

PDRs4All

XXI. *Haute Couture*: Spectral stitching of JWST MIRI-IFU cubes with matrix completion

Amélie Canin¹, Cédric Févotte², Nicolas Dobigeon², Dries Van De Putte³, Takashi Onaka⁴, and Olivier Berné¹

¹ Institut de Recherche en Astrophysique et Planétologie (IRAP), CNRS, CNES, Université de Toulouse, France

² Institut de Recherche en Informatique de Toulouse (IRIT), CNRS, Toulouse INP, Université de Toulouse, France

³ Department of Physics & Astronomy, The University of Western Ontario, Canada

⁴ Department of Astronomy, Graduate School of Science, The University of Tokyo, Japan

ABSTRACT

The Mid-Infrared Instrument (MIRI) is the imager and spectrograph covering wavelengths from 4.9 to 27.9 μm on board the *James Webb* Space Telescope (JWST). The MIRI Medium-Resolution Spectrometer (MRS) consists of four integral field units (IFU) and two wheels with three rotation settings, producing four spectral channels and three spectral sub-channels. The 12 resulting spectral data cubes have different fields of view, spatial coverages, and spectral resolutions. The wavelength range of each cube partially overlaps with the neighboring bands, and the overlap regions typically show flux mismatches that have to be corrected by spectral stitching methods. The aim of stitching methods is to produce a single data cube incorporating the data of the individual sub-channels, which requires matching the spatial resolution and the flux discrepancies. We present *Haute Couture*, a novel stitching algorithm that uses nonnegative matrix factorization to perform a matrix completion, where the available MRS data cubes are treated as 12 sub-matrices of a larger incomplete matrix. Prior to matrix completion, we also introduced a novel preprocessing method to homogenize the global intensities of the 12 cubes. Our preprocessing consists of jointly optimizing a set of global scale parameters that maximize the fit between the cubes where spectral overlap occurs. We applied our novel stitching method to JWST data obtained as part of the PDRs4All observing program of the Orion Bar and produced a uniform cube reconstructed with the best spatial resolution over the full range of wavelengths.

1. Introduction

The *James Webb* Space Telescope (JWST; Gardner et al. 2006) is an infrared telescope launched in 2021. Four instruments are on board: the Mid-Infrared Instrument (MIRI; Rieke et al. 2015; Wright et al. 2023), the Near Infrared Camera (NIRCam), the Near Infrared Spectrograph (NIRSpec), and the Near Infrared Imager and Slitless Spectrograph (NIRISS). MIRI provides images and spectroscopic data through different observing modes covering 4.9 to 27.9 μm . The Medium-Resolution Spectroscopy (MRS) mode of MIRI (Wells et al. 2015; Argyriou et al. 2023) uses four integral field units (IFUs), with each one covering a portion of the wavelength range and referred to as a channel. Each of these channels is subdivided into three sub-channels (short, medium, and long). Consecutive channels and bands share some information thanks to spectral overlaps. Along the channels, the field of view increases, while the spatial resolution decreases. The science calibration pipeline provided by the Space Telescope Science Institute for the processing of JWST data produces a total of 12 data cubes with different spatial coverages and resolutions. Another challenge raised by the MIRI data acquisition process results from intensity gaps that affect the spectral measurements between adjacent cubes. These intensity gaps, visible from typical spectra reproduced in Fig. 1, may be due to calibration mismatches. To make MIRI data easier to use, a convenient data product would be a singular data cube free of spectral intensity gaps spanning the full spectral range and covering a common field of view, which preserves the highest spatial resolution attainable in each band.

The process of assembling individual spectral data cubes is often referred to as stitching. Because of the aforementioned challenges resulting from the unconventional acquisition implemented by MIRI, stitching raises questions regarding (i) how to assemble intensity gap-free spectra from the 12 data cubes and (ii) how to deal with the distinct spatial sampling (which varies with $1/\lambda$). A commonly employed stitching method consists of re-projecting all the spectral data onto the poorest resolution spatial grid. For instance, this method was applied on data from the Infrared Spectrograph (Houck et al. 2004) taken during the *Spitzer* mission (Werner et al. 2004). This was done by first convolving all the individual cubes with the point spread function (PSF) of the largest wavelength (corresponding to the lowest resolution) in order to reach the same spatial resolution over the full spectral range (see, e.g., Berné et al. 2007; Sandstrom et al. 2010). This procedure, referred to as “coarse stitching” hereafter, is illustrated in Fig. 2.a when considering two data cubes. Unfortunately, this strategy leads to a regrettable loss of valuable spatial information. Alternatively, one can reconstruct the cube over the full wavelength range using a defined pixel size, as is done by the Drizzle algorithm provided by STScI for MIRI-MRS (Law et al. 2023). However, in this case, the cube is reconstructed at the spatial resolution of JWST for the given wavelength, that is, not at the highest JWST angular resolution.

To overcome the limitations inherent to the coarse stitching strategy, this paper presents a novel method named *Haute Couture* that enables stitching while preserving spatial resolution. The proposed method performs the reconstruction of a full data cube at the best spatial resolution over the full spectral range, as illustrated in Fig. 2.b. We first show that the data recorded in the

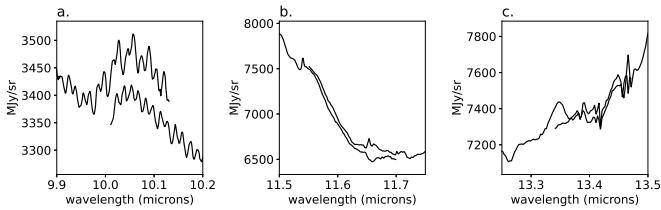


Fig. 1. Examples of intensity gaps observed in overlapping areas of adjacent cubes, obtained from MIRI-MRS data collected within the PDRs4All program (see Sect. 3.1 for more details). The spectra displayed correspond to the average spectra over the entire field of view for two overlapping sub-channels. Panel a: Channel 2-medium and 2-long. Panel b: Channel 2-long and 3-short. Panel c: Channel 3-short and 3-medium. Fringes can be seen in Panel a, corresponding to a well-known phenomenon already documented by Argyriou et al. (2020).

Table 1. Name and wavelength range of the MIRI-MRS channels and sub-channels.

	Sub-channel	(c, b)	λ -range (μm)
channel 1	short	(1, s)	4.90 - 5.74
	medium	(1, m)	5.66 - 6.63
	long	(1, l)	6.53 - 7.65
channel 2	short	(2, s)	7.51 - 8.77
	medium	(2, m)	8.67 - 10.13
	long	(2, l)	10.01 - 11.70
channel 3	short	(3, s)	11.55 - 13.47
	medium	(3, m)	13.34 - 15.57
	long	(3, l)	15.41 - 17.98
channel 4	short	(4, s)	17.70 - 20.95
	medium	(4, m)	20.69 - 24.48
	long	(4, l)	24.40 - 27.90

12 sub-channels can be rearranged as 12 sub-matrices of a larger matrix with missing values. Interestingly, recovering these missing values amounts to reconstructing a data cube at the finest spatial resolution over the full spectral range. *Haute Couture* exploits the overlapping spectral information between adjacent channels and bands to frame the stitching task as a matrix completion problem, which can then be solved by nonnegative matrix factorization (NMF). To avoid the issues caused by intensity gaps between consecutive channels and bands, prior to matrix completion, *Haute Couture* applies a novel preprocessing to homogenize the global intensities along the 12 cubes. This preprocessing consists of jointly optimizing a set of global scaling parameters to maximize the fit between the cubes where spectral overlap occurs. It is worth noting that this preprocessing is an independent procedure that can be combined with any stitching method.

Following the presentation of the methodological compounds of *Haute Couture* in Sect. 2, we present MIRI-MRS data observed as part of the PDRs4All program on the Orion Bar in Sect. 3 and report the application of our stitching method in Sect. 4. In particular, we demonstrate the ability of *Haute Couture* to reconstruct a uniform cube with the best spatial resolution over the full range of wavelengths.

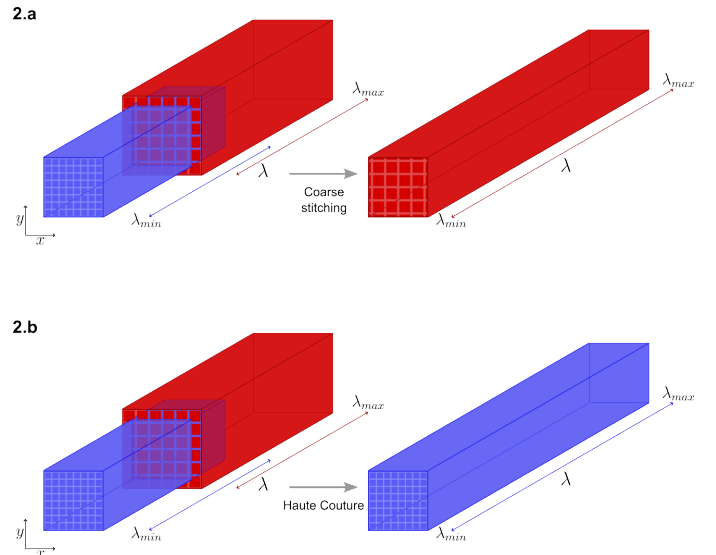


Fig. 2. Illustration of the stitching problem for two cubes. On the left, one can observe that the shortest wavelength cube in blue has a better spatial resolution (as illustrated by the finer grid) and a smaller field of view than the longest wavelength cube in red. The space-wavelength supports of the cubes partially overlap. The coarse stitching procedure shown in Fig. 1a sacrifices spatial resolution, while *Haute Couture* shown in Fig. 1b enables the reconstruction of a cube with the best spatial resolution over the full range of wavelengths.

2. Haute Couture

2.1. Spectral stitching as a matrix completion problem

As introduced in Sect. 1, MIRI-MRS acquires spectra in four channels indexed by c ($c = 1, \dots, 4$). Each channel is divided into three sub-channels, referred to as short, medium, and long and shortened as s, m, and l, respectively. The spectral ranges of each sub-channel are reported in Table 1, as obtained from the STScI Jdocs.¹ Each sub-channel produces a space-wavelength cube as schematized in Fig. 2, where two of these cubes are depicted in blue and red. In the remainder of the paper, the spatial organization of the data is ignored, and we unfold each cube with respect to the spatial dimension (i.e., pixels). Thus the data associated with each cube can be represented as a matrix indexed by pixels and wavelengths. Formally, the matrix that gathers the data collected in sub-channel $b \in \{s, m, l\}$ of channel c ($c = 1, \dots, 4$) is denoted as $\mathbf{X}^{c,b} \in \mathbb{R}_+^{\Lambda^{c,b} \times P^{c,b}}$, where its rows index wavelengths while its columns index pixels. Specifically, the matrix $\mathbf{X}^{c,b}$ associated with the sub-channel (c, b) gathers spectra composed of $\Lambda^{c,b}$ wavelengths and acquired over $P^{c,b}$ spatial pixels.

The resulting 12 sub-matrices can be arranged according to a suitable manner to form a larger matrix $\mathbf{X} \in \mathbb{R}_+^{\Lambda \times P}$, as represented in Fig. 3, where $P = \sum_{c=1}^4 \sum_{b \in \{s, m, l\}} P^{c,b}$ and Λ is the total number of wavelengths over the spectral range covered by MIRI-MRS. It is worth noting that the spectral range of adjacent sub-matrices overlap for a few wavelengths, i.e., $\Lambda \neq \sum_{c=1}^4 \sum_{b \in \{s, m, l\}} \Lambda^{c,b}$. This results in an almost block diagonal structure of the matrix \mathbf{X} . In this figure, the white parts of the matrix correspond to unavailable (i.e., unobserved) data, specified by the symbol \emptyset . Because

¹ <https://jwst-docs.stsci.edu/jwst-mid-infrared-instrument/miri-observing-modes/miri-medium-resolution-spectroscopy>

of these missing coefficients, the matrix \mathbf{X} is said to be incomplete.

The key rationale of the proposed stitching approach rests on the following insight: Recovering all the missing coefficients of this matrix \mathbf{X} would provide a comprehensive but redundant spatial-spectral description of the scene observed by MIRI-MRS. In particular, recovering the values of the sub-matrix \mathbf{S} corresponding to the blue shaded area in Fig. 3 would produce a data cube at the spatial resolution of the channel 1-long (i.e., $c = 1$ and $b = 1$) over the full spectral extent offered by MIRI-MRS, which is the main objective of this work. In other words, the spectral stitching task can be formulated as a matrix completion problem, i.e., recovery of the missing data in the matrix \mathbf{X} . In this work, we target the channel 1-long to define the spatial resolutions of the stitched data because this leads to the best trade-off between high spatial resolution and a high signal-to-noise ratio. However, this arbitrary choice can be lifted since the proposed method will complete the full matrix \mathbf{X} . In particular, we conducted some experiments while choosing the channel 1-short and 1-medium spatial resolutions. We obtained similar yet noisier structures than when targeting channel 1-long.

The following paragraphs detail the proposed two-step procedure to stitch the MIRI-MRS data while accommodating for intensity gaps between adjacent data cubes. This procedure consists of first rescaling the individual sub-matrices $\mathbf{X}^{c,b}$ homogeneously to correct the intensity gaps (see Sect. 2.2) and then filling the matrix \mathbf{X} composed of the resulting intensity-corrected sub-matrices (see Sect. 2.3).

2.2. Intensity gaps correction

To mitigate the intensity gaps highlighted in Fig. 1, we introduced a novel preprocessing to homogenize the global intensities of the 12 cubes. This preprocessing consists of jointly adjusting a set of global scale parameters to maximize the fit between the parts of the cubes where spectral overlaps occur. With $\alpha^{c,b}$, we denote the scale parameter to be applied to the entries of the matrix $\mathbf{X}^{c,b}$. As an example, referring to Fig. 3, the scale parameter, say $\alpha^{2,1}$, is computed by maximizing the fit between, on the one hand, the first rows of $\mathbf{X}^{2,1}$ and the last rows of $\mathbf{X}^{2,m}$ and, on the other hand, the last rows of $\mathbf{X}^{2,1}$ and the first rows of $\mathbf{X}^{3,s}$. Because the matrices $\mathbf{X}^{2,m}$ and $\mathbf{X}^{3,s}$ themselves should be rescaled by the unknown scale parameters $\alpha^{2,m}$ and $\alpha^{3,s}$, the optimal values of the whole set of scale parameters are inter-dependent, and they have to be adjusted jointly. Thankfully, when the fit is measured by the squared Euclidean distance between the portions of shared spectra, this optimization problem has a closed-form solution.

In practice at least one sub-channel must be chosen as an intensity reference to make the problem well-posed (otherwise, a trivial but inappropriate solution would be setting $\alpha^{c,b} = 0$ for all channels and sub-channels). In the experimental results conducted in Sect. 3, we fixed $\alpha^{1,s} = \alpha^{4,1} = 1$, and maximized the fit with respect to the ten remaining scale parameters. This amounts to trusting the calibration of the data observed in the lowest and highest ends of the wavelength range and updating the scales of the cubes in between. We note, however, that any other choice is possible. Technical details and the expressions of the resulting optimal scale parameters are reported in Appendix A.

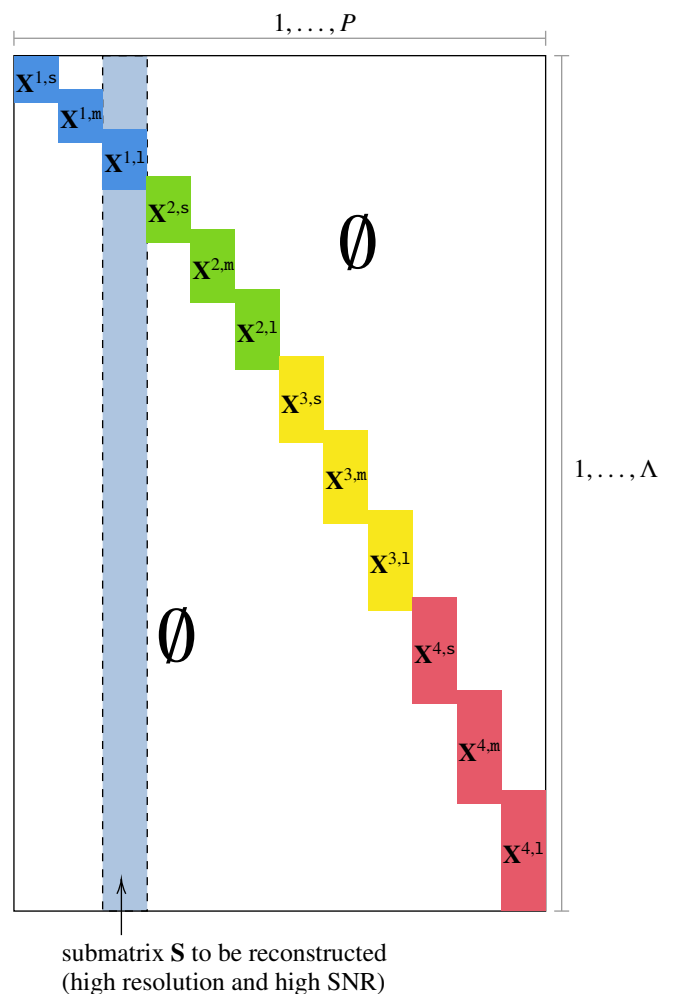


Fig. 3. Twelve unfolded cubes provided by the 12 sub-channels organized as sub-matrices of a larger matrix \mathbf{X} with an almost block-diagonal structure. The wavelength range of the consecutive blocks overlaps for a few rows, corresponding to spectral bands observed by consecutive sub-channels. Blue blocks represent channel 1, green ones channel 2, yellow ones channel 3, and red ones channel 4. White and light blue parts correspond to unobserved coefficients. The light blue area corresponds to the area we are specifically interested in reconstructing.

2.3. Matrix completion

As explained in Sect. 2.1, recovering the missing coefficients in the full matrix \mathbf{X} displayed in Fig. 3 can be interpreted as a matrix completion problem. Matrix completion aims at predicting missing values based on given observed values and the assumption of a latent matrix structure (see, e.g., Chi et al. 2019). In our case, and as in many other settings, it makes sense to assume that \mathbf{X} has a "low-rank" structure, an assumption that has already underpinned several techniques of infrared spectroscopic data processing such as source separation (Rapacioli et al. 2005; Berné et al. 2007) or data fusion (Berné et al. 2010; Guilloteau et al. 2020a,b). It consists of imposing that the columns of \mathbf{X} can be explained by linear combinations of elementary spectra embedded in noise. To further exploit the nonnegativity of the observed spectra, we propose performing this matrix completion by resorting to the technique of NMF (Lee & Seung 1999; Smaragdis et al. 2014). In other words, the missing entries are approximated by $x_{ij} \approx [\mathbf{WH}]_{ij}$, where \mathbf{W} and \mathbf{H} are nonnegative matrices of size $\Lambda \times K$ and $K \times P$ estimated from the low-rank ap-

proximation of the observed entries. More precisely, we denote by O the set of row and column indices of the observed values in \mathbf{X} , i.e., the non-white part of \mathbf{X} in Fig. 3 (or equivalently the supports of the sub-matrices $\mathbf{X}^{c,b}$). The two matrices \mathbf{W} and \mathbf{H} are estimated by solving the following optimization problem:

$$\min_{\mathbf{W}, \mathbf{H} \geq 0} \sum_{(i,j) \in O} d(x_{ij} | [\mathbf{WH}]_{ij}), \quad (1)$$

where $d(u|v)$ denotes a discrepancy measure between the non-negative numbers u and v . In this study we leverage the work of Févotte & Idier (2011), who proposed easy-to-implement and efficient multiplicative updates for estimating \mathbf{W} and \mathbf{H} when $d(\cdot)$ is chosen as a so-called β -divergence. The β -divergence is a continuous family of measures of fit governed by a single shape parameter $\beta \in \mathbb{R}$ that takes well-known divergences as special cases, namely the generalized Kullback-Leibler and Itakura-Saito divergences ($\beta = 1$, $\beta = 0$, respectively) and the square Euclidean distance ($\beta = 2$). Given a couple of matrices $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$ that solve the minimization problem (1), the missing coefficients of \mathbf{X} can be reconstructed as $\hat{x}_{ij} = [\hat{\mathbf{W}}\hat{\mathbf{H}}]_{ij}$ for $(i, j) \notin O$.

Like most NMF techniques, the algorithm tailored by Févotte & Idier (2011) is a descent algorithm that relies on alternating updates of \mathbf{W} and \mathbf{H} . Because of the bilinearity induced by the product \mathbf{WH} , the objective function to minimize in Eq. (1) is nonconvex, and the algorithm is likely to produce local solutions that depend on the chosen starting point. As such, the initialization of the descent algorithm is an issue. We therefore investigate several initialization strategies in Sect. 3.2.

3. Application to MIRI-MRS data of the PDRs4All ERS program

3.1. Data and preprocessing

Data. The observations used in this work are part of the Early Release Science (ERS) program PDRs4All (Berné et al. 2022) and were obtained in January 2023. This program observed the Orion Bar with MIRI-MRS in a 1×9 pointing mosaic in the four channels. The mosaic is depicted in white in Fig. 4. Data were acquired with 47 groups per integration, one integration, and four dithers using the FASTR1 readout. The data were presented by Chown et al. (2024) and Van De Putte et al. (2024). The experiments reported in this paper used a single mosaic tile containing the protoplanetary disk d203-506 (highlighted in red in Fig. 4), which was also previously studied by Berné et al. (2023) and Zannese et al. (2024). Figure 5 presents the MIRI-MRS original data cubes with the initial field of view in each channel.

Data reduction pipeline. To illustrate the relevance of the proposed stitching method on a set of data representative of the current quality of JWST data products, the MIRI MRS data were reprocessed starting from the “uncal” files downloaded from MAST. For this purpose we used the scripts provided by the PDRs4All data reduction team and available from a public repository² in combination with the JWST pipeline version 1.17.1 and CRDS context version “jwst_1322.pmap.” The standard pipeline stages were applied first with the default settings for stage 1 (detector1) and the following settings for stages 2 and 3. In stage 2, the image-to-image background subtraction method was applied (“bkg_subtract”) using the “rate” files

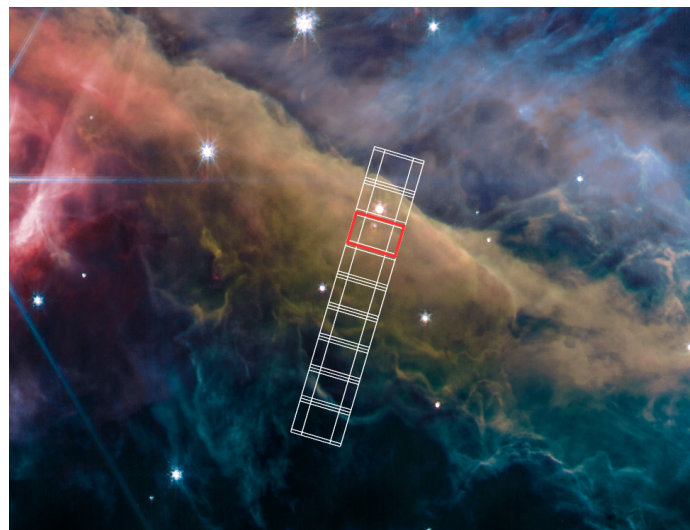


Fig. 4. Colorized NIRCам image of the Orion Bar of the PDRs4All ERS program (Berné et al. 2024). Filters F140M and F210M are in blue; F277W, F300M, F323N, and F335M are in green; F405N is in orange; and F444W, F480M, and F470N are in red. The pattern in white is the MIRI-MRS mosaic footprint. The red box corresponds to the pointing used in this article.

obtained by applying stage 1 to the dedicated background observations for this program. The “residual_fringe” step was enabled, which reduced the residuals of fringes due to the reflections within the detector (Argyriou et al. 2020). In stage 3, we enabled the “outlier_detection” step and disabled the “cube_build” step so that the final products were the fully processed 2D IFU images (“crf” files). We then ran the “cube_build” step, which is based on the Drizzle algorithm (Law et al. 2023), and it led to a common spatial grid across all wavelengths for the data to be stitched. This choice was motivated by two main reasons: (i) Focusing on the exact intersection of fields of view with distinct spatial pixel resolutions would have led to a significant loss of information. (ii) Because of the lower spatial resolution, the data at longer wavelengths carry less information (in terms of numbers of pixels), which would have biased the results. This resampling could have been performed after data reduction during the preprocessing step (see below) by a naive spatial interpolation. Instead, this resampling was performed during the data reduction to benefit from the functionalities offered by the pipeline, which performs this interpolation more efficiently, using both spatial and spectral information.

Therefore, the “cube_build” step was performed with custom settings starting from the stage 3 “crf” files. The full specification of these settings was as follows. The position angle was set to “cube_pa=250.4,” which aligned with the orientation of the 9×1 mosaic. The pixel scale and field of view were set to those of channel 1 using “scale_xy=0.13,” “nspax_x=33,” and “nspax_y=39,” and the center coordinates were set to “ra_center=83.834782,” “dec_center=-5.418207.” For channel 1, the resulting cube had nearly identical properties to the one built with the default coordinate system. For the longer wavelength channels, the pixel scale of these custom cubes substantially oversampled the instrumental resolution; this was the expected type of input for the method presented in this work.

Preprocessing. The *Haute Couture* stitching method proposed in Sect. 2 assumes that the spectral grids between two adjacent cubes are the same for the shared wavelength range, which is

² <https://github.com/PDRs4All/PDRs4All>

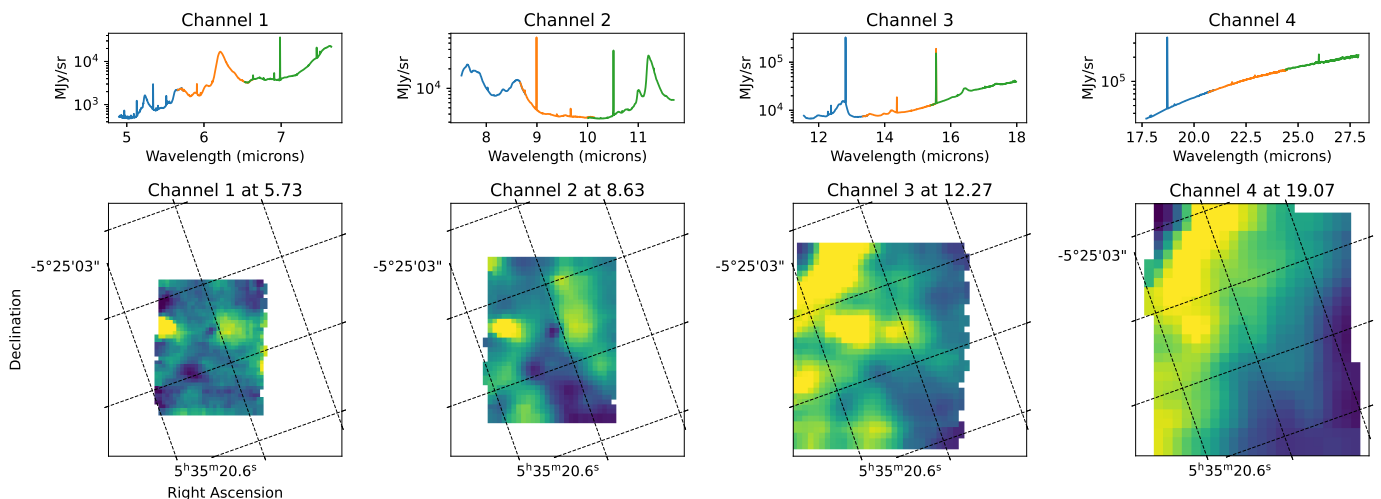


Fig. 5. First row: Average spectra over the entire field of view of each channel (shown in the second row). A different color (blue, orange, green) is used for each sub-channel (short, medium, long). Second row: MIRI-MRS images at different wavelengths (the images are aligned). It can be observed that the field of view increases with the channel but the spatial resolution deteriorates.

not the case for the data provided by the reduction pipeline. As such, the data were spectrally resampled on a common wavelength grid with a consistent step using a linear 1D interpolation.

Additionally, the data contained saturated frames, spikes, and so-called bad pixels (i.e., outliers) that needed to be cleaned. Instead of merely discarding the saturated frames, they were replaced by corrected frames resulting from a spectral linear interpolation to avoid gaps with no data for some spectral bands. This step was shown to be of major importance. If the data cubes to be stitched contained one saturated frame, the stitched result might be significantly distorted. For instance, it may contain unexpected spatial patterns and significant spectral bias. Regarding the remaining spikes and bad pixels in each frame, they were automatically detected and corrected. A significant number of those anomalies exhibited structures with high intensity. To identify them, we implemented a robust anomaly detector following the strategy proposed by Anderson & Gordon (2011). In each frame $\mathbf{X}_{i,:}^{c,b}$ associated with a given channel c and sub-channel b the median pixel value $\bar{x}_i^{c,b} = \text{median}(\mathbf{X}_{i,:}^{c,b})$ was first computed as a robust counterpart of the mean pixel values. Each pixel value significantly far from this robust mean was then identified as an anomaly. More precisely, a pixel value $x_{ij}^{c,b}$ such that $|x_{ij}^{c,b} - \bar{x}_i^{c,b}| > t \times \text{MAD}(\mathbf{X}_{i,:}^{c,b})$ was flagged as an anomaly, where $\text{MAD}(\mathbf{X}_{i,:}^{c,b}) = \text{median}(x_{i,j}^{c,b} - \bar{x}_i^{c,b})$ computes the median absolute deviation and t is a cutoff parameter that needs to be adjusted for each channel and sub-channel. The pixel values detected as anomalies were finally replaced by the average of their four spatially nearest neighbors.

Finally, we created a mask containing only the most informative pixels. A pixel value was flagged as not carrying any information if its value was NaN. For each cube, the mask corresponds to the spatial pixels that have more than 99% spectral information (i.e., that are not NaN). In the mask, the star at the top of our field of view was removed because it is saturated in most of the wavelength range and the spectra are not exploitable.

3.2. Choice of the parameters and initialization

As discussed in Sect. 2, *Haute Couture* requires a few hyperparameters to be set by the end user. A first hyperparameter is the value of β , i.e., the shape parameter of the divergence used as a measure of the fit in NMF. As thoroughly discussed by Févotte & Idier (2011), the β -divergence is a log-likelihood in disguise, and choosing β is similar to making a noise assumption. The range of practical values is generally $\beta \in [0, 2]$, with $\beta = 0$ corresponding to multiplicative Gamma noise and $\beta = 2$ corresponding to additive Gaussian noise. Choosing $\beta = 1$ (generalized Kullback-Leibler divergence), corresponding to a Poisson-like continuous distribution, has been shown to offer an excellent trade-off in many applications. In particular, it was advocated for in the work by Berné et al. (2007) for source separation of similar mid-IR data using NMF.

A second hyperparameter to set is K , i.e., the rank of the factorization (the common dimension of \mathbf{W} and \mathbf{H}). A suitable value needs to capture the latent structure in the data while preventing overfitting (i.e., modeling the noise rather than the inherent phenomena that explain the data). We followed a trial and error approach by assessing the quality of the reconstruction with a set of candidate values for K , namely $K \in [2, 4, 6, 10, 15]$. We considered only channel 2 and 3 in order to save computing time. For every couple of estimated matrices \mathbf{W} and \mathbf{H} , we then calculated the correlation between the reconstruction and the MIRI-MRS spectra (spectra after preprocessing in order to compare) for each sub-channel and the average of the sub-channel. For all K , the average correlation was 0.935 ± 0.0006 . The algorithm was not very sensitive to the value of K in our example, and thus, we chose $K = 6$ since it maximized the correlation.

In addition to choosing the hyperparameters β and K , *Haute Couture* requires choosing an initialization strategy for \mathbf{W} and \mathbf{H} in the NMF algorithm. We empirically assessed the performance of three methods: (i) random initialization, (ii) using the outputs of the K-means clustering algorithm, and (iii) using the maximum angle source separation (MASS) technique proposed by Boulais et al. (2021). As for the adjusting of the hyperparameters β and K , we conducted an exhaustive set of experiments to assess the performance of each of these three methods in reaching the best fit, with $\beta = 1$, $K = 6$, again using chan-

nel 2 and 3 to save computing time. Our experiments revealed that K-means and MASS return similar results and a much better performance than random initialization. We favored MASS in the experiments, however, as it was initially designed for infrared astronomical hyperspectral cubes.

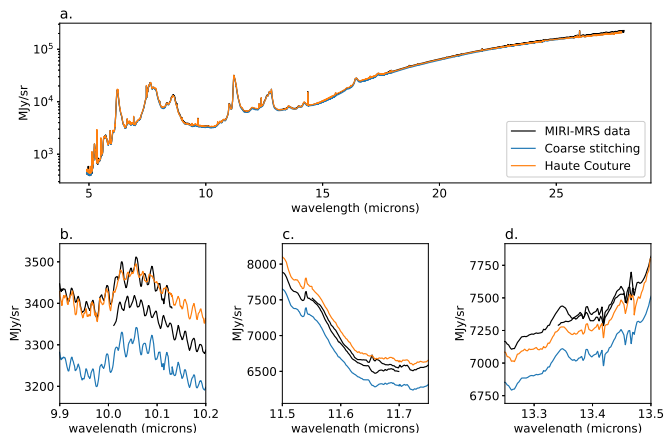


Fig. 6. Top: Average spectrum of the MIRI-MRS data after removing the saturated frames, spikes, and bad pixels (black curve); average of the stitched data recovered by *Haute Couture* (orange curve); and average of the result of the coarse stitching procedure (blue curve). Bottom: Zoom-ins on the spectral ranges where overlap occurs between contiguous sub-channels (as already depicted in Fig. 1).

4. Results

4.1. Stitching results

Figure 6 (top) displays the average spectrum of the MIRI-MRS data over the entire field of view after removing the saturated frames, the spikes, and the bad pixels, as explained in Sect. 3.1. It depicts the average spectrum of the stitched data \hat{S} recovered by *Haute Couture* (corresponding to the values in channel 1-long of the sub-matrix highlighted in the blue shaded area of Fig. 3). This result is compared to the product given by the coarse stitching procedure explained in Appendix B. The figure also displays three zoom-ins of the spectral ranges where overlap occur between contiguous sub-channels (bottom). By design, the *Haute Couture* stitched data average spectrum is continuous, i.e., the spectral gaps exhibited in the original MIRI-MRS data have been removed. It is also worth noting that the fit between the original and *Haute Couture* reconstructed data is rather tight over the full spectral range.

Figure 7 displays the preprocessed MIRI-MRS data before stitching and compares the final products resulting from the coarse stitching procedure and from the proposed *Haute Couture* method. This figure also depicts the *Haute Couture* product after undergoing a convolution with a Gaussian filter whose standard deviation $\sigma = \frac{1.22\lambda}{2\sqrt{2\ln 2}D}$ has been specifically designed to mimic the effect of the MIRI PSF, where λ is the considered wavelength and D is the mirror diameter. The resulting convolved product can then also be fairly compared to the original MIRI-MRS data as a validation step.

In the first row, corresponding to data and results at $5.61\mu\text{m}$ (channel 1), the *Haute Couture* image appears as significantly denoised and exhibits the same spatial resolution as the original data. This was expected since, as motivated in Sect. 2.1, the *Haute Couture* stitching procedure targets the spatial resolutions

of channel 1-long (blue shaded area in Fig. 3). Moreover, the observed denoising effect can be attributed to the low-rank factor model underlying *Haute Couture*. By implicitly representing the spectral data into a K -dimensional subspace, the proposed stitching preserves the essential physical information while effectively filtering out most of the noise. The second row, associated with results at $9.29\mu\text{m}$ (channel 2), shows a lightly denoised and more contrasted *Haute Couture* image compared to the MIRI-MRS image. At $12.27\mu\text{m}$ (channel 3), the *Haute Couture* image is sharper and more contrasted than the MIRI-MRS data. Interestingly, the spatial structure of the d203-506 protoplanetary disk, which is known to be present in the considered astrophysical scene Berné et al. (2024), is well reconstructed. The MIRI-MRS image at $20.42\mu\text{m}$ (channel 4) does not contain any particular spatial structure and the d203-506 protoplanetary disk is no longer visible. In the *Haute Couture* image, the spatial structures are well recovered.

Conversely, in all channels, the product resulting from the coarse stitching procedure is affected by a worse contrast and a clear underestimation of the spectral flux density already highlighted in Fig. 6. It is also worth noting that the loss in spatial resolution is significant. This was expected since the coarse stitching procedure targets the poorest resolution offered by the data cubes to be stitched.

The most straightforward way to validate an algorithm is to test it against ground truth data, as is often possible with Earth observation data. However, in the case of astronomical data, obtaining ground truth is impossible. A widely used alternative consists of validating the algorithm using simulated data. This strategy has been followed for instance by Guilloteau et al. (2020b) to validate data fusion algorithms by generating simulated NIRCcam and NIRSpect data. Simulating MIRI-MRS data, however, is more challenging (Hadj-Youcef et al. 2017). The Pandeia tool (Pontoppidan et al. 2016) can produce simulated MIRI-MRS data, but it only allows simulation of simple scenes. The MRISim tool (Klaassen et al. 2021) allows for more complex scenes. However, to our knowledge, the public version of this software has not been updated since the launch of JWST, and therefore the tool may not provide data that are directly comparable to real observations. In this context, as a sanity check, we compare in Fig. 7 the results of degraded *Haute Couture* data after undergoing the Gaussian smoothing (Col. 4) and those of the original MIRI-MRS data (Col. 1). It can be seen that there is a good qualitative agreement in terms of the spatial textures between images of Cols. 1 and 4, especially in channels 1 and 2. At longer wavelengths, the images differ notably in the right part of the scene. This can be explained by the masking procedure implemented during the preprocessing, which removes pixels associated with the star located in this area (see Sect. 3.1).

4.2. Error propagation

The MIRI-MRS cubes provided by the pipeline as detailed in Sect. 3.1 and stitched in Sect. 4.1 are provided with so-called error cubes (specified by the extension ERR). More precisely, each data cube $\mathbf{X}^{c,b}$ composing the matrix \mathbf{X} depicted in Fig. 3 is granted with an error matrix denoted $\Delta\mathbf{X}^{c,b}$. To quantify the relative amount of error contained in the original MIRI-MRS data to be stitched, we first evaluated the signal-to-error ratio (SER) defined as

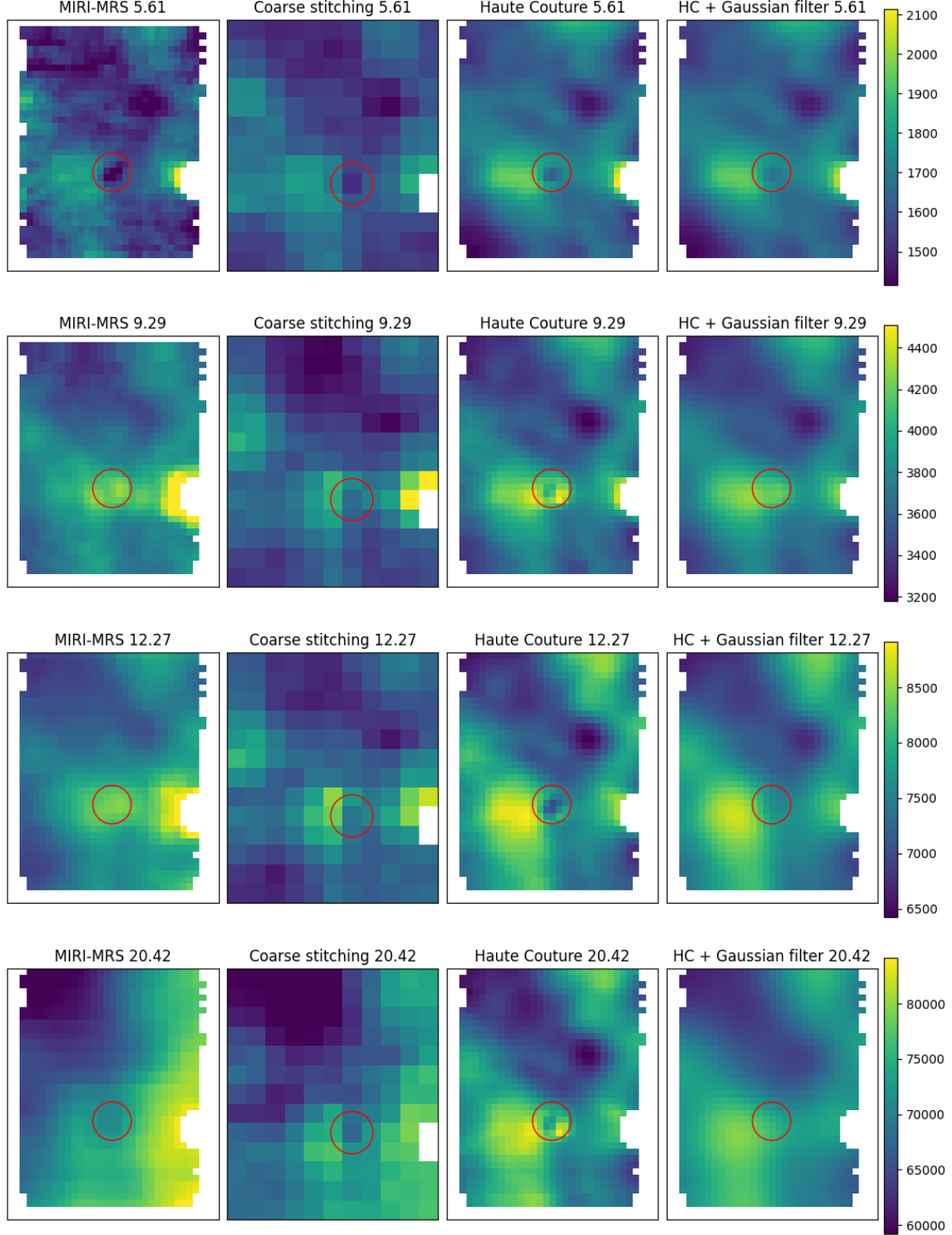


Fig. 7. Comparison between the MIRI-MRS data obtained using the JWST pipeline `cube_build` (Col. 1), the result of the coarse stitching procedure (Col. 2), the stitched data produced by *Haute Couture* (Col. 3), and this final product convolved with a Gaussian filter (Col. 4) in each channel at a given wavelength. The same color bar (in MJy/sr) is used in each row, and the data in Cols. 1, 3, and 4 share the same mask. The red circles correspond to the emplacement of the d203-506 protoplanetary disk. The loss of spatial resolution is also clearly visible when performing a coarse stitching.

$$\text{SER} = 10 \log \frac{\sum_{c,b} \|\mathbf{X}^{c,b}\|_2^2}{\sum_{c,b} \|\Delta\mathbf{X}^{c,b}\|_2^2} = 53.09\text{dB}. \quad (2)$$

We then assessed the propagation of this error when performing stitching. To do so, we introduced the set of data matrices

$\mathbf{X}_+^{c,b} = \mathbf{X}^{c,b} + \Delta\mathbf{X}^{c,b}$ and $\mathbf{X}_-^{c,b} = \mathbf{X}^{c,b} - \Delta\mathbf{X}^{c,b}$ corresponding to the original data positively and negatively perturbed by the error term in each channel c and sub-channel b . These sub-matrices were rearranged to form the incomplete matrices \mathbf{X}_+ and \mathbf{X}_- as perturbed counterparts of the original incomplete matrix \mathbf{X} introduced in Sect. 2 (see also Fig. 3). The completion procedure

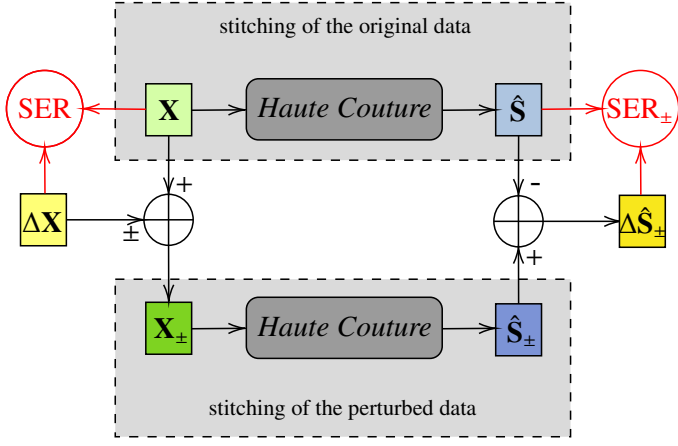


Fig. 8. Quantification of the error propagation. The amount of error, SER, before stitching was evaluated from the original data contained in \mathbf{X} (light green box) and the error matrix $\Delta\mathbf{X}$ (light yellow box). The original data \mathbf{X} were processed by *Haute Couture* to produce the stitched spectra $\hat{\mathbf{S}}$ (light blue box). The error matrix $\Delta\mathbf{X}$ was combined with the original data \mathbf{X} to produce the perturbed data \mathbf{X}_\pm (dark green box). The perturbed data \mathbf{X}_\pm were processed by *Haute Couture* to produce the stitched spectra $\hat{\mathbf{S}}_\pm$ (dark blue box), yielding the propagated error $\Delta\hat{\mathbf{S}}_\pm$ (dark yellow box). The amount of error SER_\pm after stitching was evaluated from the stitched original spectra $\hat{\mathbf{S}}$ and the propagated errors $\Delta\hat{\mathbf{S}}_\pm$.

detailed in Sect. 2.3 was applied on these two incomplete matrices \mathbf{X}_+ and \mathbf{X}_- , yielding the pair of stitched data $\hat{\mathbf{S}}_+$ and $\hat{\mathbf{S}}_-$ recovered by the proposed method *Haute Couture*. Finally, the propagation of the error was evaluated by quantifying how much the original amount of error translates into the perturbed data after undergoing stitching. Formally, we computed the SER associated with the positively and negatively perturbed data as

$$\text{SER}_+ = 10 \log \frac{\|\hat{\mathbf{S}}\|_2^2}{\|\Delta\hat{\mathbf{S}}_+\|_2^2} \quad \text{and} \quad \text{SER}_- = 10 \log \frac{\|\hat{\mathbf{S}}\|_2^2}{\|\Delta\hat{\mathbf{S}}_-\|_2^2}, \quad (3)$$

where $\Delta\hat{\mathbf{S}}_\pm = \hat{\mathbf{S}}_\pm - \hat{\mathbf{S}}$ measures the discrepancy between the stitched spectra $\hat{\mathbf{S}}$ resulting from the original data and the stitched spectra $\hat{\mathbf{S}}_\pm$ resulting from the perturbed data. $\Delta\hat{\mathbf{S}}_\pm$ is considered as the propagated error. We obtained $\text{SER}_+ = 33.43\text{dB}$ and $\text{SER}_- = 34.25\text{dB}$. The overall procedure is sketched out in Fig. 8. It appears that the SER deteriorates by about 20dB, which may be explained by the fact that the error terms $\Delta\mathbf{X}^{c,b}$ do not likely follow the NMF model underlying the completion procedure detailed in Sect. 2.3. Yet, more importantly, the amount of additional error induced by *Haute Couture* remains of low magnitude with respect to both the original error and the energy of the stitched data.

5. Conclusion

In this paper we have presented a new stitching algorithm named *Haute Couture* that provides near-optimal assembly of MIRI-MRS spectral cubes. *Haute Couture* allows spatial information to be preserved throughout the overall spectral range covered by JWST. By applying the algorithm to real observations, we showed that *Haute Couture* is able to recover spatial structures that are not preserved by a coarse stitching procedure. Our proposed approach relies on the formulation of the stitching task as a matrix completion problem, which is subsequently solved by NMF. In principle, the versatility of this formulation makes the proposed approach applicable to spectral cubes provided

by other spatial instruments, such as NIRSpec (also on board JWST) or by ground-based IFUs.

Data availability

The source code for this method is available through <https://github.com/acanin/haute couture>. The *Haute Couture* cube presented in this paper is available in electronic form at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via <http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/>.

Acknowledgements. This work is partially supported by the Artificial and Natural Intelligence Toulouse Institute (ANITI), funded by the France 2030 program under the grant agreement ANR-23-IACL-0002. This work is funded by the Centre National d'Etudes Spatiales (CNES) through the APR program.

References

- Anderson, R. E. & Gordon, K. D. 2011, *PASP*, 123, 1237
- Argyriou, I., Glasse, A., Law, D. R., et al. 2023, *A&A*, 675, A111
- Argyriou, I., Wells, M., Glasse, A., et al. 2020, *A&A*, 641, A150
- Berné, O., Habart, É., Peeters, E., et al. 2022, *PASP*, 134, 054301
- Berné, O., Habart, E., Peeters, E., et al. 2024, *Science*, 383, 988
- Berné, O., Helens, A., Pilleri, P., & Joblin, C. 2010, in 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, IEEE, 1–4
- Berné, O., Joblin, C., Deville, Y., et al. 2007, *A&A*, 469, 575
- Berné, O., Martin-Drumel, M.-A., Schroetter, I., et al. 2023, *Nature*, 621, 56
- Boulais, A., Berné, O., Faury, G., & Deville, Y. 2021, *A&A*, 647, A105
- Chi, Y., Lu, Y. M., & Chen, Y. 2019, *IEEE Transactions on Signal Processing*, 67, 5239
- Chown, R., Sidhu, A., Peeters, E., et al. 2024, *A&A*, 685, A75
- Févotte, C. & Idier, J. 2011, *Neural Computation*, 23, 2421
- Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, *Space Sci. Rev.*, 123, 485
- Guilloteau, C., Oberlin, T., Berne, O., & Dobigeon, N. 2020a, *IEEE Transactions on Computational Imaging*, 6, 1362
- Guilloteau, C., Oberlin, T., Berné, O., Habart, É., & Dobigeon, N. 2020b, *AJ*, 160, 28
- Hadj-Youcef, M. A., Orioux, F., Fraysse, A., & Abergel, A. 2017, in 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, 503–507
- Houck, J. R., Roellig, T. L., van Cleve, J., et al. 2004, *ApJS*, 154, 18
- Klaassen, P. D., Geers, V. C., Beard, S. M., et al. 2021, *MNRAS*, 500, 2813
- Law, D. R., E. Morrison, J., Argyriou, I., et al. 2023, *AJ*, 166, 45
- Lee, D. D. & Seung, H. S. 1999, *Nature*, 401, 788
- Perrin, M. D., Sivaramakrishnan, A., Lajoie, C.-P., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9143, *Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave*, ed. J. M. Oschmann, Jr., M. Clampin, G. G. Fazio, & H. A. MacEwen, 91433X
- Pontoppidan, K. M., Pickering, T. E., Laidler, V. G., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9910, *Observatory Operations: Strategies, Processes, and Systems VI*, ed. A. B. Peck, R. L. Seaman, & C. R. Benn, 991016
- Rapacioli, M., Joblin, C., Boissel, P., Calvo, F., & Spiegelman, F. 2005, in *ESA Special Publication*, Vol. 577, *ESA Special Publication*, ed. A. Wilson, 409–410
- Rieke, G. H., Wright, G. S., Böker, T., et al. 2015, The mid-infrared instrument for the James Webb Space Telescope, I: Introduction, Tech. Rep. Technical Report JWST-STScI-000001, STScI
- Sandstrom, K. M., Bolatto, A. D., Draine, B. T., Bot, C., & Stanimirović, S. 2010, *ApJ*, 715, 701
- Smaragdīs, P., Févotte, C., Mysore, G. J., Mohammadiha, N., & Hoffman, M. 2014, *IEEE Signal Processing Magazine*, 31, 66
- Van De Putte, D., Meshaka, R., Trahin, B., et al. 2024, *A&A*, 687, A86
- Wells, M., Pel, J.-W., Glasse, A., et al. 2015, *PASP*, 127, 646
- Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, *ApJS*, 154, 1
- Wright, G. S., Rieke, G. H., Glasse, A., et al. 2023, *PASP*, 135, 048003
- Zannese, M., Tabone, B., Habart, E., et al. 2024, *Nature Astronomy*, 8, 577

Appendix A: Joint scale parameter estimation

This appendix details the approach introduced in Sect. 2.2 and implemented to mitigate intensity gaps observed between the sub-channels of MIRI-MRS data cubes (see Fig. 1). To ease the presentation of the method, we introduce the one-to-one mapping between the couple $(c, b) \in \{1, \dots, 4\} \times \{s, m, l\}$ which indexes a given sub-channel b of a given channel c and the integer $i \in \{1, \dots, 12\}$ which indexes the sub-matrix composing \mathbf{X} according to the particular order represented in Fig. 3. Thus in what follows, the sub-matrix $\mathbf{X}^{c,b}$ is simply denoted as \mathbf{X}^i , leading to the alternate representation of the matrix \mathbf{X} depicted in Fig. A.1.

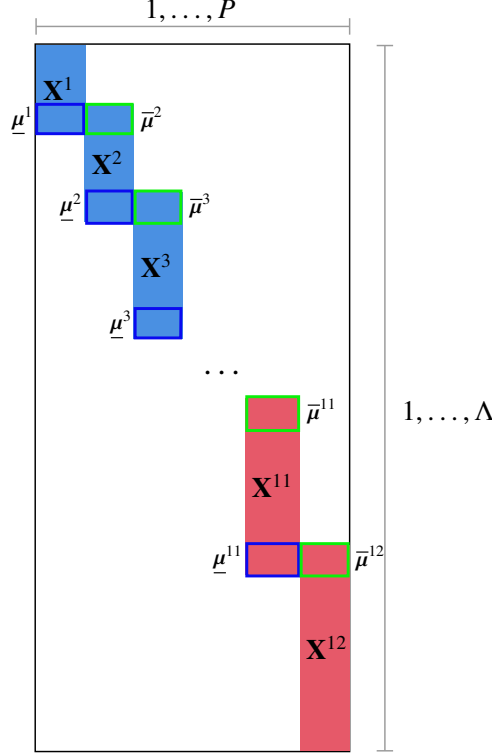


Fig. A.1. Matrix \mathbf{X} as of Fig. 3 with alternative indexing of the sub-matrices.

The intensity gap correction aims at estimating a set of 12 scale parameters $\alpha = [\alpha^1, \dots, \alpha^{12}]^T$ to be applied to the 12 data matrices $\{\mathbf{X}^1, \dots, \mathbf{X}^{12}\}$, respectively. By the very specific nature of the data acquisition, the wavelength range associated with the top rows (resp. bottom) of the matrix \mathbf{X}^i emphasized as green (resp. blue) boxes in Fig. A.1 overlaps with the bottom rows of the matrix \mathbf{X}^{i-1} (resp. top of \mathbf{X}^{i+1}) emphasized as blue (resp. green) boxes in Fig. A.1. It is worth noting that contiguous green and blue boxes frame portions of spectra that correspond to the same spatial locations at the same wavelengths. Thus the proposed approach consists in jointly adjusting the scale parameters by ensuring that the portions of pixel spectra located in the blue boxes match as much as possible the portions of the pixel spectra in the contiguous green boxes. Formally, the spatial average of the top (resp. bottom) rows of \mathbf{X}^i where such overlap occurs is denoted by $\bar{\mu}^i$ (resp. $\underline{\mu}^i$), see Fig. A.1. Maximizing the fit between $\alpha^i \underline{\mu}^i$ and $\alpha^{i+1} \bar{\mu}^{i+1}$, for $i = 1, \dots, 11$, boils down to solving the optimization problem

$$\min_{\alpha} \sum_{i=1}^{11} \left\| \alpha^i \underline{\mu}^i - \alpha^{i+1} \bar{\mu}^{i+1} \right\|_2^2, \quad (\text{A.1})$$

with $\alpha = [\alpha^1, \dots, \alpha^{12}]^T$. The problem as formulated by Eq. (A.1) is ill-posed since one optimal solution is $\alpha^* = [0, \dots, 0]^T$, i.e., setting all the sub-matrices to zero. To remove this trivial, inappropriate solution, one or more sub-channel serve as intensity references, with scale parameters fixed to an arbitrary value. In particular, in this work we assume that channels 1 and 12 have been properly calibrated and set $\alpha^1 = \alpha^{12} = 1$, but the proposed methodology may apply with any other arbitrary choices. Thus, we introduced the sub-vector of scale parameters $\tilde{\alpha} = [\alpha^2, \dots, \alpha^{11}]^T$ and reformulated the problem (A.1) as

$$\min_{\tilde{\alpha}} \mathcal{J}(\tilde{\alpha}) \quad \text{with} \quad \mathcal{J}(\tilde{\alpha}) = \sum_{i=2}^{10} \left\| \alpha^i \underline{\mu}^i - \alpha^{i+1} \bar{\mu}^{i+1} \right\|_2^2. \quad (\text{A.2})$$

The criterion $\mathcal{C}(\tilde{\alpha})$ to be minimized is smooth and convex. Problem (A.2) has a unique solution that obeys

$$\nabla_{\tilde{\alpha}} \mathcal{J}(\tilde{\alpha}) = 0, \quad (\text{A.3})$$

where ∇ denotes the gradient operator. Trivial algebra shows that Eq. (A.3) is equivalent to solving the linear problem

$$\mathbf{M}\tilde{\alpha} = \mathbf{b}, \quad (\text{A.4})$$

where \mathbf{M} and \mathbf{b} are the square matrix and vector of sizes ten given by

$$\mathbf{M} = \begin{pmatrix} 1 & -\frac{\langle \mu^2, \bar{\mu}^3 \rangle}{\|\mu^2\|_2^2 + \|\bar{\mu}^2\|_2^2} & 0 & \dots & 0 \\ -\frac{\langle \mu^2, \bar{\mu}^3 \rangle}{\|\mu^3\|_2^2 + \|\bar{\mu}^3\|_2^2} & 1 & -\frac{\langle \mu^3, \bar{\mu}^4 \rangle}{\|\mu^3\|_2^2 + \|\bar{\mu}^3\|_2^2} & \ddots & \vdots \\ 0 & -\frac{\langle \mu^3, \bar{\mu}^4 \rangle}{\|\mu^4\|_2^2 + \|\bar{\mu}^4\|_2^2} & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{\langle \mu^{10}, \bar{\mu}^{11} \rangle}{\|\mu^{10}\|_2^2 + \|\bar{\mu}^{10}\|_2^2} \\ 0 & \dots & 0 & -\frac{\langle \mu^{10}, \bar{\mu}^{11} \rangle}{\|\mu^{11}\|_2^2 + \|\bar{\mu}^{11}\|_2^2} & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \frac{\langle \mu^1, \bar{\mu}^2 \rangle}{\|\mu^2\|_2^2 + \|\bar{\mu}^2\|_2^2} \\ 0 \\ \vdots \\ 0 \\ \frac{\langle \mu^{11}, \bar{\mu}^{12} \rangle}{\|\mu^{11}\|_2^2 + \|\bar{\mu}^{11}\|_2^2} \end{pmatrix}.$$

The optimal scale parameters are thus given by $\tilde{\alpha}^* = \mathbf{M}^{-1}\mathbf{b}$.

Appendix B: Coarse stitching

In this section we present how we produce the coarse stitching presented as comparison in Sect. 4. We use the MIRI-MRS original data cubes presented in Sect. 3.1 and Fig. 5.

Field of view. The first step is to have a common field of view corresponding to the intersection of each field of view. A exact common spatial grid is impossible due to the pixel size of each channel.

Convolution with PSF. Each channel and sub-channel is convolved with the PSF of the channel 4-long using the library `stpsf` (Perrin et al. 2014). We use the Fourier transform to perform the convolution. Then we calibrate the flux in order to have the same flux than before because there is a side effect when convoluting the data that loses flux. This effect is particularly visible at large wavelength because these are fewer pixels.

Reprojection. To compare the same pixels, we need a common spatial grid for every channel and sub-channel. Hence, those are reprojected in the spatial grid of the channel 4-long using the function `griddata` of the library `SciPy`.

Stitching. The stitching is performed individually for each spatial pixel. The channel 1-short is used as reference for the calibration. For each pixel the average of the fluxes are calculated on the spectral overlap between two successive sub-channel. Then the spectrum that is not the reference is scaled using the ratio of the average.