PROCEEDINGS OF SCIENCE

ngDIFMAP : new generation DIFMAP for Modelfitting Interferometric Data

Agniva Roychowdhury[a,*] **and Eileen T. Meyer**[a]

[a] *Department of Physics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA*

E-mail: agniva.physics@gmail.com

We present a new generation version of `DIFMAP` (`ngDIFMAP`) to model and fit interferometric closure quantities. `ngDIFMAP` uses a global optimization algorithm based on simulated annealing, which results in more accurate parameter estimation especially when the number of parameters is high. Using this package we demonstrate the ramifications of amplitude and phase errors, and loss of $u-v$ coverage on estimated parameters, in terms of variance, bias and correlations between parameters. In addition to increasing correlations, we find that amplitude and phase errors can cause significant increase in variance and bias for all parameters for a simple jet model, and we provide prescriptions to take this into account in reporting errors on best-fit model parameters. This software serves as the pillar to our new program called the 'Catalogue of proper motions in extragalactic jets from Active galactic Nuclei with Very large Array Studies' or CAgNVAS, designed with the objective of using archival and new VLA observations to measure proper motions of jets on kpc scales. Since this objective requires extremely high accuracy in component localization, and very robust understanding of source structure, `ngDIFMAP` is ideally suited for this purpose. Our results are, in general, a pathway to understanding noisy interferometric data more accurately than ever before.

15th European VLBI Network Mini-Symposium and Users' Meeting (EVN2022)
11-15 July 2022
University College Cork, Ireland

*Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

PoS(EVN2022)059

1. Introduction

A model-independent measurement of the bulk Lorentz factor in black hole jets is essential to constrain open questions related to jet launching, acceleration and origin of high-energy emission. Until now, most studies concentrated on pc scale proper motions with VLBI of bright compact knots. We present here the program 'Catalogue of proper motions in extragalactic jets from Active galactic Nuclei with Very large Array Studies' or CAgNVAS, with the objective of using archival and new VLA observations to measure proper motions of jets on kpc scales. This objective requires extremely high accuracy in component localization. Fitting of the complex visibilities in radio interferometry with a very popular tool called `DIFMAP` (Difference Mapping, [7]), which is fairly common to deduce underlying source morphology without the biases introduced from image plane analysis and is generally used for VLBI source structure. However, the `modelfit` function inside DIFMAP, that is used for this purpose, uses a set of simple model components by a local minimizer, or a Levenberg-Marquadt optimization algorithm. Furthermore, even though radio interferometric data can have multiple sources of error, the uncertainties on best-fit parameters are determined simply from the curvature of the χ^2 surface. Additionally, the local minimizer is not ideal when high accuracies are needed. We remedy the disadvantages of model fitting in original `DIFMAP` by modifying it to create a new generation `DIFMAP` (`ngDIFMAP`) that contains a global optimizer which fits interferometric closure quantities. The global optimizer allows more accurate determination of best-fit values of the source structure parameters. Usage of closure quantities partly removes the issue of dealing with antenna-based gain errors. Additionally, in order to determine the biases, variances and correlations between these parameters, a new functionality inside `ngDIFMAP` allows determination of approximate effects of errors in interferometric data on best-fit parameters. All of these improvements make more accurate positional measurement of knots in large-scale jets possible, in addition to moving towards a more robust methodology to understand source structure.

In this paper, we introduce `ngDIFMAP` and discuss major sources of bias and related uncertainties in determining mean values of best-fit parameters in model-fitting radio interferometric datasets, with a special focus on jets from radio-loud active galactic nuclei (AGN).

2. Closure Quantities

Closure quantities are "good" interferometric observables independent of antenna-based gain errors and hence a robust estimator of the source structure. For example, in presence of an amplitude and phase error, the observed visibility $\tilde{V}$, as a function of the true visibility V, would be given as:

$$\tilde{V}_{ij} = (1 + a_i)(1 + a_j)e^{i(\phi_i - \phi_j)} A_{ij} e^{i\Psi_{ij}} + \epsilon_{ij}(t) \tag{1}$$

where a_i, a_j are the amplitude errors and ϕ_i, ϕ_j are the phase errors per antenna. The true visibility $V_{ij} = A_{ij}e^{j\Psi_{ij}}$ with true amplitude and phase A_{ij} and Ψ_{ij} respectively. If a triangle is formed from three such baselines, the observed "closure" phase around these antennas would be given by the cyclic sum of the *observed* phases:

$$\tilde{\Psi}_{ijk}(t) = [\tilde{\Psi}_{ij} + \tilde{\Psi}_{jk} + \tilde{\Psi}_{ki}] = [\Psi_{ij} + \Psi_{jk} + \Psi_{ki}] + [\epsilon'_{ijk}] \tag{2}$$

PoS(EVN2022)059

which is clearly independent of the phase errors ϕ and is a representation of the "true" phase, to a first order in a thermal noise ϵ'.

Similarly, if a closure quadrangle is formed from four baselines, a possible closure amplitude is given by:

$$A_{ijkl} = \frac{|\tilde{V}_{ij}||\tilde{V}_{kl}|}{|\tilde{V}_{ik}||\tilde{V}_{jl}|} = \frac{|V_{ij}||V_{kl}|}{|V_{ik}||V_{jl}|} \tag{3}$$

As evident, multiplicative errors but *specific* to every antenna can be disposed of using this method, making it a very robust estimator of the source structure. The information about the absolute flux density and position are of course lost in the process, which can otherwise be obtained from default methods. Therefore, using closure quantities to image or fit models to radio observations is a so-called "calibration-insensitive" way to constrain intrinsic source properties, as used by the EHT team for example [e.g., 3].

3. Code Description

`DIFMAP`, written in C, is highly modular and has functionalities ranging from plotting the visibility data to deconvolution given in separate source code files. Particularly relevant for this purpose are `modfit.c` and `lmfit.c` which are dedicated to model fitting and `uvf_read.c` that "read" in a `uvfits` dataset. Hence we mainly edited three source codes to suit our purpose in `ngDIFMAP`. The main edits can be summarized as follows in the corresponding subsections. We edited `DIFMAP` and did not write a code anew is because of how accepted and well tested it is in the field: a new code necessarily entails a number of disadvantages, including bug fixing and maintaining it over time. In contrast, it was much more suitable to use the engine of `DIFMAP` and only add/modify functionalities.

3.1 Closure Quantity Fitting using Simulated Annealing

Assuming a chosen model using which visibilities are generated and closure quantities are computed, a new function `getnextclp()` inside `modfit.c` returns the data minus model residuals for closure amplitudes and phases to the model fitting function inside `lmfit.c`. We have followed the prescriptions of [3] and [1] for this purpose. The corresponding likelihood regularizer used is :

$$\chi^2 = \frac{1}{N_{\mathrm{cl-ph}} + N_{\mathrm{cl-amp}}} \left[\sum_{n=1}^{N_{\mathrm{cl-amp}}} \frac{|A_{\mathrm{n,obs}} - A_{\mathrm{n,model}}|^2}{\sigma^2_{\mathrm{n,cl-amp}}} + \sum_{n=1}^{N_{\mathrm{cl-ph}}} \frac{|e^{i\tilde{\psi}_{\mathrm{n,obs}}} - e^{i\tilde{\psi}_{\mathrm{n,model}}}|^2}{\sigma^2_{\mathrm{n,cl-ph}}} \right] \tag{4}$$

where the two sums are over only independent closure amplitudes and phases respectively with the corresponding subscripts. This of course only works when the closure quantities are negligibly correlated, implying the covariance structure is ignored. Although this can cause severe discrepancies when the number of antennas is low [1], this assumption is safe for arrays like the VLA,

PoS(EVN2022)059

where $N = 27$. For determining errors, we have employed other techniques as discussed later. The variances $\sigma_{\rm cl-ph}$ and $\sigma_{\rm cl-amp}$ follow from the same set of references. However, the corresponding closure amplitude error can diverge for low SNR [3] and deviate significantly from a Gaussian distribution, which will invalidate the used least-squares assumption. This is not as severe for the closure phase errors. Using log closure amplitudes makes the error distribution resemble more of a Gaussian distribution [3] and the least-squares estimator can be used. The resulting χ^2 is better written as [3]:

$$\chi^2_{\rm cl-amp} = \frac{1}{N_{\rm cl-amp}} \left[\sum_{n=1}^{N_{\rm cl-amp}} \frac{A^2_{\rm n,obs}}{\sigma^2_{\rm n,cl-amp}} \left(\log \left| \frac{A_{\rm n,obs}}{A_{\rm n,model}} \right| \right)^2 \right] \tag{5}$$

In the above, it is implicitly assumed that $A_{model} \neq 0$, and for $0 < A_{model} \ll 1$, the initial choice of model is not very far away from the observation, otherwise a logarithmic divergence is possible. It essentially boils down to practice which one of the above prescriptions may be chosen and the user may choose any inside `ngDIFMAP`. If the latter, then the corresponding part in Equation 4 is replaced by the above. Furthermore, inside `lmfit.c` we added few functions dedicated to "Simulated Annealing" (SA), which is a global optimization algorithm [e.g., 4]. It is a variant of the Metropolis algorithm, where it starts from an initial parameter set and a "temperature" and evaluates the neighbouring parameter space randomly for a number of iterations. For each case, it measures the corresponding Gibbs probability given by $p \sim \exp(-\chi^2/T)$. While the solution with the highest probability is chosen, worse solutions are not always rejected (with a threshold) to prevent falling into possible local minima. This algorithm is repeated for a given number of times (or function evaluations) and at each step the temperature dropoint sourceand the probability of rejections decreases, when the algorithm slowly converges to the correct global minima. The speed and accuracy of this algorithm is dependent on a number of "tuning" parameters, whose description can be found in the code documentation and in [4]. The χ^2 is fed to this optimizer inside `lmfit.c` and it accordingly searches for the global minima. The χ^2 can be user-described, which can be complex visibilities, or closure quantities. However, since this is a global optimizer, it takes considerably more time than the present Levenberg-Marquadt (LM) optimizer inside `DIFMAP`. For VLA data, it can take anywhere between one and seven days, which is heavily dependent on the size and simplicity of the dataset. An option for partial parallelization using `openmp` has been provided in `lmfit.c` but it only reduces the total time by at most 10%. Parallelization using the Message Passing Interface (MPI) is not possible without rewriting entire `DIFMAP`. The user has the option to use any of the above (SA or LM), but the LM optimizer has not been modified to incorporate closure quantity fitting since it requires completely rewriting the existing the fitting module of the code. This will be done in a future endeavour.

PoS(EVN2022)059

3.2 Editing Visibility Data as desired

This source code reads the complex visibilities from a `uvfits` file through function `get_uvdata()`. Therefore, the data being fed to `modelfit` can be directly modified by editing `uvf_read.c`. This is the crux of the modification: manipulate the data in any desired way. Furthermore, a few additions inside `modfit.c` allows to print the model visibilities to an ASCII file, which can then be read through `uvf_read.c` in a second run *instead* of the observed uvfits file. We have added this functionality, implying *any* ASCII model visibility can be loaded inside `DIFMAP` using `uvf_read.c`, be modified according to *any* prescription (edit $u-v$ coverage, change normalization, add errors, etc.) and the corresponding effects on the `clean` image as well as `modelfitting` can be noted. This opens up a vast sea of possible analyses with *any* kind of interferometric data. This functionality, although very straightforward, is unfortunately not available in any publicly radio data analysis software. For most cases, one needs to develop their own code, like the Event Horizon Telescope (EHT) collaboration, in which case they are not always publicly available [e.g., 2]. Our addition to `DIFMAP` makes the radio data analysis suite complete for any radio astronomer who has visibility data observed by *any* telescope.

4. Demonstration of Capabilities

4.1 Simulating real $u-v$ coverage and gain errors

The type of datasets tested in this work relate to high-frequency ($\gtrsim$ 5 GHz) Very Large Array (VLA) observations that generally have minimal background noise with majorly the target in the field of view. A compact source ($\sim$ few arcsec) implies that a simple Fourier Transform relation between visibilities and the sky brightness exists.

For any pair of antennas represented by i, j, the actual observed visibility $\tilde{V}$ can be represented as:

$$\tilde{V}_{ij} = H_{ij} G_i G_j^* [V_{ij} + \epsilon_{ij}] \tag{6}$$

where V is the true visibility and ϵ represents additive thermal noise. G_i and G_j are the corresponding final antenna gains, $H_{ij}(t)$ can represent a baseline-based correction, which are estimated *after* flux and bandpass calibration. For this study we shall define a "real" dataset as one which has residual gain errors as well as imperfect $u-v$ coverage. For ease of modelling these errors, modifying the prescription of [3], we have more generally modelled $\tilde{V}$ with time-dependent errors as:

$$\tilde{V}_{ij} = [1 + Y_{ij}(t)] e^{j\,\delta\phi_{ij}} [V_{ij} + \epsilon_{ij}] \tag{7}$$

where Y_{ij} is a random variable $\sim \mathcal{N}(0, \sigma_g)$, where σ_g represents the level of gain error (1 for 100%). $\delta\phi_{ij}$ is a random variable $\sim \mathcal{N}(0, \sigma_p)$, with σ_p referring to the phase error in radians. Y_{ij} and $\delta\phi_{ij}$ are drawn for every (u, v or i, j) point that is sampled. In general, it is impossible to have a single accurate prescription that would produce the imperfections of the radio array, consequent calibration and self-calibration in the antenna gains. Hence, to first approximation, we have used

PoS(EVN2022)059

a Gaussian distribution, which will be found to mimic the scatter in real data adequately (see next subsection).

Even if a proper visibility covariance matrix is prescribed for the fit, the best-fit parameters may be correlated. A Monte Carlo "bootstrap" method is an effective way to "repeat" the same observation numerous times without writing observing proposals, to understand uncertainties and bias in estimation of best-fit parameters of the model. It is vaguely similar to FR-RSS (Flux Randomization and Random Subset Selection; [6]) or bootstrapping to understand uncertainties in VLBI rotation measure mapoint sourceof jets [5]. This can be used further if one uses a pre-defined model in a well sampled $u - v$ plane as a starting point and progressively "worsens" the $u - v$ coverage *randomly*, and then follows the spread or bias in the final parameter estimation using the resulting parameter covariance matrix.

4.2 Estimating degree of errors in real datasets

To mimick a real dataset is to model its possible errors. While it is very difficult from first principles since the antenna complex gains are a priori unknown, Equation 7 rather provides a simple prescription to provide a rough estimate of the same. Assuming that the real dataset is best described by a specific model, the deviations from the model, or essentially the average scatter of the dataset, must encode the average property of the errors it is being afflicted by. In this regard, we define a "realness" parameter R, of any synthetic dataset, that characterizes how close its *average* visibility scatter is to a given observed dataset for the same science target. It can be defined as $R = \sigma_{real}/(\sigma_{real} - \sigma_{model})$, where σ is the standard deviation per bin in the total visibility data. This quantification is strictly visual and is also based on the fact that real data are more "noisy", and thus more scattered. Figure 1 demonstrates this idea. It shows a sample visibility amplitude data of 3C 78 observed at 15 GHz in 1985 with the VLA, which shows considerable scatter. The best-fit model for the source structure is also given on the right, which when acted upon by amplitude errors of $\sim 8\%$, begins to resemble the observed visibilities, on an average. This is verified by the plot of the Realness parameter, which peaks at $\sigma_g \sim 8\%$.

A similar method can be used for the visibility phases and the corresponding σ_p determined. Equipped with an estimate of the possible σ_g and σ_p, it is then straightforward to begin with the synthetic dataset for the starting model, apply corresponding errors and inspect the variance, bias and correlation plots. The bias can be then used to predict how deviant the best fit model parameters are from the intrinsic source description, while the variance will provide an estimate of the spread in the best-fit parameter. In contrast, since the correlations between different parameters are also dependent on the errors, they also must be quoted when mentioning the true best-fit parameters. The only real caveat in this approach is the assumption of a model for the antenna gain errors. While it will most likely not affect the variances, it is a possible exercise to determine the dependence of the biases and correlations on the prescription of the gain error. Note that error prescriptions, different from what we considered here, might create correlations between visibility data, in which case the corresponding covariance matrix for the model fit for every Monte Carlo run needs to be modified. However, this task is beyond the scope of this work and may be discussed in an upcoming paper.

PoS(EVN2022)059

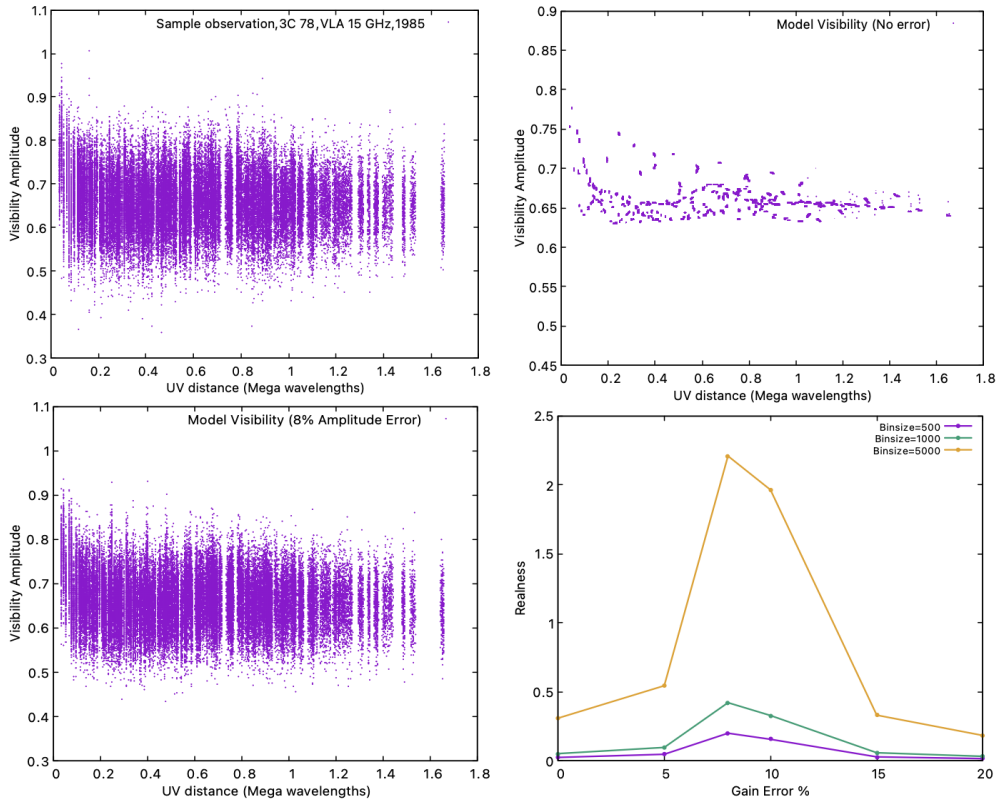

Figure 1: Figure shows how a synthetic dataset can be made noisy to represent a real dataset. Top left: Visibility amplitude v/s UV distance for a VLA 15 GHz observation of radio galaxy 3C 78. Top right: Corresponding best-fit model of the data. Bottom Right: Values of the Realness parameter between the model visibility amplitude and the observed visibility amplitude for different strengths of amplitude error. It peaks at 8% amplitude error. Bottom left: The synthetic dataset (model) *after* application of 8% amplitude error, which looks very similar to the real visibility amplitudes, on an average.

4.3 Effects of errors on uncertainties

We start from a simple model of a jet that consists of a bright point source in addition to a knot modelled by a two-dimensional Gaussian. We use a range of parameters for the core and the knot, and for each corresponding configuration, we check the correlations between the parameters and their variances by creating $N = 1000$ realizations per given $u - v$ coverage. The $u - v$ coverage is worsened by choosing visibilities randomly in progressively smaller fractions and the bias in parameters noted. We describe the entire setup using a covariance matrix σ, where we check the dependence of the off-diagonal as well as the diagonal (variance) terms on the $u - v$ coverage and choice of the initial starting model for different sets of parameters.

Figure 2 shows the effect on the parameter space for dataset with a non-ideal $u - v$ coverage (denoted by $\beta/\beta_0 = 0.1$ or 0.0005, with β_0 being any given prescription of the default coverage), an amplitude error σ_g of 15% and a phase error of $\sigma_p = 0.2$ radian. The other parameters $d = 200$ mas and $w = 200$ mas refer to the distance between the Gaussian (knot) and the core (point source), and the Gaussian width, respectively. $F = 10$ mJy represents the Gaussian flux density. Plots like these can be produced for any set of initial conditions. Further, plots of the correlation (σ_{xy}), bias

PoS(EVN2022)059

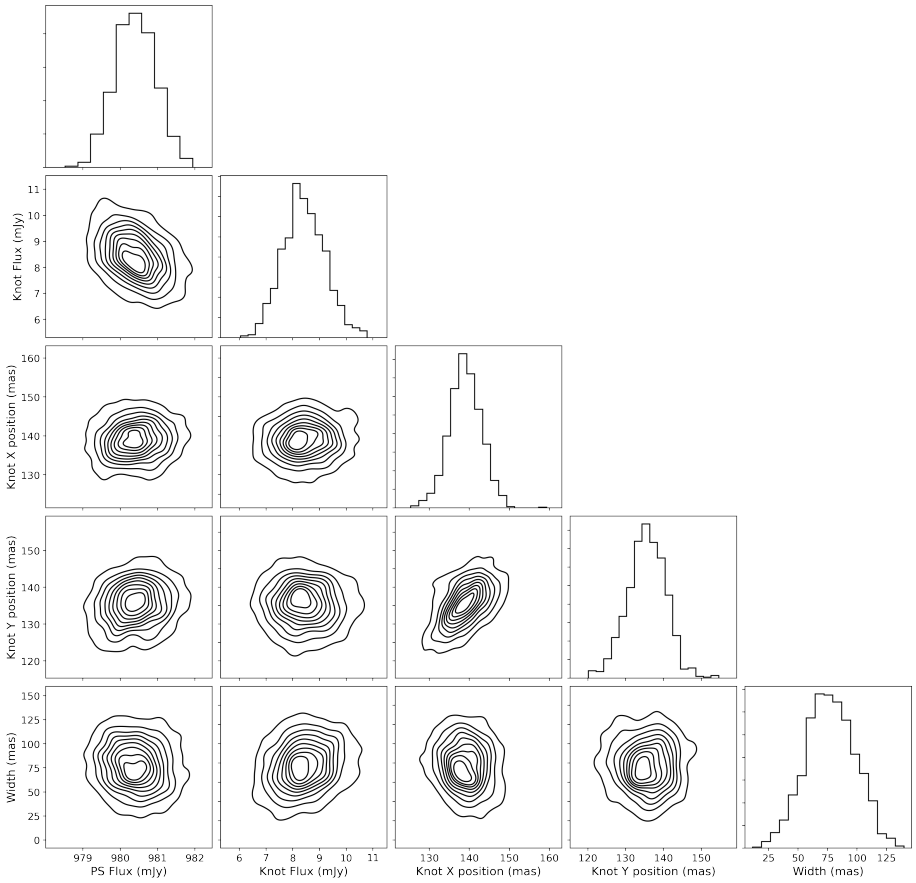

Figure 2: Marginal probability distributions for selected parameters obtained from bootstrapping, for configuration with d=200, w=200, F=10 mJy, $\beta/\beta_0 = 0.1$, $\sigma_g = 0.15$ and $\sigma_p = 0.2$. Since the scales are not similar, it is difficult to discern correlations if there is any. The Gaussian X and Y positions are clearly correlated, since they are equally affected by amplitude and phase errors. The Gaussian flux density and the unresolved point source flux density are negatively correlated, as observed previously too. Contours were made from a kernel density estimator and are *not* representative of any confidence region.

$((x_{obs}/x_{expected} - 1)$ for any parameter x) and variance (σ_{xx}) can be separately plotted for any configuration for a specified independent parameter that is tuned accordingly.

In contrast, for $\beta/\beta_0 = 0.0005$, the biases and variances become absurdly large for the worse case of the amplitude and phase error. Figure 3 shows the effects of errors on bias and variances at $\beta/\beta_0 = 0.0005$ as a function of the flux density. While the flux density and positional standard deviations are thousands in mJy and mas respectively, the bias is between few tens and few hundred percent for the lower flux density cases. It is clear that while such a poor $u - v$ coverage by default works, in a real dataset with combined amplitude and phase errors, the dataset can be essentially considered irretrievable. The limiting β until which substantial information can still be recovered is dependent on the source structure and the $u - v$ coverage *pattern*. However, in the limit of highest flux density, the biases and variances are much lower. This result shows that the biases and variances are more heavily affected by $u - v$ coverage worsening when there are already amplitude and phase errors present. Future work in this direction will throw light on applicability of model

PoS(EVN2022)059

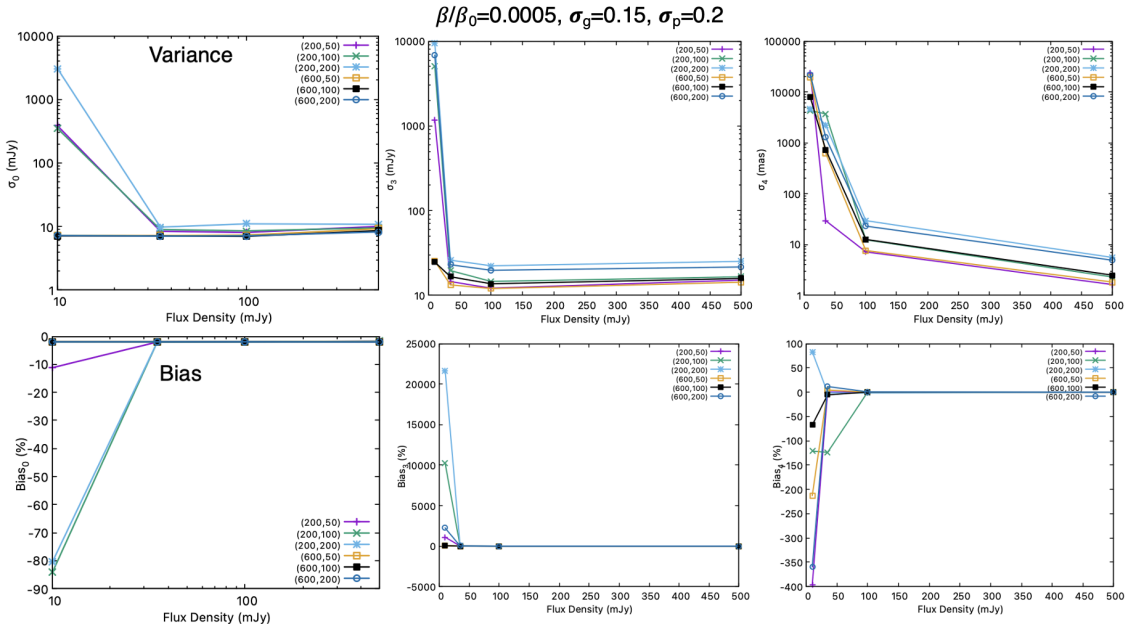

Figure 3: Variances and biases of parameters 0 (point source flux density), 3 (Gaussian flux density) and 4 (Gaussian X position) at $\beta/\beta_0 = 0.0005$, $\sigma_g = 0.15$ and $\sigma_p = 0.2$, as a function of the flux density. All of them are absurdly high $\sim$ orders of magnitude at lower flux densities. This shows that the data is practically unusable if the flux density is not high enough. The results hint at the limitations of fitting models to visibilities in the low SNR regime.

fitting in the low SNR regime using VLBI where the $u - v$ coverage is poor.

5. Conclusion and Future Goals

This conference proceeding is part of a larger article from the CAgNVAS project, which aims to create a catalogue of proper motions of AGN jets using the Very Large Array (VLA) archive. Motivated by the need for accurately measuring VLA proper motions and associated biases, we have presented `ngDIFMAP`, created by adding closure quantity fitting, a global optimizer and added functionalities to `DIFMAP` to make parameter estimation from model-fitting in the $u - v$ plane more robust. Particularly, we have demonstrated in this paper the effects of interferometric errors on the visibility data and subsequent parameter estimation, which were not demonstrated in any previous publication. `DIFMAP` is by default written in C and runs very efficiently for almost any interferometric data. We have capitalized on this feature to add new functionalities, which would be useful to any radio astronomer. We are currently attempting to add more features, that include more complex models inside `modelfit` and a possible multi-resolution CLEAN. `ngDIFMAP` hasn't been made public yet since it requires further structuring for a public distribution. We will provide the code and the documentation on a personal request basis until then.

6. Acknowledgment

The authors acknowledge National Science Foundation (NSF) Grant 12971 for the support.

PoS(EVN2022)059

References

[1] Blackburn, L., Pesce, D. W., Johnson, M. D., et al. 2020, , 894, 31, doi: 10.3847/1538-4357/ab8469

[2] Broderick, A. E., Gold, R., Karami, M., et al. 2020, , 897, 139, doi: 10.3847/1538-4357/ab91a4

[3] Chael, A. A., Johnson, M. D., Bouman, K. L., et al. 2018, , 857, 23, doi: 10.3847/1538-4357/aab6a8

[4] Goffe, W. L., Ferrier, G. D., & Rogers, J. 1994, Journal of Econometrics, 60, 65, doi: https://doi.org/10.1016/0304-4076(94)90038-8

[5] Pashchenko, I. N. 2019, , 482, 1955, doi: 10.1093/mnras/sty2654

[6] Peterson, B. M., Wanders, I., Horne, K., et al. 1998, , 110, 660, doi: 10.1086/316177

[7] Shepherd, M. C., Pearson, T. J., & Taylor, G. B. 1994, in Bulletin of the American Astronomical Society, Vol. 26, 987–989

PoS(EVN2022)059