Reduction and Calibration Pipelines for e-MERLIN and COBRaS

Luke Peck[*]

Department of Physics & Astronomy, University College London, UK
E-mail: lwp@star.ucl.ac.uk

Danielle Fenech

Department of Physics & Astronomy, University College London, UK
E-mail: dmf@star.ucl.ac.uk

As part of the COBRaS e-MERLIN Legacy project, SERPent is a reduction and Radio Frequency Interference (RFI) -mitigation script which has been produced to form part of a bigger automated reduction and calibration pipeline. Written in the Parseltongue scripting language, it accesses interferometric data via the AIPS program. It can be easily added and executed in other pipelines as a module, or run alone. The script flags RFI affected visibilities, antenna dropouts and Lovell phase calibrator stationary scans inherent to the e-MERLIN system.

Here the flagging and computational performances of SERPent are presented with e-MERLIN commissioning datasets for both L and C band observations. The refining of automated reduction and calibration procedures is essential for the e-MERLIN legacy projects, where the vast data sizes ($>$ TB) mean traditional astronomer interactions with the data are unfeasible.

---

[*]Speaker.

1. e-MERLIN and the Cygnus OB2 Radio Survey (COBRaS) Legacy Project

The expanded - Multi Element Radio Linked Network (e-MERLIN) is an array comprising of seven antenna located across the UK, the largest of which is the 76m Lovell telescope at Jo-drell Bank. The upgrade consists of a new optical fibre network, receivers, analogue and digital electronics and correlator providing:

- Observing Bands: 1.3 - 1.8 GHz (L Band), 4 - 8 GHz (C Band), 22 - 24 GHz (K Band)

- Bandwidth: 4 GHz (C band)

- Sensitivity: $\sim 1\ \mu$Jy

- Resolution: 10 to 150 mas.

The Cygnus OB2 Radio Survey (COBRaS) is a UCL-led (P.I. Prof. Raman Prinja) inter-national project, utilising the recently upgraded e-MERLIN. As one of the e-MERLIN legacy projects, it has been awarded approximately 300 hours of observing time, to produce the most detailed radio census of the most massive OB association in the northern hemisphere. This study will investigate a range of inter-related astrophysical themes which have implications to a wide astrophysical audience including:

- Mass loss and evolution of massive stars,

- The formation, dynamics and content of massive OB associations,

- The frequency of massive binaries and the incidence of non-thermal radiation.

The observations are split between multiple L and C Band pointings, which once combined will produce a highly detailed mosaic map of the core region of Cygnus OB2 as figure 1 demon-strates (C Band pointings).
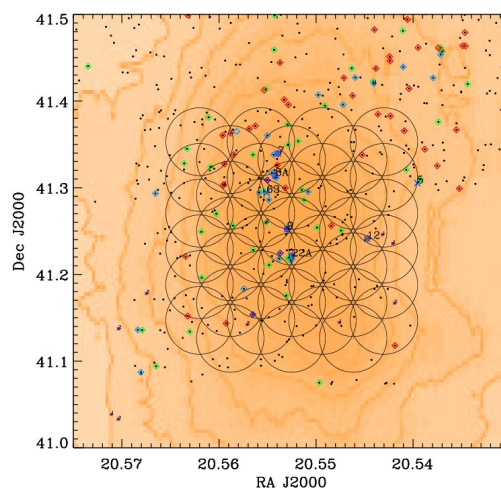

Figure 1: COBRaS Legacy C band Mosaicking map of the core region of Cyg OB2.

2. SERPent

In preparation of full COBRaS Legacy data, a series of software tools has been developed with the goal of fully automating the reduction and calibration procedures for radio interformetric datasets. With the advent of modern arrays (LOFAR, ALMA) and upgrades to existing arrays (e-MERLIN, EVLA), the observation data sizes are reaching 1 - 1000's Terabytes, making traditional astronomer interactions with the data unfeasible.

The **S**cripted **E**-merlin **R**fi-mitigation **P**ypelin**E** for i**NT**erferometry (SERPent) is an automated reduction and RFI-mitigation pipeline written in the popular Parseltongue scripting language. During the commissioning of e-MERLIN, SERPent has evolved from a RFI-mitigation procedure to a full data-editing pipeline. It will run on data from any interformeter loaded within AIPS, however, it has specialised passages for e-MERLIN.

2.1 Lovell - Mk2 Baseline Removal

Two of the seven antennas in e-MERLIN are located at the same site at Jodrell Bank. The proximity of the Lovell and Mk2 antenna results in the sidelobes from each beam to interfere with one another causing confusion in the data. SERPent automatically flags this baseline.

2.2 Lovell Stationary scans

A problem unique to the e-MERLIN array is the Lovell stationary scan. Due to the size of the Lovell telecope and the subsequent slew time, the Lovell telescope only participates in every alternative phase-cal scan, remaining stationary on the target for the other scans. The other antennas in the array are not affected. This results in the visibilities from baselines containing the Lovell telescope to have two different amplitude levels for the phase-cal, which can be seen in figure 2.
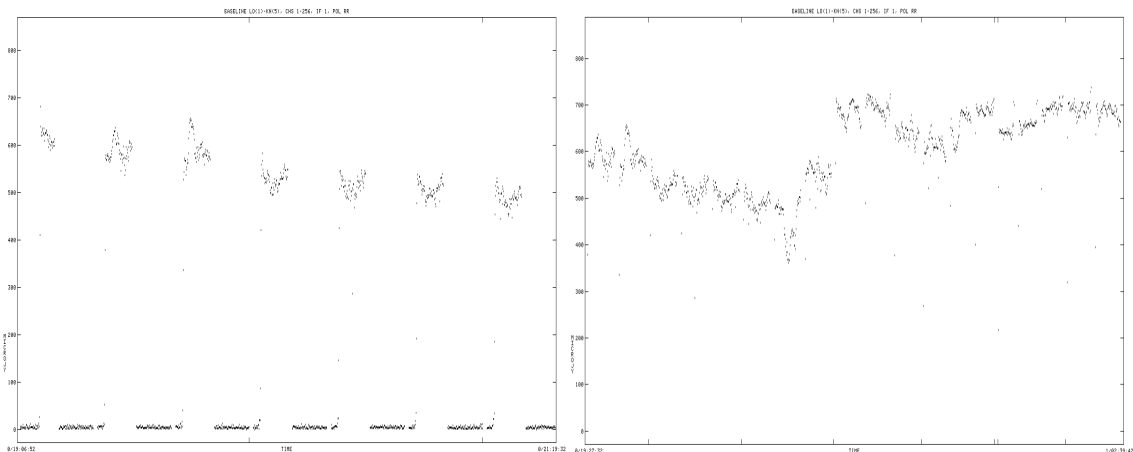

Figure 2: AIPS IBLED task window, displaying the phase calibrator source: 2007+404. The left image is before flagging and the Lovell stationary scans can be seen around 0 Jy level, and the right image is after flagging. Each screen shows a small selection of scans for clarity.

2.3 Zero Level in-scan amplitude flagging

Early COBRaS commissioning data revealed zero level in-scan amplitudes for some visibilities, as a result of a system failure or telescope slew errors. These zero level amplitudes reside

within scans containing good data and thus need their own passage within SERPent to be flagged. Figure 3 shows another before and after image demonstrating how SERPent has flagged the zero level amplitudes within a scan.
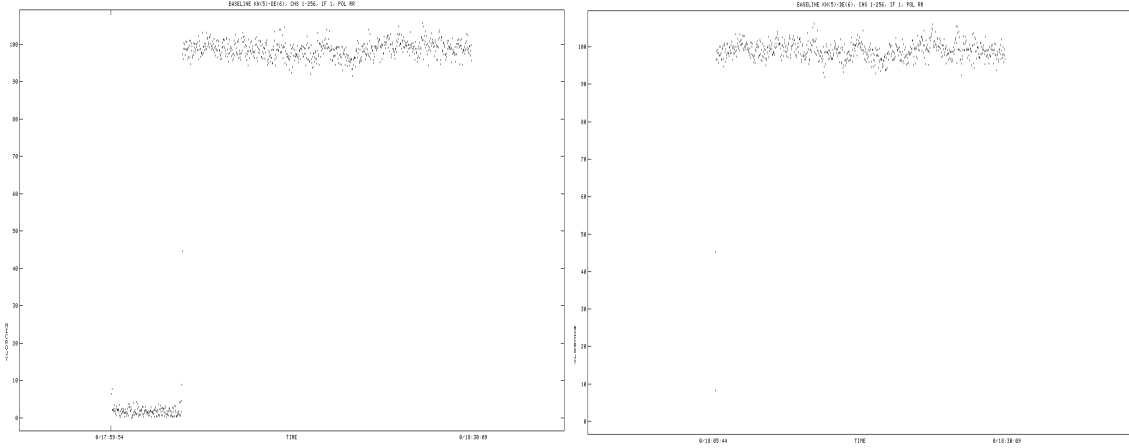

Figure 3: AIPS IBLED task window, displaying the phase-cal source: 2007+404. This before image (left) of a scan shows two visibility levels at 100 mJy and 0 mJy representing good and bad data respectively and the after SERPent flagging image (right) shows this in-scan zero level amplitude part removed.

2.4 RFI-mitigation Performance

The main component of SERPent is the implementation of a variation of the SumThreshold sequence (Offringa et al. 2010 [1]), a post correlation detection method which searches for RFI in the time-frequency domain, i.e. time scan vs channel forming a 2D array. As a form of combinatorial thresholding, a combination of samples are tested against some threshold limit, which decreases as the number of visibilities in the sample increases with successive runs. The number of samples tested, or subset i, increases binarily to a certain size, e.g. i = 1, 2, 4, 8, 16, 32, 64 for performance purposes. If a visibility is above the given threshold, it is flagged and set to the threshold levels in the succeeding runs, as to not bias the subset.

The first threshold level can be calculated by a range of methods and statistics. The variance of a sample is an important component for this threshold and various methods are described and tested by Fridman (2008) [2].

The Median Absolute Deviation (MAD) is used as the variance estimator in SERPent and is defined by equation 2.1, where $median_i(x_i)$ is the median of the original population. Each sample of the population is then modified by the absolute value of the median subtracted from each sample. The median of this new population is taken and multiplied by a constant 1.4286 to make this estimation consistent with that of an expected Gaussian distribution.

$$MAD = 1.4826 \, median_j\{|x_j - median_i(x_i)|\} \tag{2.1}$$

The first threshold level χ_1 is thus determined by an estimate of the mean $\bar{x}$, the variance σ and an aggressiveness parameter β (equation 2.2) (Niamsuwan, Johnson & Ellingson 2005) [3]. Since the median is less sensitive to outliers, it is preferred to the traditional mean in this equation (thus

$\bar{x} = median$) and the MAD to the traditional standard deviation for the variance for similar reasons ($\sigma = MAD$).

$$\chi_1 = \bar{x} + \beta \sigma \tag{2.2}$$

The subsequent threshold levels are determined by equation 2.3 where $\rho = 1.5$, this empirically works well for the SumThreshold method (Offringa et al. 2010 [1]) and defines how 'coarse' the difference in threshold levels will be, and i is the subset value.

$$\chi_i = \frac{\chi_1}{\rho^{log_2 i}} \tag{2.3}$$

Figures 4 and 5 show the results of SERPent's implementation of the SumThreshold method for L and C band e-MERLIN commissioning data.
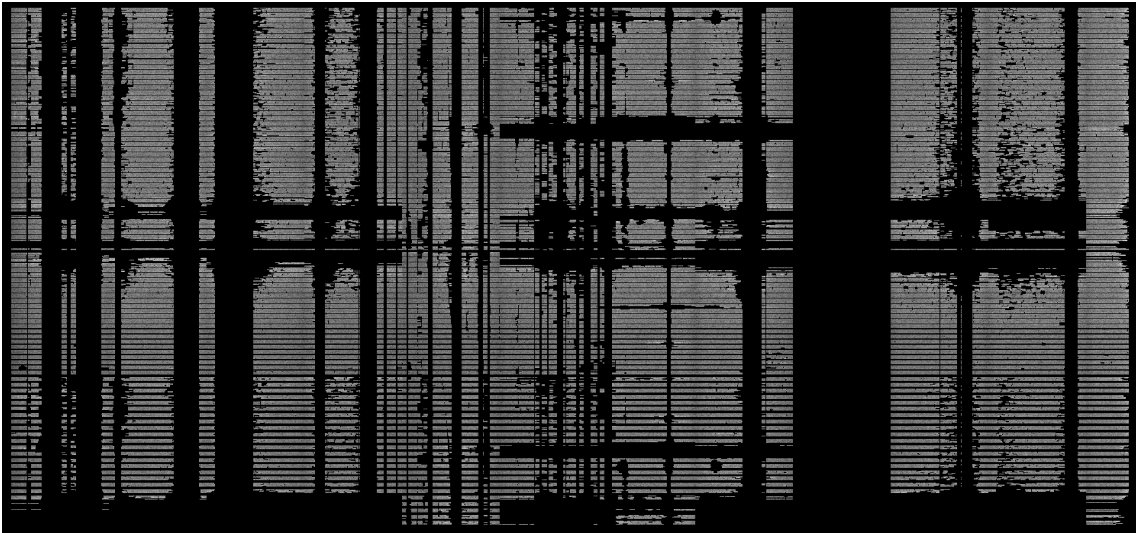

Figure 4: AIPS SPFLG image of a target pointing for COBRaS: 2033+4113, L band, baseline $6 - 8$, all stokes combined, IFs: 12. The AIPS task REFLG was also deployed in this image. The vertical axis is time and horizontal axis is frequency.

2.5 Computational Performance

SERPent has been parallelized to increase the performance by splitting the data via baselines and IFs and distributing these 'jobs' over multiple CPUs. Figure 6 demonstrates the benefit of this optimization on some commissioning L band test data on a couple of computing systems.

A typical SERpent run on a single COBRaS commissioning C band observation (5 sources including calibration sources, 132 GB raw, 25 GB averaged, 14 hours total) took 6 hours on a 100 GB, 16 CPU computer.
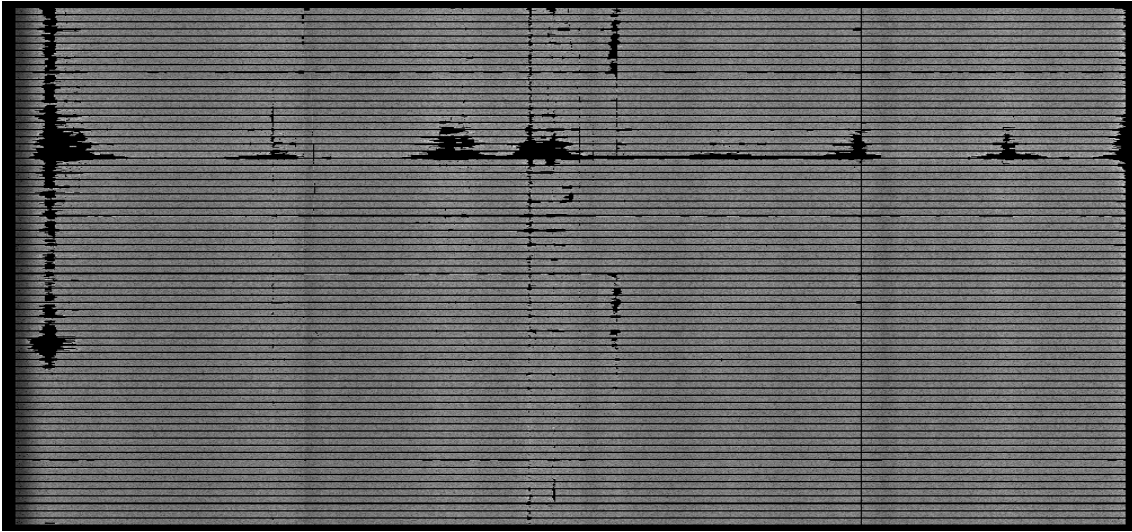
Acknowledgments

Figure 5: AIPS SPFLG image of a target pointing for COBRaS: 2032+4113, C band, baseline $1-5$, all stokes combined, IFs: 4. The AIPS task REFLG was also deployed in this image. The vertical axis is time and horizontal axis is frequency.
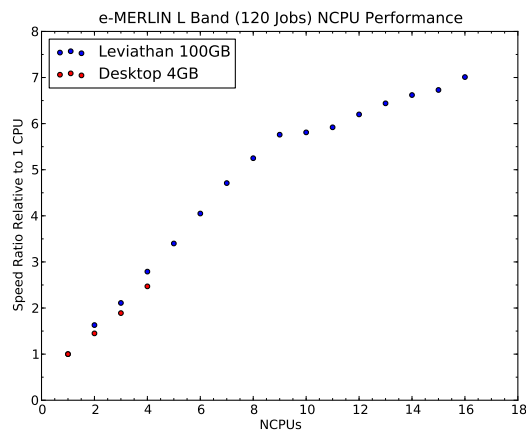

Figure 6: showing the time taken to flag the RFI TEST dataset (1.63 GB) using a common Desktop computer (4 GB, 4 CPUs) and Leviathan (100 GB, 16 CPUs) over a range of CPUs. Each point is an average of 3 runs using the same number of CPUs.

test data, Megan Argo and everyone who tested the performance of SERPent.

References

[1] A.R. Offringa, A.G. de Bruyn, M. Biehl, S. Zaroubi, G. Bernardi, V.N. Pandey, *Post-correlation radio frequency interference classification methods*, *MNRAS*, 2010MNRAS.405..155O, [arXiv:1002.1957]

[2] P.A. Fridman, *Statistically Stable Estimates of Variance in Radio-Astronomy Observations as Tools for Radio-Frequency Interference Mitigation*, *AJ*, 2008AJ....135.1810F, [arXiv:1009.5655]

[3] N. Niamsuwan, J.T. Johnson, S.W. Ellingson, *Examination of a simple pulse-blanking technique for radio frequency interference mitigation*, *Radio Science*, 2005RaSc...40.5S03N