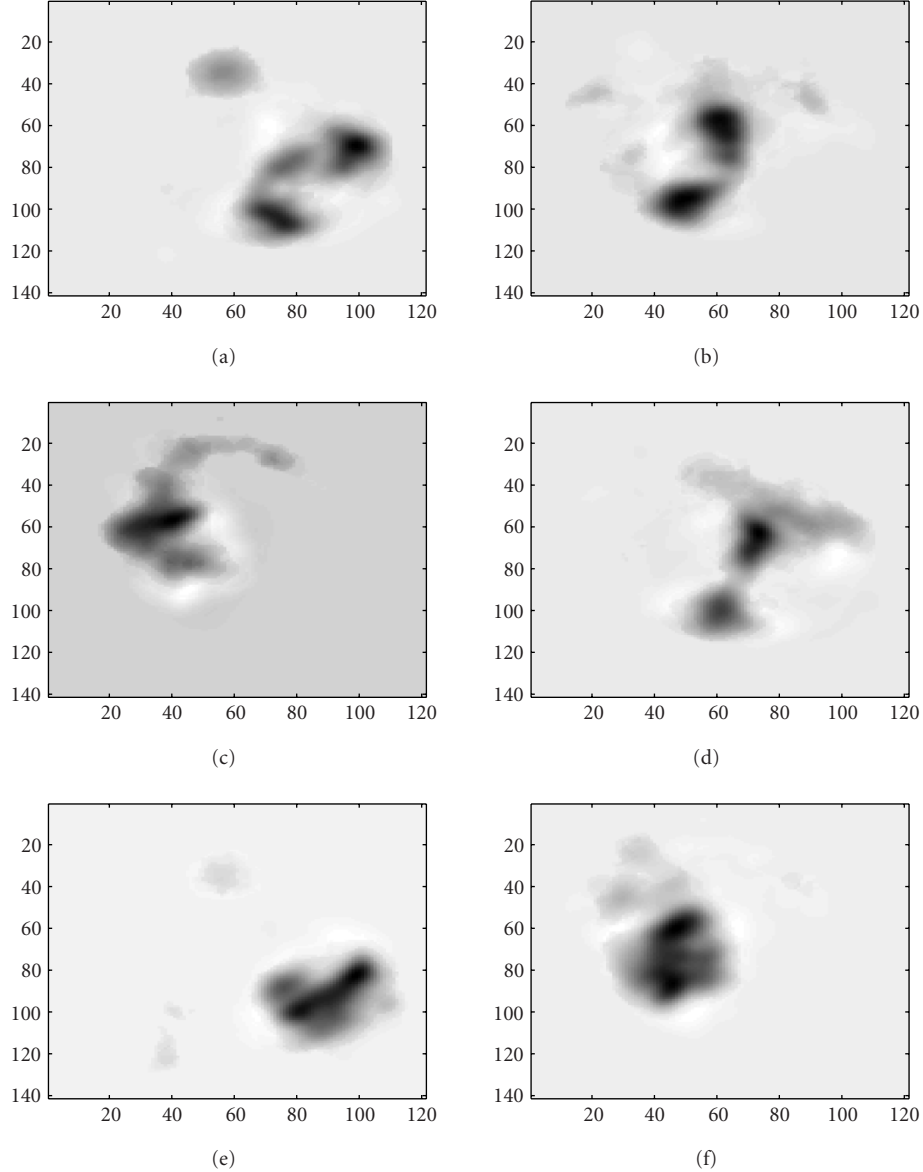


FIGURE 8: Maps of the weights of the 6 Gaussians with fixed mean values and variances. A region of artificial line profiles creating a region of double-line profiles is added (center at (60, 35)), which is clearly visible in map (8a). The x - and y -axes are the astronomical coordinates.

multiple-line profiles. We think that this is very useful for an astronomer who has to handle complex 3D data cubes. The further investigation is simplified by the knowledge and possible masking of regions of multiple line profiles, which might be carried out with more sophisticated methods applied to the original data cube.

5. SUMMARY AND CONCLUSION

A new method for the segmentation of multiband images of astronomical radio data cubes is presented. The observed spectra are first fit by a weighted combination of Gaussian components. The parameters of the Gaussians are the input of a Markovian segmentation algorithm. Due to the curse of

dimensionality (Hughes phenomenon), the number of input parameters is limited to 9. In our approach, we chose to fix the number of Gaussians at 6 and to set their mean values and variances to their 6 most representative values. The weights of these 6 Gaussians with fixed mean value and variance are determined by again fitting the observed spectra. In this way, the original number of channels of the data cube is reduced to 6 effective channels. A Markovian image segmentation is then done on the 6 maps of the weights of the Gaussians. The final result is a segmentation map where regions of similar spectral line profiles are assembled into different classes. The number of classes has to be determined by the user. The optimum number of classes depends on the complexity of the data cube.

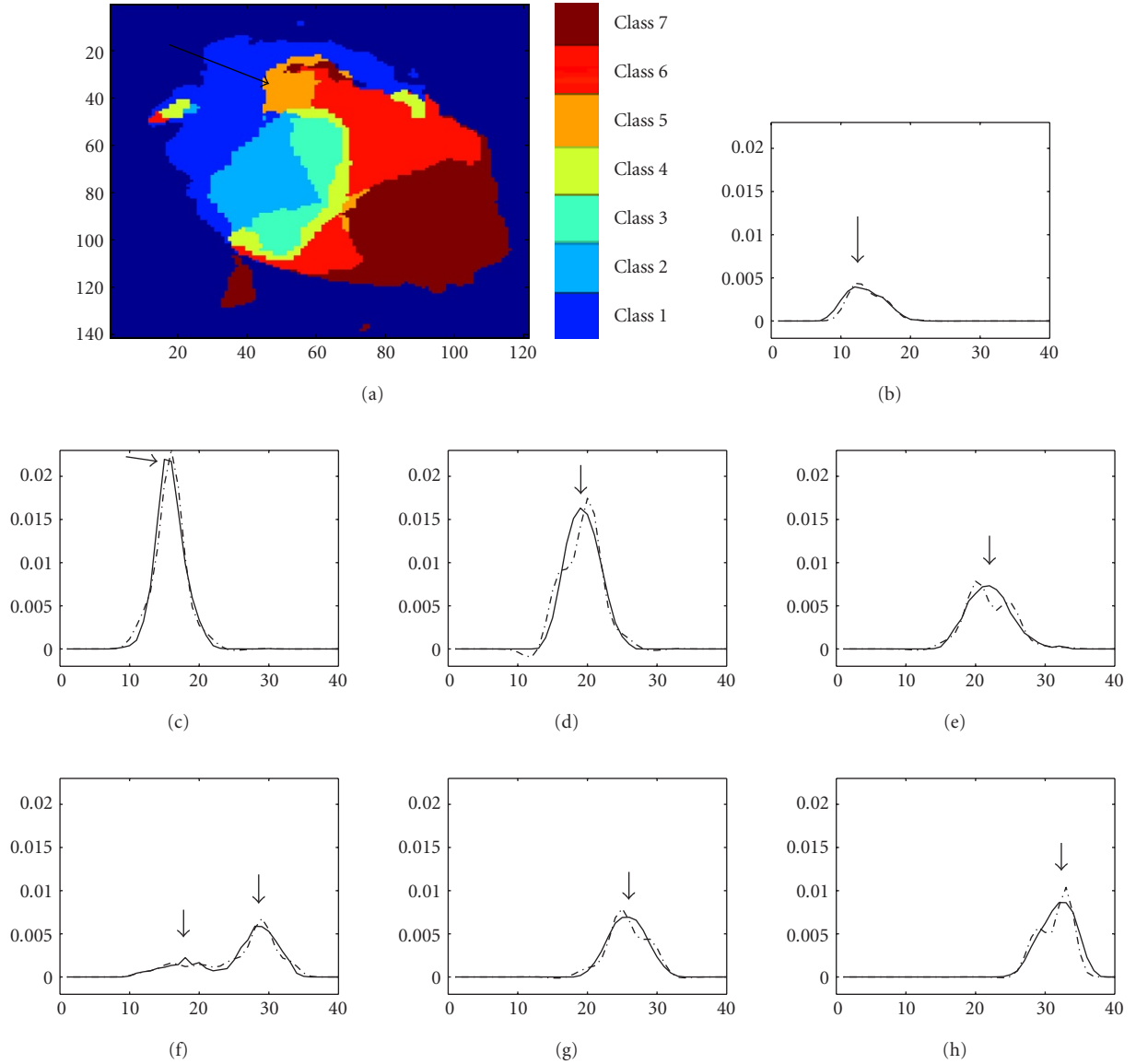


FIGURE 9: Results of segmentation of the maps shown in Figure 8. (a) Segmentation map. The x - and y -axes are the astronomical coordinates. (9b)–(9h) Average spectra for each class (solid: observed spectrum, dashed: model spectrum). The velocity channels are represented on the x -axis, the intensity on the y -axis (in Jy). The peaks of the line profiles are indicated by an arrow.

This procedure is applied to the HI 21 cm line data cube of the Virgo cluster spiral galaxy NGC 4254 [1]. Due to the intensity structure in the data cube, that is, the nature of the object, the 6 most representative mean values are almost equidistant and the 6 most representative variances are almost constant. Thus, the multivariate image of the weights of the 6 Gaussians corresponds to maps of 6 effective channels. For the segmentation, it turned out that the main features of the data cube are visible in the final segmentation map with 7 different classes. Since the final segmentation map contains regions of similar spectral line profiles, it resembles at first sight a binned velocity field. However, different classes contain lines of different maxi-

um intensity. Increasing the number of classes leads to new classes of about the same central velocity with different maximum intensities. We optimized the number of classes by comparing the segmentation map to the channel and moment maps. We plan to investigate if there is a way to determine the optimum number of classes in a more objective way.

In a second approach, we added a region of an artificial line to the data cube, creating a region of a double-line profile. This region is clearly detected by our segmentation algorithm as a new class with a double-line. This makes us confident that our method can give useful information which is complementary to the traditional moment maps.

The proposed method aims at helping astronomers to handle complex data cubes where an inspection of the channel maps by eye is a difficult task. Once the region of interest is identified by our method, a mask can be easily produced to inspect only this region in the channel maps

The two free parameters of our method are the number of Gaussians (with a maximum number of 9) and the number of classes. We plan to investigate the optimization of these parameters in an objective way by applying our method to different radio data cubes of various complexity.

ACKNOWLEDGMENTS

We would like to thank the anonymous referees for helping us to improve this paper substantially. This research has benefited from financial support from the French Ministère de la Recherche (ACI-GRID IDHA (2001–2004)).

REFERENCES

- [1] B. Phookun, S. N. Vogel, and L. G. Mundy, "NGC 4254: a spiral galaxy with an $m = 1$ mode and infalling gas," *Astrophysical Journal*, vol. 418, no. 1, part 1, pp. 113–122, 1993.
- [2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, NY, USA, 2001.
- [3] P. J. Huber, "Projection pursuit with discussion," *Annals of Statistics*, vol. 13, no. 2, pp. 435–525, 1985.
- [4] R. A. Perley, F. R. Schwab, and A. H. Bridle, "Synthesis imaging in radio astronomy," in *Astronomical Society of the Pacific Conference*, vol. 6 of *ASP Conference Series*, San Francisco, Calif, USA, 1989.
- [5] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, New York, NY, USA, 1997.
- [6] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Tech. Rep. ICSI-TR-97-021, University of Berkeley, Berkeley, Calif, USA, 1997.
- [7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2001.
- [9] E. Walter and L. Pronzato, *Identification of Parametric Models from Experimental Data*, Communications and Control Engineering Series, Springer, London, UK, 1997.
- [10] D. A. Van Dyk, "Hierarchical models, data augmentation and MCMC," in *Statistical Challenges in Modern Astronomy III*, G. J. Babu and E. D. Feigelson, Eds., pp. 41–56, Springer, New York, NY, USA, 2002.
- [11] M. Mignotte, C. Collet, P. Pérez, and P. Bouthemy, "Bayesian inference and optimization strategies for some detection and classification problems in sonar imagery," in *Nonlinear Image Processing X*, vol. 3646 of *Proceedings of SPIE*, pp. 14–27, San Jose, Calif, USA, January 1999.
- [12] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, no. 6, pp. 721–741, 1984.
- [13] C. Graffigne, F. Heitz, P. Pérez, F. J. Prêteux, M. Sigelle, and J. B. Zerubia, "Hierarchical Markov random field models applied to image analysis: a review," in *Neural, Morphological, and Stochastic Methods in Image and Signal Processing*, vol. 2568 of *Proceedings of SPIE*, pp. 2–17, San Diego, Calif, USA, July 1995.
- [14] J.-N. Provost, C. Collet, P. Rostaing, P. Pérez, and P. Bouthemy, "Hierarchical Markovian segmentation of multispectral images for the reconstruction of water depth maps," *Computer Vision and Image Understanding*, vol. 93, no. 2, pp. 155–174, 2004.
- [15] W. Pieczynski, "Statistical image segmentation," *Machine Graphics and Vision*, vol. 1, no. 1-2, pp. 261–268, 1992.
- [16] J.-M. Laferté, P. Pérez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 390–404, 2000.
- [17] M. Luetgten, *Image processing with multiscale stochastic models*, Ph.D. thesis, MIT Laboratory of Information and Decision Systems, Cambridge, Mass, USA, May 1993.
- [18] W. Pieczynski, "Champs de Markov cachés et estimation conditionnelle itérative," *Traitement du signal*, vol. 11, no. 2, pp. 141–153, 1994.
- [19] E. Salerno, A. Tonazzini, E. E. Kuruoglu, L. Bedini, D. Heranz, and C. Baccigalupi, "Source separation techniques applied to astrophysical maps," in *Proc. 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES '04)*, R. J. Howlett, M. Gh. Negoita, and L. C. Jain, Eds., vol. 3 of *Lecture Notes in Artificial Intelligence* 3215, pp. 426–432, Springer, Wellington, New Zealand, September 2004.
- [20] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. 11th European Signal Processing Conference (EUSIPCO '02)*, vol. 1, pp. 561–564, Toulouse, France, September 2002.

Farid Flitti received the State Engineer degree in 1996, the M.s. degree in 1999, both in electrical engineering from the École Nationale Polytechnique d'Alger, Algiers, and the DEA degree in signal processing from the Université Paris-Sud in 2001. He is currently pursuing the Ph.D. degree in image processing at the Université Louis Pasteur (Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédetection, UMR-CNRS 7005), Strasbourg, France. He is interested in signal and image processing, mainly in statistical models, multispectral image reduction, image segmentation and fusion.



Christophe Collet was born in 1966 in France. He graduated from the Université Paris-Sud, Orsay, in 1989. He received the M.S. degree in signal processing, and received a Ph.D. degree in image processing from the University of Toulon in 1992. He spent 8 years at the French Naval Academy and was the Chairman of the Laboratory GTS "Groupe de Traitement du Signal" from 1994 to 2000, where he developed hierarchical Markovian approaches for sonar image analysis. Since 2001, he has held a Full Professor position at Strasbourg University (LSIIT, UMR-CNRS 7005). His major research interests include multi-image segmentation and classification with hierarchical approaches (wavelet decomposition, multigrid optimization, multiscale modeling), Bayesian inference, Markovian approaches for pattern recognition, Bayesian networks, with a particular focus on astronomy (hyperspectral) and medical (multimodal) images.



Bernd Vollmer He received his Ph.D. degree in 2000 at Paris Observatory. Between 2000 and 2003, he was a Postdoc at the Max-Planck-Institut für Radioastronomie, Bonn, Germany. Since 2005, he has been a Staff Astronomer at the Centre de Données Astronomiques de Strasbourg (CDS), France. He has been a Theorist and Numerical Modeler, as well as an Observer. He has done numerous N-body models of the interaction of cluster spiral galaxies with their environment, including both gravitational interactions and ram pressure stripping. He compares simulations with observed HI gas distributions and velocity fields of Virgo cluster galaxies. He works also on the gas dynamics in the Galactic Center.



François Bonnarel studied astronomy at Marseilles' Observatory where he got a Ph.D. degree in 1986. After being a Postdoc in Max-Planck-Institut für Astronomie in Heidelberg (FRG), he got a position of Computer Engineer in Strasbourg Observatory in 1987 where he has been working so far. His main interests and activities deal with astronomical databases, image databases, astronomical images management, image compression, image processing for astronomy, virtual observatory, and image services interoperability. He is the main author, or coauthor of more than seventy publications in astronomy.

