ARTICLE IN PRESS

Statistical Methodology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Optimal Bayesian fusion of large hyperspectral astronomical observations

M. Petremand [a], A. Jalobeanu [b], C. Collet [a,*]

[a] *LSIIT UMR CNRS 7005, University of Strasbourg, France* [1]

[b] *Centro de Geofísica, University of Évora, Portugal*

ARTICLE INFO

Keywords:
Hyperspectral images
Bayesian inference
Data fusion
Astronomy
B-splines
Uncertainties

ABSTRACT

New generation integral-field spectrographs (IFS) such as MUSE will soon start observing distant astronomical objects with much higher spectral and spatial resolutions than today's instruments. The new hyperspectral observations will represent a huge amount of scientific data (up to 1.2 GB per each MUSE raw acquisition) whose analysis requires the development of dedicated processing methods. In addition, in the framework of long acquisition sessions (typically 80 h for deep-field observations acquired with the MUSE instrument) spread over several nights and split into one-hour exposures, the same field of view is observed under varying atmospherical and physical conditions (*e.g.* seeing, noise, spectral and spatial shifts, etc.) thus producing multiple hyperspectral cubes to analyze. In this paper, we propose a new data fusion method that aims at reconstructing a single data cube from this large CCD (raw) data set (up to 100 GB for a MUSE acquisition session) taking into account acquisition parameter variations and whose analysis becomes easier for astronomers. This last point clearly means that we need to produce a fused hyperspectral image together with associated uncertainties on each fused pixel. On the one hand, a forward model precisely describes the complex MUSE acquisition process and, on the other hand, the inverse problem is solved to yield the fused cube from the whole observation set. The reconstruction and the fusion of large raw observations are performed in a sequential way so as to minimize computing time and memory usage. This challenging task is performed in

* Corresponding author.
E-mail addresses: m.petremand@unistra.fr (M. Petremand), jalobeanu@uevora.pt (A. Jalobeanu), c.collet@unistra.fr (C. Collet).

[1] http://lsiit-miv.u-strasbg.fr/paseo.

1572-3127/$ – see front matter
doi:10.1016/j.stamet.2011.04.007

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

the rigorous Bayesian framework, which provides a fusion scheme that is more optimal (in the statistical sense) than existing 3D reconstruction methods used in astronomy while giving access to uncertainties (in the form of a precision matrix) associated with the fused image: results can then be reused by astronomers for further analysis. The global fusion scheme is validated on small-size simulated observations with varying acquisition conditions while real MUSE raw observations will not be available before 2013.

1. Introduction

The new integral-field spectrograph MUSE (Multi Unit Spectroscopic Explorer) [11] has been retained by the European Southern Observatory (ESO) as a second generation VLT (Very Large Telescope) instrument. MUSE, operational in 2012 on the VLT site at Paranal (Chile), will be mainly dedicated to the observation of deep universe thanks to its wide-field mode whose available spectral and spatial resolutions have never been reached so far. Indeed, a reconstructed MUSE hyperspectral observation is about $300 \times 300 \times 3600$ pixels and covers a $1' \times 1'$ field of view from 465 nm to 930 nm. However, the increasing resolution for deep-field observations naturally leads to huge data cubes (1.2 GB per MUSE observation) which complicates the analysis task and implies the use of dedicated and well-designed processing algorithms. Thus, the DAHLIA (Dedicated Algorithms for HyperspectraL Imaging in Astronomy) project federates 4 French research laboratories around the development of new tools for MUSE hyperspectral images (*e.g.* fusion [10], visualization [14], deconvolution, source separation and segmentation) in order to be able to make scientific measurements with this new generation of IFS. In the framework of the observation and the study of distant and faint astronomical objects, a MUSE acquisition session may last up to 80 h (*i.e.* 80 nights) so as to maximize the photon number collected by MUSE's sensors. However, the presence of cosmic rays [6] randomly hitting CCD matrices reduces the maximal exposure time to 1 h and thus may lead to 80 observation sessions of 1 h each. In the end, a deep-field raw (CCD) observation may then be composed of nearly 80×1.2 GB of data to be handled, analyzed and visualized which is challenging for modern processing algorithms. In addition, the 80 raw images have been acquired under varying observation conditions such as PSF (Point Spread Function), noise variances, spatial and spectral shifts, dead or saturated CCD pixels, sampling lattices, exposure times…because of different atmospheric perturbations, atmospheric diffractions, geometric disturbances, sensors and acquisition noises. Observations of the same field of view cannot therefore be easily compared, averaged or jointly processed and benefits of deep-field surveys with such IFS are lost. A fused version of this large, complex and heterogeneous observation set can nevertheless be properly built so as to reflect the complementarity of the 80 acquisition sessions in a single consistent image easier to analyze.

For the time being, hyperspectral fusion methods used within the astronomical community consist of combining and averaging observations in the model space without explicitly using any image formation models. For instance, a trilinear interpolation method [17] is currently implemented in the Data Reduction Software (DRS) of MUSE and will be systematically applied on raw observations as soon as they are acquired. Another widely used method combines 2D undersampled observations to form a super-resolved image [4] but takes into account neither the complex physical image formation process nor observation parameters such as PSF. Moreover, super-resolution for MUSE data is not necessary as this instrument is designed so that observations are well sampled (or oversampled) with respect both to average seeing conditions at the VLT site as well as IFS optics.

Our goal is then to propose a new fusion scheme whose optimality (in the statistical sense) is expected to be higher than that of existing ones thanks to the inversion, in the rigorous Bayesian framework [5,13], of the data acquisition process with respect to varying observation parameters. The maximum *a posteriori* approach is used to solve the inverse problem and goes through an accurate description of the image formation model as well as the choice of a prior model suited to large data. The uncertainties associated with the fused hyperspectral image are available in the Bayesian context

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

and can then be used by astronomers as a relevant prior for further investigations on the fusion result (*e.g.* denoising, source separation, segmentation, astrometry and photometry computations). This is a challenging but required approach to correctly manage MUSE observations and take advantage of large data sets acquired by this instrument. Such an algorithm has already been developed and validated for the reconstruction and the fusion of 2D images [9] but its application to the hyperspectral case is not straightforward due to data complexity and size which are critical parameters for Bayesian approaches.

The first step of the developed fusion algorithm all comes down to precisely writing the forward model describing the image formation process specific to each raw observation (Section 2.1) and then to define both reachable spectral and spatial resolutions of the fused image we aim at building (see Sections 2.2 and 2.3). In a second step, the forward model is inverted in a Bayesian context (Section 3.1) to reconstruct and fuse the n observations one by one. This sequential approach, detailed in Section 3.2, allows to deal with large hyperspectral observations and acquisition parameters but does not exempt algorithms and data structures from being designed so as to optimize memory usage while minimizing computationally expensive file access. Uncertainties related to the fused image are directly obtained from this sequential algorithm in the form of a precision matrix that can be quickly inverted to yield a covariance matrix (Section 3.3). Finally, our fusion method is validated on simulated raw observations in Section 4 whereas conclusions and perspectives for further developments are given in the last section.

2. Image formation for MUSE observations

2.1. Forward model

The forward model formally lists the set of atmospherical and optical disturbances affecting the acquired astronomical scene T (also called the ground truth) and leading to its sampled, noisy and disturbed (*i.e.* observed) version Y on instrument's sensors. Its writing in the form of an observation equation then requires an accurate knowledge of MUSE design and specifications [11] as well as all physical phenomena playing a part in the acquisition process.

The observed ground truth T, a continuous function of space (x, y) and wavelength λ whose both spatial and spectral resolutions are infinite, lies in the model space (also corresponding to the reconstruction and fusion space) whereas the ith observation Y^i lies in the sensor space. The following list summarizes the image formation process from model to sensor space:

1. T is firstly spatially convolved by the FSF (Field Spread Function—spatial PSF) including both atmosphere and instrument effects and then spectrally convolved by the LSF (Line Spread Function—spectral PSF) only induced by MUSE optics. LSF and FSF width depends both on spatial (x, y) and spectral positions λ. Moreover, unlike the FSF, the LSF remains constant between each acquisition session. In the end, one obtains a continuous blurred version of T ($\star$ denotes the convolution operator):
$$T'(u) = \left(T \star h_u^i\right)(u) \quad \text{where } u = (x, y, \lambda) \quad \text{and} \quad h_u^i = \mathrm{FSF}_u^i \times \mathrm{LSF}_u \tag{1}$$
2. T' is sampled on the 3D grid u_p^i which encodes both spatial and spectral sampling geometries and maps a sensor position p to a model space location u. The correspondence between p and u is totally known for each observation session i thanks to mapping tables called "pixtables" and mainly depends on telescope pointing and orientation as well as spectrally variable spatial shifts induced by the atmospheric diffraction. This sampling step then yields:
$$I_p^i = T'(u_p^i) = \left(T \star h_{u_p^i}^i\right)(u_p^i) \tag{2}$$
3. an additive signal-dependent noise B_p^i resulting from the acquisition process and supposed to follow a zero-mean Gaussian distribution with its standard deviation σ_p^i (set to $+\infty$ for dead or saturated CCD pixels) leads to:
$$Y_p^i = I_p^i + B_p^i \quad \text{where } B_p^i \sim \mathcal{N}\left(0, \sigma_p^i\right). \tag{3}$$

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

2.2. Bandlimited resampling scheme for data fusion

In the proposed approach, an important assumption is made as we consider that all observed signals (*i.e.* raw observations) are spatially and spectrally bandlimited. Indeed, during the acquisition process, the high frequency content of T has been definitively lost because of convolutions with LSF and FSF acting as low-pass filters in the Fourier domain. There is therefore no point in trying to recover T from the set of bandlimited Y^i and their related observation parameters. The fusion algorithm then aims at recovering a bandlimited version of the ground truth $F = T \star \varphi$ where φ is a 3D separable convolution kernel ($\varphi = \varphi_x\varphi_y\varphi_\lambda$) which totally determines both finite spatial and spectral resolutions of F. The kernel φ models the PSF of an ideal instrument whose resolution remains spatially and spectrally constant while being higher than that of the MUSE instrument.

In the framework of the Nyquist–Shannon sampling theory [8], the sine cardinal function should be considered as an ideal PSF φ but one would face several drawbacks related to the infinite support of this function (*e.g.* slow implementation or Gibbs oscillations in reconstructed signals). Cubic B-spline functions are an elegant alternative to solve these issues while being a good compromise between accuracy and computational complexity [21]. Indeed, B-splines are nearly bandlimiting functions (*i.e.* F is a good approximation of a bandlimited signal [21]) whereas their footprint is small and finite implying a fast implementation when it comes to interpolation [20]. F can then be interpolated at each model space position u by the cubic B-spline basis function [20]:

$$F(u) \simeq \sum_m L_m \varphi(u - m), \quad m \in \mathbb{Z}^3 \tag{4}$$

where L is a set of discrete B-spline knots (also called interpolation coefficients) weighting the B-spline kernel function φ. A discrete version $X = s \star L$ (s being the discrete B-spline kernel such as $s(m) = \varphi(m)$) of F is rather aimed at in our fusion method, as a simple interpolation step allows to restore the continuous signal from its discrete samples. Thus, X corresponds to an ideal observation acquired by the ideal instrument modeled by the constant PSF φ.

A second assumption is made in our model and states that, as observations, FSF and LSF are bandlimited too, *i.e.* they are wider than φ and can also be approximated by a discrete sum of B-spline kernel functions weighted by interpolation coefficients. With such hypothesis, Eq. (2) can be rewritten:

$$I_p^i = \sum_m L_m \alpha_{pm}^i \quad \text{with } \alpha_{pm}^i = h_{u_p^i}^i (u_p^i - m) \tag{5}$$

where α_p^i is a set of *rendering coefficients* for the pixel p, directly calculated from observation parameters. α_p^i quantifies the contribution of each model space point to the sensor pixel p for the observation i (see Fig. 1(a)). Theoretically, as many rendering coefficients as model space points must be computed per pixel p which leads to the huge size of 750 PB for a single MUSE raw observation. In practice, most of these coefficients can be neglected as they quickly decrease to zero but the remaining ones (about 500 per pixel p) still represent about 1.2 TB of data per observation. One can also mention that their number mainly depends both on PSF sizes and sampling grid of the model space (see Section 2.3).

Fig. 1(b) summarizes the image formation model in the form of a directed acyclic graph where dependences between model variables are shown.

2.3. Model space sampling

Because of atmospheric and instrument disturbances, the resolution of a MUSE observation (*i.e.* LSF and FSF widths) varies according to spatial and spectral positions in the model space. In addition, this variation does not remain constant between several consecutive acquisitions due to different atmospherical conditions. The fusion of such heterogeneous observations on a common sampling grid with constant resolution (given by φ) thus becomes problematic and the choice of the pixel size over the fused image is crucial to avoid both useless oversampling or dramatic undersampling for astrometry computation and introduction of aliasing artifacts.

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

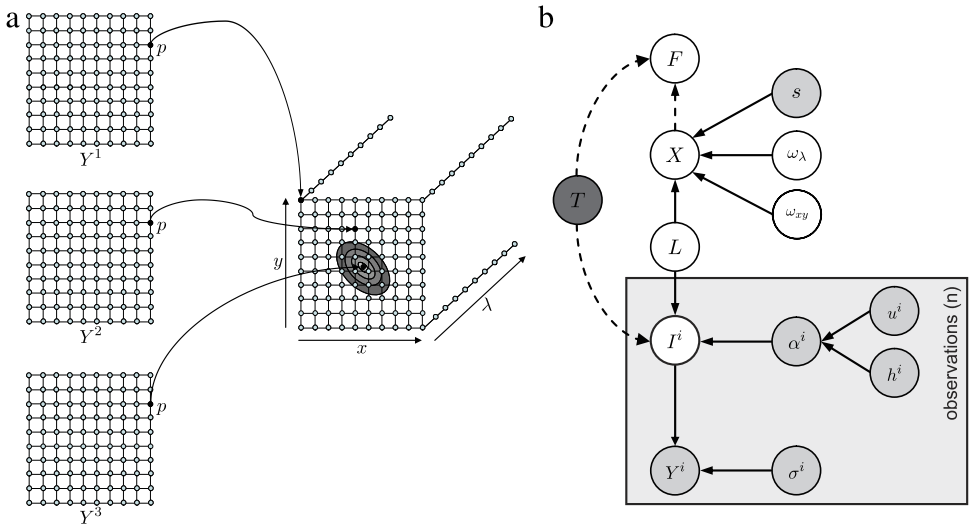

Fig. 1. (a): Mapping between the same pixel p of three MUSE raw observations (acquired under different conditions) and model space elements. Y_p^1, Y_p^2 and Y_p^3 are mapped to different locations in the model space and thus have received different model contributions (*i.e.* different set of rendering coefficients). Surrounded points in the model space correspond to the set of rendering coefficients α_p^3. Most of the model space point contributions are negligible. (b): Directed acyclic graph of the image formation model. Shaded nodes represent fixed variables whereas blank ones stand for random variables to be estimated (ω_{xy} and ω_λ hyperparameters are described in Section 3). The dark gray node is the irretrievable ground truth and dashed lines correspond to generative links.

The resolution of X must then be chosen with respect to that of the observation set, for instance so as to maximize either signal-to-noise ratio or both spatial and spectral resolutions. One can also reparametrize the wavelength axis of the model space so that LSF remain constant as mentioned in [10]. For the spatial domain, a wavelength-dependent spatial resolution for the model space could be implemented but resulting fused images would not be practical to handle and process. In the proposed fusion algorithm, the model space resolution is configurable and first real MUSE observations in 2013 will give additional information on FSF and LSF variations across the field of view and will help to properly set this model parameter.

3. Bayesian hyperspectral fusion

3.1. Bayesian formulation of the inverse problem

For the sake of clarity, matrix notation, where bold and plain letters refer to matrices and vectors, respectively, is adopted for the rest of the paper. The aim of our fusion method is to recover an estimation $\hat{L}$ of underlying interpolation coefficients from the set $\{Y^i, \boldsymbol{\alpha}^i\}_i$ of independent observations and thus compute an estimate of the ideal image $\hat{X} = \boldsymbol{S}\hat{L}$ (where $\boldsymbol{S}$ stands for the B-spline operator). We propose to use the Bayesian inference [5] for the forward model inversion that all comes down to writing and maximizing the following posterior probability deduced from Eqs. (3) and (5):

$$P\left(L|\{Y^i\}_i, \omega\right) \propto \prod_i P\left(Y^i|L\right) \times P\left(L|\omega\right) \tag{6}$$

with:

1. $P\left(Y^i|L\right)$ a Gaussian data-driven (likelihood) term such as, from Eqs. (3) and (5), $Y^i|L \sim \mathcal{N}\left(\boldsymbol{\alpha}^i L, \boldsymbol{P}^{i^{-1}}\right)$ where $\boldsymbol{P}^i$ stands for the diagonal inverse covariance matrix (precision matrix) made of $1/\sigma_p^{i2}$;

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

2. $P(L|\omega)$ a prior probability weighted by the hyperparameter ω and defined either on L or on $X = \boldsymbol{S}L$.

The mandatory choice of the prior model is crucial as it regularizes the final estimate and avoids the noise amplification during the forward model inversion. Due to the large size of handled cubes, we propose to start with a simple and computationally efficient prior modeled by a quadratic Markov random field [12] that imposes a certain spatial (resp. spectral) smoothness of X governed by the regularization parameter ω_{xy} (resp. ω_λ):

$$P\left(L|\omega_{xy},\omega_\lambda\right) \propto \exp -\left[\omega_{xy}\left(\|\boldsymbol{D}_x\boldsymbol{S}L\|^2+\|\boldsymbol{D}_y\boldsymbol{S}L\|^2\right)+\omega_\lambda\left(\|\boldsymbol{D}_\lambda\boldsymbol{S}L\|^2\right)\right] \tag{7}$$

where $\boldsymbol{D}_x$, $\boldsymbol{D}_y$ and $\boldsymbol{D}_\lambda$ stands for horizontal, vertical and spectral first order difference operators respectively.

The regularization parameters ω_{xy} and ω_λ are fixed throughout the inversion algorithm as the use of existing hyperparameter estimation techniques based on marginalized maximum likelihood [2] still remains problematic on such large cubes. In addition, this simple prior might not be relevant for MUSE hyperspectral data and replaced, in the future, by more appropriate (*i.e.* well suited to astronomical objects of different shapes and sizes on a dark and smooth background) but computationally expensive ones such as sparse priors [18]. Finally, thanks to the availability of the precision matrix over $\hat{X}$, denoising algorithms such as wavelet [18] or Pixon [15] based methods can also be implemented on the fused image afterward to remove the extra noise introduced during the maximization of the posterior probability (Eq. (6)). With the simple quadratic prior (Eq. (7)), its optimum is estimated by minimizing its corresponding energy function $U(L) = -\log P\left(L|\{Y^i\}_i,\omega_{xy},\omega_\lambda\right)$ whose gradient and Hessian are given by:

$$\nabla_L U = \left(\nabla_L^2 U\right)L - \sum_i \boldsymbol{\alpha}^{i^T}\boldsymbol{P}^i Y^i \quad \text{and} \quad \nabla_L^2 U = \left(\sum_i \boldsymbol{\alpha}^{i^T}\boldsymbol{P}^i\boldsymbol{\alpha}^i\right) + \underbrace{2\left(\omega_{xy}\boldsymbol{S}\boldsymbol{Q}_{xy}\boldsymbol{S}+\omega_\lambda\boldsymbol{S}\boldsymbol{Q}_\lambda\boldsymbol{S}\right)}_{\mathbf{Q}} \tag{8}$$

with $\boldsymbol{Q}_{xy} = \boldsymbol{D}_x^T\boldsymbol{D}_x + \boldsymbol{D}_y^T\boldsymbol{D}_y$ and $\boldsymbol{Q}_\lambda = \boldsymbol{D}_\lambda^T\boldsymbol{D}_\lambda$.

3.2. Energy minimization

$U(L)$ is a quadratic form whose minimum is determined by solving $\nabla_L U(L) = 0$ with a classical optimization method such as a conjugate gradient approach [16]. However, due to large hyperspectral images and matrix products in Eq. (8), the full evaluation of $\nabla_L U$ cannot be directly performed for each iteration of the optimization algorithm. Thus, a sequential approach is preferred so as to build $\nabla_L U$ while reconstructing and merging the observations one by one. Eq. (8) is then rewritten:

$$\nabla_L U = \left(\boldsymbol{\alpha}^f + \mathbf{Q}\right)L - \Lambda^f \quad \text{with } \Lambda^f = \sum_i \underbrace{\boldsymbol{\alpha}^{i^T}\boldsymbol{P}^i Y^i}_{\Lambda^i} \quad \text{and} \quad \boldsymbol{\alpha}^f = \sum_i \boldsymbol{\alpha}^{i^T}\boldsymbol{P}^i\boldsymbol{\alpha}^i. \tag{9}$$

This simplified expression introduces two new large matrices Λ^f and $\boldsymbol{\alpha}^f$ but whose evaluations become computationally tractable:

1. Λ^i stands for the reconstructed and weighted image associated with Y^i. Indeed, Y_p^i is firstly multiplied by $1/\sigma_p^{i2}$ (product with $\boldsymbol{P}^i$) and then projected in the model space (with respect to both spatial and spectral shifts of Y^i) where its contribution to each pixel of Λ^i is induced by the blur kernel $\boldsymbol{\alpha}^{i^T}$. Λ^i is a blurred but geometrically consistent reconstruction of Y^i and its expression reminds us of the unnormalized drizzling equation [4]. In the end, the reconstructed hyperspectral cubes are fused together which yields Λ^f;
2. Λ^f is affected by a blur operator characterized by $\boldsymbol{\alpha}^f$ and is deconvolved during the gradient descent algorithm as determining the zero of $\nabla_L U$ is equivalent to infer $\hat{L}$ from the simplified equation of observation using the prior defined in Eq. (7):

$$\boldsymbol{\Lambda}^f = \boldsymbol{\alpha}^f L + \mathcal{N}\left(0,\boldsymbol{\alpha}^f\right). \tag{10}$$

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

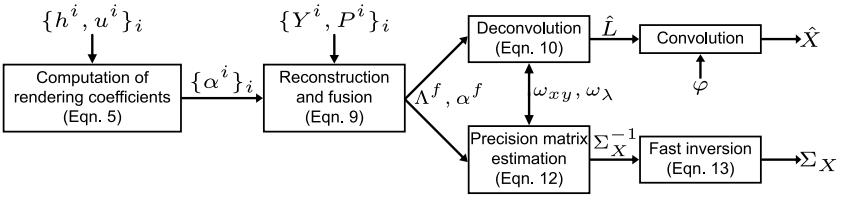

Fig. 2. Global fusion algorithm.

The deconvolution step is performed in a very simple way but more sophisticated methods [19] may be used. Once the optimization algorithm has converged, an estimate $\hat{L}$ is obtained and then leads to $\hat{X}$.

3.3. Uncertainty estimation

The covariance matrix $\mathbf{\Sigma}_X$ encodes the uncertainties (*i.e.* variances and covariances) associated with the fused image $\hat{X}$ and its inverse can be evaluated thanks to a Gaussian approximation of the posterior distribution $X|\{Y^i\}_i, \omega_{xy}, \omega_\lambda$. Indeed, if we assume that the posterior probability $P\left(X|\{Y^i\}_i, \omega_{xy}, \omega_\lambda\right)$ (cf. Eq. (6) with $L = \boldsymbol{S}^{-1}X$) locally follows a Gaussian distribution around its optimum $\hat{X}$, one can write the energy $U = -\log P\left(X|\{Y^i\}_i, \omega_{xy}, \omega_\lambda\right)$:

$$U(X) \simeq \frac{1}{2}\left(X - \hat{X}\right)^T \mathbf{\Sigma}_X^{-1}\left(X - \hat{X}\right) + \text{const.} \tag{11}$$

where $\mathbf{\Sigma}_X^{-1}$ stands for the precision matrix of X. $\mathbf{\Sigma}_X^{-1}$ is then computed as follows:

$$\mathbf{\Sigma}_X^{-1} = \nabla_X^2 U(X) = \boldsymbol{S}^{-1}\boldsymbol{\alpha}^f\boldsymbol{S}^{-1} + 2\left(\omega_{xy}\boldsymbol{Q}_{xy} + \omega_\lambda\boldsymbol{Q}_\lambda\right). \tag{12}$$

$\mathbf{\Sigma}_X^{-1}$ is thus linked to the observation parameter set through $\boldsymbol{\alpha}^f$ from which it is quickly computed as the inverse B-spline operator $\boldsymbol{S}^{-1}$ benefits from a fast implementation [21]. This precision matrix can be used as an input for further Bayesian processing algorithms applied on $\hat{X}$ (several examples are given in [9]) but its direct inversion yielding $\mathbf{\Sigma}_X$ remains intractable due to its large size. We rather propose to invert a local precision matrix $\mathbf{\Sigma}_{X_m}^{-1}$ restricted to a fused pixel m and its neighborhood by solving the following quadratic equation with a classical conjugate gradient method:

$$\arg\min_{\mathbf{\Sigma}_{X_m}} \left\| \mathbf{\Sigma}_{X_m}^{-1}\mathbf{\Sigma}_{X_m} - \mathbb{I} \right\|^2 \tag{13}$$

where $\mathbb{I}$ is the identity matrix and $\mathbf{\Sigma}_{X_m}$ an estimate of the local covariance matrix restricted to m. This fast method, less than 1 s per pixel m, allows to estimate covariances on the fly.

The whole fusion pipeline, from rendering coefficient computations to uncertainty estimation is summarized in Fig. 2.

4. Preliminary results

At this time, neither real MUSE raw observations nor accurate astronomical simulations (currently in development at the astronomical observatory of Lyon) are yet available. To validate the fusion pipeline, we then developed a simple simulation tool to generate synthetic raw observations with respect to the image formation model described in Section 2.1. The acquisition process of a continuous parametric astronomical scene T composed of simple objects modeled by Dirac (stars) and Gaussian (galaxies) functions is reproduced for varying observation parameters: LSF and FSF widths linearly increasing along spectral axis but remaining spatially constant, constant spatial and spectral shifts, noise and sampling grids. In addition, the ideal image X is obtained by convolving T by the B-spline function φ and thus compared with the estimated $\hat{X}$ after fusion.

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

Table 1
Acquisition parameters for each observation. LSF and FSF values correspond to FWHM in pixels, λ_0 (resp. λ_1) stands for the wavelength at the beginning (resp. end) of the model space. Shifts are given in pixels and mean signal-to-noise ratio for stars and galaxies in dB.

Y	FSF (λ_0, λ_1)	LSF (λ_0, λ_1)	Shifts (x, y, λ)	SNR (stars, galaxies)
1	(1.4, 1.96)	(1.7, 1.8)	(0, 0, 0)	(19, 13)
2	(1.6, 2.24)	(1.4, 1.46)	(0.8, 0.6, 1.2)	(18.5, 12.5)
3	(1.4, 1.96)	(1.8, 1.9)	(1.2, 0.5, 0.5)	(20, 12)
4	(1.7, 2.38)	(1.7, 1.8)	(0.2, 1.1, 0.8)	(18, 12)

We simulated 4 raw observations from an underlying scene composed of 2 stars and 2 elliptical galaxies with different acquisition parameters (see Table 1). Noise variances, sampling steps and pixel sizes remain constant for all observations and each raw acquisition is composed of 4 CCD images of 256×32 pixels each. Then, each observation has been sequentially reconstructed and fused in the model space whose size is $32 \times 32 \times 32$ with the same pixel size as observations. The computing time for the whole fusion pipeline from rendering coefficient computation to covariance estimation is about 20 min on a desktop computer (Intel Core 2, Quad Processor 2.50 GHz, 4 GB of RAM).

Fig. 3 presents fusion results centered on a star (point source) and an elliptical galaxy (diffuse object) for our Bayesian inversion method and classical one-step reconstruction algorithms using trilinear and B-spline interpolations corresponding to methods currently in use in astronomy for the reconstruction of hyperspectral data. These results are compared with the ideal image X as well as a (x, y, λ) version of Y^1. Indeed, as there is no shift between spectra, each observation can be simply re-arranged without any reconstruction step in a data cube for display and comparison purposes. Drizzling method cannot be used because it tends to increase the blur level in reconstructed images (as for Λ^f).

Spectra and objects get more blurred with increasing λ as shown on object thumbnails (band 4 and 29) and spectra for Y^1 because of LSF and FSF width variations. Objects reconstructed with interpolation methods still exhibit blur compared to the Bayesian fusion result, especially for high band numbers where spatial and spectral resolutions in the observation set are the lowest. Indeed, the deconvolution step implemented in our approach takes into account all blur information encoded in α^f. For point sources (top part of Fig. 3), highest spatial and spectral frequency content lost during the acquisition process cannot be totally recovered with the Bayesian method due to the simple prior we defined, as shown on thumbnails for band 29 and on cross-sections. Diffuse and smooth objects like galaxies (bottom part of Fig. 3) are less concerned by this issue but their lower brightness lead to a lower signal-to-noise ratio (see Table 1) and then to the introduction of noise on the Bayesian fusion result (see galaxy thumbnails, cross-sections and spectra of Fig. 3). Indeed, unlike interpolation methods which combine observations and introduce some blur, the inversion technique proposed in this paper performs a de-blurring and inevitably leads to a slight noise amplification, which is not totally controlled by the too simple regularization method.

The uncertainties $\mathbf{\Sigma}_X$ associated with the fusion result are presented in the Fig. 4 for the spectrum located at the position (16, 16) in the fused image. Covariances are closely related to the set of observation parameters (see Eq. (12)) and therefore remain spatially constant while being spectrally variable as it is the case for both FSF and LSF in our simulation: the covariance increase with λ thus only reflects that of LSF and FSF widths as the noise level is fixed. Variances can be considered as a good estimation of the fusion quality (*i.e.* the wider the PSF is, the higher the variance is) whereas covariances represent dependences between model coefficients. Moreover, the spatial dependence is higher than the spectral one because the FSF width increases faster than the LSF one in the simulated data set (see Table 1).

5. Conclusion and perspectives

In this paper, we proposed a new hyperspectral fusion method that aims at building a summarized version of a heterogeneous and large set of MUSE raw observations of the same field of view, each acquired under varying conditions. The accurate knowledge of the MUSE instrument brought by

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

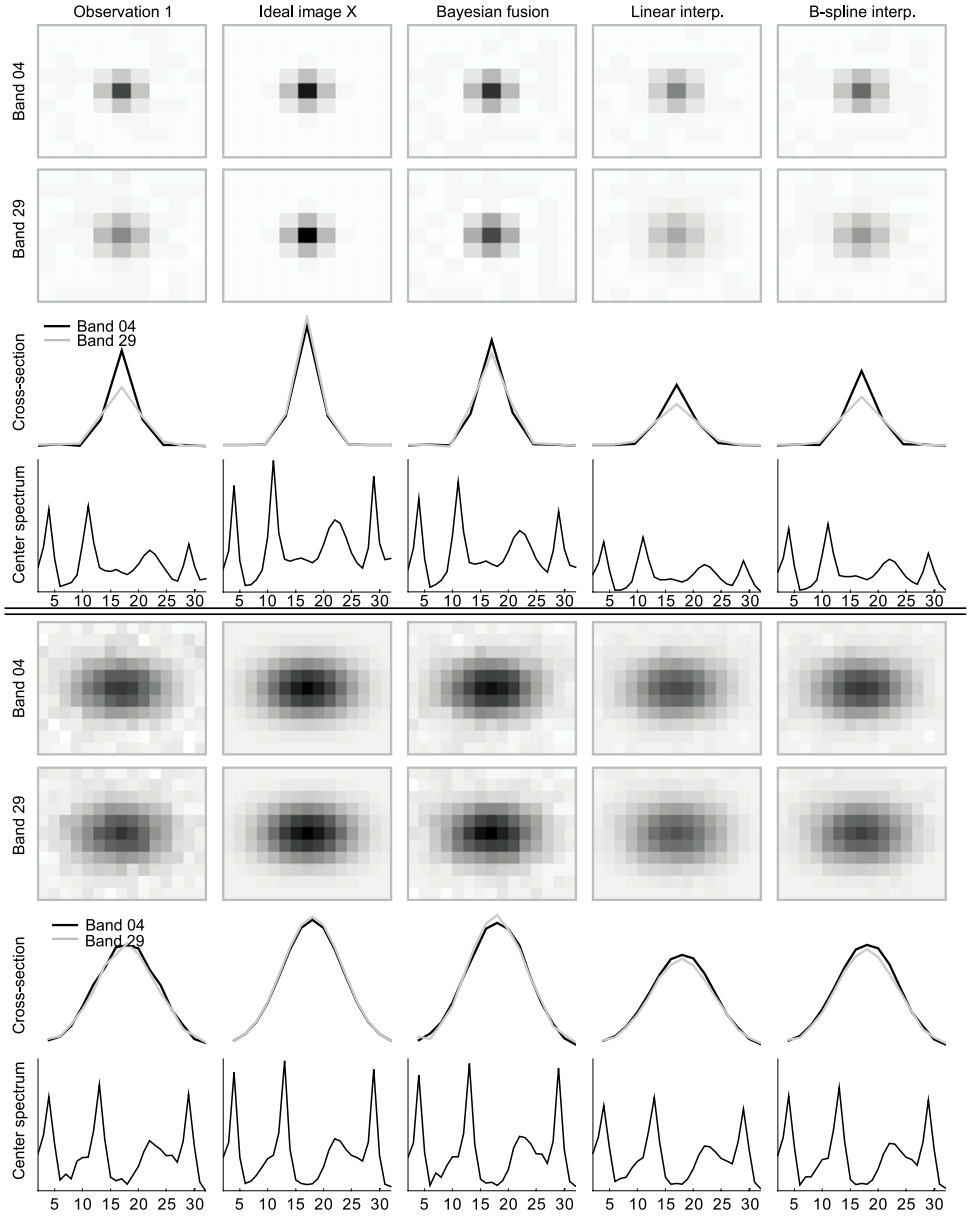

Fig. 3. Results from simulated data ($32 \times 32 \times 32$ pixels) centered on a star (top part) and a galaxy (bottom part) for: (x, y, λ) version of Y^1 (raw data has been re-arranged to form a 32×32 pixels frame for each wavelength), ideal image X, Bayesian fused image $\hat{X}$, classical trilinear and B-spline interpolation results.

the astronomical community allows to precisely write the complex image formation process in the form of an observation equation describing all physical and optical phenomena (different for each observation) involved throughout the acquisition process and encoded in rendering coefficients. Then, the inversion of this forward model with respect to raw observations and their associated parameters is performed within a rigorous framework, guaranteed by the Bayesian inference, and yields an estimate of a data cube whose resolution is limited by a spatially and spectrally constant PSF

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

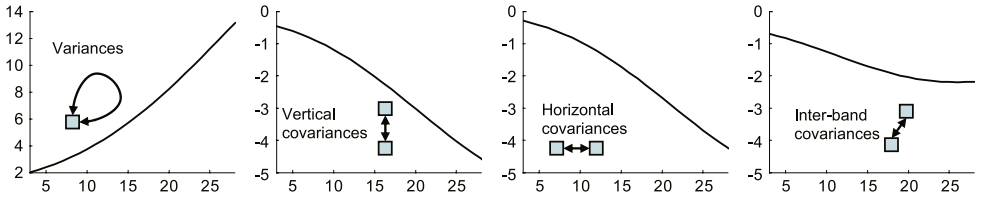

Fig. 4. Variances and covariances related to the Bayesian fusion result along the wavelength axis for the location (16, 16).

modeled by a *B*-spline kernel. This Bayesian approach also gives access to uncertainties related to the inferred resulting image in the form of a precision matrix whose local inversions lead to variances and covariances for each fused pixel. These algorithm outputs can also be used as inputs for a new fusion whenever new raw observations are available afterward.

Even without uncertainties, the fusion product is expected to outperform the standard DRS output, as the Bayesian approach does not only remap sensor-space data onto the model space, but it also inverts the data acquisition process while taking into account all the available measurements and stacking them in a more optimal statistical way.

This innovative approach has been validated on $32 \times 32 \times 32$ pixels simulations but is also intended to be applied on large hyperspectral cubes: up to 100 GB for 80 MUSE raw observations to fuse. It then has been devised so as to optimize both memory usage and file access thanks to a sequential reconstruction algorithm but its extension to large scale data in few years still requires additional optimizations, especially on the management and the computation of $\boldsymbol{\alpha}^i$ and $\boldsymbol{\alpha}^f$ matrices whose sizes quickly increase with that of the data. For instance, to avoid their total computation, we propose to pre-sample a high number of blur kernels (Gaussian ones for instance) on different shifted lattices so as to build a kernel library. Each local blur kernel (*i.e.* row of $\boldsymbol{\alpha}^f$ or $\boldsymbol{\alpha}^i$) may thus be approximated by a combination of these pre-computed functions. Such a method is currently under investigation but its final implementation requires a very accurate knowledge of MUSE acquisition conditions that will obviously be available in 2013 with first real raw observations. We also plan to use code parallelization and graphical processing unit programming to speed up the total processing time (I/O operations, computations and storage) that is constrained by the total acquisition time (up to 80 h) in order to avoid inconvenient delays. Processing and visualization tools should also be delivered to the astronomical community to facilitate the joint analysis of fused images and their uncertainties. For instance, the visualization of covariances together with hyperspectral images in the form of videos is under development [14] and will help astronomers to focus their analysis on low variance regions in images.

During an acquisition session, cosmic rays impact several neighboring CCD pixels at random locations and are affected neither by atmosphere nor by instrument PSF. Both surface and intensity of such impacts depend on the angle of incidence between incoming cosmic rays and CCD sensors and their number is proportional to the exposure time. Their rejection is a crucial task to propose a fusion method robust to outliers (*i.e.* cosmic rays) but the straightforward detection on the 80 MUSE raw observations in the sensor space is not conceivable because of varying observation parameters. The presented sequential approach offers the opportunity to detect cosmic rays in the model space as a Gaussian time series of $\tilde{\Lambda}^i$, the normalized versions of Λ^i (see Eq. (9)), is available for each model location. In that case, the cosmic ray detection is closely related to the robust estimation topic [7] because it all comes down to estimating the mean of the time series by neglecting data points considered as cosmic rays. Methods based on a simple intensity thresholding according to the median absolute deviation [1] and on the robust Bayesian estimation of time series parameters [3] are currently investigated. In the end, detected cosmic rays would be flagged and ignored in the sequential computation of Λ^f and $\boldsymbol{\alpha}^f$ to yield an estimate $\hat{L}$ exempt from most suspicious contributions. Finally, $\hat{L}$ may be used to generate mean raw observations $\bar{I}^i$ predicted by the model ($\bar{I}^i = \boldsymbol{\alpha}^i \hat{L}$) and thus directly allows the rejection of cosmic rays on CCD images by comparing $\bar{I}^i$ to its associated observation Y^i.

In the developed approach, the reconstruction and the fusion of a set of variable resolution observations into a constant resolution image requires an efficient deconvolution step with a well-

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007

designed regularization term. We thus propose a very simple quadratic prior weighted by two fixed hyperparameters (for both spatial and spectral sides) which proves to be unsuitable for high blur and noise levels. Besides improving the used regularization as described in Section 3.1, the reconstruction and the fusion of the observation set into an image having an appropriate resolution for each wavelength is also under investigation. This interesting approach will lead to pyramid-shaped fused cubes as the spatial resolution tends to decrease with increasing λ. It will greatly ease the storage of large data cubes in term of data quantity but will require dedicated tools for proper spectrum extraction or frame comparison.

Finally, the fusion technique can also apply to any kind of astronomical hyperspectral images since raw images at the sensor level as well as acquisition parameters are available.

Acknowledgments

This work was partially funded by the French Research Agency (ANR) as part of the DAHLIA project (*grant* #ANR-08-BLAN-0253). Additional information on DAHLIA and MUSE projects can be found on their respective websites: http://dahlia.oca.eu and http://muse.univ-lyon1.fr.

References

[1] L. Davies, U. Gather, The identification of multiple outliers, Journal of the American Statistical Association 88 (1993) 782–792.
[2] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B. Statistical Methodology 39 (1) (1977) 1–38.
[3] R. Fransens, C. Strecha, L. Van Gool, Robust estimation in the presence of spatially coherent outliers, in: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, CVPRW'06, IEEE Computer Society, 2006.
[4] A.S. Fruchter, R.N. Hook, Drizzle: a method for the linear reconstruction of undersampled images, Publications of the Astronomical Society of the Pacific 114 (2001) 144–152.
[5] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, second ed., Chapman & Hall, 2003.
[6] D. Groom, Cosmic rays and other nonsense in astronomical CCD imagers, Experimental Astronomy 14 (2002) 45–55.
[7] P.J. Huber, E.M. Ronchetti, Robust Statistics, 2nd ed., Wiley-Blackwell, 2009.
[8] A.K. Jain, Fundamentals of Digital Image Processing, Prentice Hall, 1999.
[9] A. Jalobeanu, J.A. Gutiérrez, E. Slezak, Multi-source data fusion and super-resolution from astronomical images, Statistical Methodology 5 (2008) 361–372.
[10] A. Jalobeanu, M. Petremand, C. Collet, Bayesian fusion of hyperspectral astronomical images, in: Proc. of 30th Workshop on Bayesian Inference and Maximum Entropy methods, Chamonix, France, 2010.
[11] F. Laurent, F. Henault, E. Renault, R. Bacon, J. Dubois, Design of an integral field unit for MUSE, and results from prototyping, Publications of the Astronomical Society of the Pacific 118 (2006) 1564–1573.
[12] S.Z. Li, Markov Random Field Modeling in Image Analysis, 3rd ed., Springer-Verlag, 2009.
[13] A. Mohammad-Djafari, et al., Inverse Problems in Vision and 3D Tomography, John Wiley and Sons, 2010.
[14] M. Petremand, M. Louys, C. Collet, V. Mazet, A. Jalobeanu, F. Salzenstein, New Bayesian fusion scheme and visualization tool for astronomical hyperspectral data cubes, in: Astronomical Data Analysis VI, Monastir, Tunisia, 2010.
[15] R.K. Piña, R.C. Puetter, Bayesian image reconstruction: the Pixon and optimal image modeling, Publications of the Astronomical Society of the Pacific 105 (1993) 630–637.
[16] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, 2nd ed., Cambridge University Press, 1993.
[17] R.J. Renka, Multivariate interpolation of large sets of scattered data, ACM Transactions on Mathematical Software 14 (1988) 139–148.
[18] J.-L. Starck, F. Murtagh, J.M. Fadili, Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity, Cambridge University Press, 2010.
[19] J.-L. Starck, E. Pantin, F. Murtagh, Deconvolution in astronomy: a review, Publications of the Astronomical Society of the Pacific 114 (2002) 1051–1069.
[20] P. Thévenaz, T. Blu, M. Unser, Interpolation revisited, IEEE Transactions on Medical Imaging 19 (2000) 739–758.
[21] M. Unser, Splines: a perfect fit for signal and image processing, IEEE Signal Processing Magazine 16 (1999) 22–38.

Please cite this article in press as: M. Petremand, et al., Optimal Bayesian fusion of large hyperspectral astronomical observations, Statistical Methodology (2011), doi:10.1016/j.stamet.2011.04.007