Radio Imaging with Information Field Theory

Philipp Arras*†, Jakob Knollmüller*, Henrik Junklewitz and Torsten A. Enßlin*
*Max-Planck Institute for Astrophysics, Garching, Germany
†Technical University of Munich, Munich, Germany

***Abstract*—Data from radio interferometers provide a substantial challenge for statisticians. It is incomplete, noise-dominated and originates from a non-trivial measurement process. The signal is not only corrupted by imperfect measurement devices but also from effects like fluctuations in the ionosphere that act as a distortion screen. In this paper we focus on the imaging part of data reduction in radio astronomy and present RESOLVE, a Bayesian imaging algorithm for radio interferometry in its new incarnation. It is formulated in the language of information field theory. Solely by algorithmic advances the inference could be speed up significantly and behaves noticeably more stable now. This is one more step towards a fully user-friendly version of RESOLVE which can be applied routinely by astronomers.**

I. Introduction

To explore the origins of our universe and to learn about physical laws on both small and large scales telescopes of various kinds provide information. An armada of telescopes including many radio telescopes all over the earth and in space collect data to be put into one consistent theoretical picture of our universe by astrophysicists. Radio interferometers are of specific interest from a data reductionist's point of view since they do not measure a direct image of the sky as optical telescopes do. As a consequence radio interferometers provide only very incomplete information about the patch of the sky they are looking at. These two factors render the problem of radio imaging non-trivial and in order to obtain high-quality images sophisticated statistical methods need to be developed and applied.

In this paper, we want to present the latest state of the art of reducing data from radio interferometers with the help of *information field theory* (IFT) [1].

IFT is a statistical field theory which enables statisticians to solve complex Bayesian inference problems which involve fields. A field is a physical quantity defined over a continuous space like a three-dimensional density field or two-dimensional flux field. Treating these fields as continuous objects IFT does not suffer from side-effects induced by introducing a pixelation scheme right from the beginning. Moreover, a theory formulated in the language of fields enables IFT statisticians to employ the machinery having been developed by field theorists.

The algorithmic idea presented here is called RESOLVE (**R**adio **E**xtended **SO**urces **L**ognormal decon**v**olution **E**stimator) and was first presented in [2]. Since then the inference machinery has evolved dramatically with subsequent speedups of a factor of around 100.

This paper is organised as follows: In section II the measurement principle of radio interferometers is outlined. Section III gives a quick introduction to information field theory followed by section IV in which the Bayesian hierarchical model used by RESOLVE is explained. We conclude with an application on real data in section V.

II. Measurement Process and Data in Radio Astronomy

Radio telescopes measure the electromagnetic sky in wavelengths from $\lambda = 0.3\,\mathrm{mm}$ (lower limit of ALMA[1]) to $30\,\mathrm{m}$ (upper limit of LOFAR[2]). This poses a serious problem. The angular resolution of a single-dish telescope $\delta\theta$ scales with the wavelength λ divided by the instrument aperture D:

$$\delta\theta = 1.22\,\frac{\lambda}{D}.$$

As an example consider $\lambda = 0.6\,\mathrm{cm}$ and $\delta\theta = 0.1\,\mathrm{arcsec}$ which are typical values for the VLA[3]. Then the size of the aperture would need to be approximately $15\,\mathrm{km}$ which is not feasible technically. Therefore, many radio telescopes apply a different measurement principle.

Radio telescopes like VLA are in fact *radio interferometers*. They consist of several antennas (a total number of 27 in the case of the VLA). The electromagnetic radio wave which arrives at each antenna is converted to a digital signal and sent to a central supercomputer, called *correlator*. As its name suggest, it correlates the signal of each antenna with every other antenna in temporal windows of typically around $10\,\mathrm{s}$. These correlation coefficients are called *visibilities*. Each visibility corresponds to the strength of excitation of a Fourier mode in image space. The distance between two antennas is proportional to the spatial frequency and the orientation of the antennas gives the orientation of the Fourier mode.

All in all, the radio interferometric measurement process is modeled by the *Radio Interferometric Measurement Equation* (RIME, [3]):

$$d_{pq} = \int I(l,m)e^{i(lu_p+mv_q)}\,dl\,dm + n_{pq}. \tag{1}$$

Put into words, the data is given by the Fourier transform of the flux distribution $I(l,m)$ where l and m are the direction cosines of the angular coordinates ϕ and θ on the sky. Please note that this formula is based on several assumptions and simplifications. First, this version of the RIME is only valid for narrow field of views since it assumes a flat sky.

[1] Atacama Large Millimeter Array, Chile
[2] Low-Frequency Array, Europe
[3] Karl G. Jansky Very Large Array, New Mexico

ISBN 978-90-827970-1-5 © EURASIP 2018

Second, it assumes that all antennas are located at the same altitude. Third, it does not account for different polarizations and assumes that the antennas simply measure Stokes I. Finally and perhaps most importantly, it assumes that the data has been perfectly calibrated for all possible instrumental and additional measurement effects (e.g. receiver instabilties, ionispheric interference, ...). In this paper we treat only radio imaging and build on top of data which is calibrated by established algorithms. In other words, it is assumed that the data is calibrated perfectly.

III. Information Field Theory

In a nutshell, IFT is information theory with fields. It is a framework which uncovers the connection between statistical field theory and Bayesian inference. Exploiting this connection enables us to translate all knowledge physicists have gathered about statistical field theory and thermodynamics to Bayesian inference.

The general idea is that given some finite data set d, it is inferred how likely different realizations of the observed physical field s is. This is done with the help of Bayes theorem which combines the likelihood $\mathcal{P}(d|s)$ with the prior knowledge $\mathcal{P}(s)$ and some normalization constant $\mathcal{P}(d)$ into the posterior distribution $\mathcal{P}(s|d)$:

$$\mathcal{P}(s|d) = \frac{\mathcal{P}(d|s)\mathcal{P}(s)}{\mathcal{P}(d)} = \frac{\mathcal{P}(s,d)}{\mathcal{P}(d)}.$$

This can be rewritten as:

$$\mathcal{P}(s|d) = \frac{1}{\mathcal{Z}(d)} e^{-\mathcal{H}(s,d)},$$

where $Z(d) := \int \mathcal{D}s\, \mathcal{P}(s,d)$ and $\mathcal{H}(s,d) := -\log \mathcal{P}(s,d)$. $\int \mathcal{D}s$ is the path integral which is defined as the continuum limit of the product of integrals over every pixel $\int \prod_i ds_i$. For details on that refer to [1, 4].

The above formula is well-known in statistical physics and inspires us to call $\mathcal{H}$ the *information Hamiltonian*. In order to obtain the maximum a-posterior estimate (MAP) of s one has to minimize $\mathcal{H}$ with respect to s because the exponential is a monotonic increasing function. Since the information Hamiltonian is given by

$$\mathcal{H}(s,d) = \mathcal{H}(d|s) + \mathcal{H}(s),$$

it knows both about the measurement process via the likelihood term $\mathcal{H}(d|s)$ and about the prior knowledge via $\mathcal{H}(s)$. Please note that additional constants in s can be dropped from $\mathcal{H}(s,d)$ since they only change the normalization of the posterior but not its shape. This will be indicated by "$\simeq$".

As an illustrative example, let us re-derived the famous Wiener filter [5]. Suppose we observe a noisy random process with known stationary signal and noise spectra and additive noise. More precisely, suppose we are given some measurement data d described by the following measurement equation:

$$d = Rs + n, \tag{2}$$

where d is a finite-dimensional vector, s is the unknown signal field and n the additive noise. s and n are assumed to be zero-centered Gaussian random fields drawn from $\mathscr{G}(s,S)$ and $\mathscr{G}(n,N)$, respectively, where the covariances S and N are known. R, the linear *Response operator*, models the measurement device and is also known. It maps the signal s defined over a continuous domain to a finite data vector d. Note that equation (1), the RIME, is of that form. Also note that in this specific case the response operator R contains a Fourier transform.

Let us compute the posterior distribution or equivalently the information Hamiltonian for this problem. The likelihood $\mathcal{P}(d|s)$ is essentially given by equation (2):

$$\mathcal{P}(d|s,n) = \delta(d - (Rs + n)).$$

Then marginalize over the noise field:

$$\mathcal{P}(d|s) = \int \mathcal{D}n\, \mathcal{P}(d|s,n)\mathcal{P}(n) = \mathscr{G}(d - Rs, N).$$

Combining this with the prior probability $\mathcal{P}(s) = \mathscr{G}(s,S)$ and taking the negative logarithm gives the information Hamiltonian:

$$\begin{aligned}\mathcal{H}(s,d) = \tfrac{1}{2}(d - Rs)^\dagger N^{-1}(d - Rs) + \tfrac{1}{2} s^\dagger S^{-1} s \\ - \tfrac{1}{2}\log|2\pi N| - \tfrac{1}{2}\log|2\pi S|,\end{aligned} \tag{3}$$

where $\cdot^\dagger$ denotes transposition and element-wise complex conjugation of a matrix or a vector. The above expression is a second order polynomial and the square in s can be completed:

$$\mathcal{H}(s,d) \simeq \tfrac{1}{2}(s - m)^\dagger D^{-1}(s - m),$$

where $m = Dj$, $j = R^\dagger N^{-1} d$ and $D^{-1} = S^{-1} + R^\dagger N^{-1} R$. In other words, the posterior probability distribution is

$$\mathcal{P}(s|d) = \mathscr{G}(s - m, D)$$

where m is called the Wiener filter solution.

In this fashion the Wiener filter turns out to be the simplest filter which can be build within the framework of IFT. Note that already here one of IFT's strength becomes apparent: Pixelation schemes have not appeared yet. This is a general feature of IFT. The theory is formulated with fields (which infinitely many degrees of freedom which are not pixelated yet). Only when the filter is implemented on the computer the fields become discretised. To this end the Python package NIFTy [6, 7, 8] provides customized functionality to implement IFT algorithms. It even enables the user to easily switch between different pixelation schemes.

IV. IFT Model for Radio Interferometers

In radio interferometry, the situation is somewhat more difficult than the Wiener filter scenario discussed so far: First, the radio sky cannot be sensibly modeled by a Gaussian random process since electromagnetic flux is always positive and varies on many different orders of magnitude: a radio source typically is many magnitudes brighter than the surrounding background flux. Second, we do not know the signal covariances S of the brightness distribution on the sky. Therefore, we need to infer it as well. And finally, the noise covariance provided by the telescope might not be entirely correct. Radio frequency

(a) Posterior mean m (logarithmic brightness).

(b) Relative error on m.

Fig. 1: Exemplary application of RESOLVE on real data which was taken in 2003 by the VLA of the source 3C405 also known as Cygnus A.

interference or calibration errors might enhance the error bars on the data significantly. Therefore, the noise level of each data point needs to be inferred as well. The underlying assumptions and priors of the following calculations are:

1) The sky obeys log-normal statistics, i.e. the measurement can be written as:

$$d = Re^s + n,$$

where s is a Gaussian field again and R is the linear response operator which maps the sky field onto visibilities.[4] This is the proper choice since it enforces positivity of the flux field and can easily vary on different scales.
2) s is drawn from a probability distribution describing a isotropic and homogeneous process.
3) Power spectra of s preferentially follow a power law. In other words, curvature on double-logarithmic scale in the power spectrum shall be punished in the inference.
4) The noise covariance matrix is diagonal: $N = \widehat{e^\eta}$, where η is a vector whose entries are the logarithms of the variance of every data point.[5]
5) Large noise covariances are punished by an Inverse-Gamma prior on η.
6) The posterior probability distribution can be approximated by $\widetilde{\mathcal{P}}(s,\tau,\eta|d) = \mathscr{G}(\xi-\xi^*,\Xi)\,\delta(\tau-\tau^*)\,\delta(\eta-\eta^*)$, where τ is the logarithm of the power spectrum, Ξ is the posterior covariance of the map estimation and the starred quantities are the means of the respective variables.

For starters let us introduce some notation. Because s is drawn from an isotropic and homogeneous probability distribution the Wiener-Khinchin theorem [9] implies that S is diagonal in Fourier space and its diagonal is given by a power spectrum $p(k)$:

$$S_{\vec{k}\vec{k}'} = (2\pi)^2 \delta(\vec{k}-\vec{k}')\, p(|\vec{k}|).$$

[4]Here and in the following, exponentials of vectors are understood to be taken element-wise.

[5]The hat operator $\widehat{e^\eta}$ denotes the diagonal operator with the vector e^η on its diagonal.

The power spectrum is a positive function, thus we can apply the same trick as for the sky map. Define:

$$p(|\vec{k}|) = e^{\tau(|\vec{k}|)}$$

For convenience define a projection operator $\mathbb{P}$ which sums all values of a field b in harmonic space which lie in one bin in the power spectrum:

$$b_{\vec{k}} = \mathbb{P}_{\vec{k}\kappa} a_\kappa = \frac{1}{\rho_k} \int_{|\vec{k}|=\kappa} p_\kappa,$$

where ρ_k is the bin volume. Defining $\mathcal{F}$ to be the Fourier transform mapping from harmonic space to signal space, the signal prior covariance S can be expressed as:

$$S = \mathcal{F}\left(\widehat{\mathbb{P}^\dagger e^\tau}\right)\mathcal{F}^\dagger.$$

Finally, we split the field s into two parts in harmonic space: $s = \mathcal{F}(A_\tau\, \xi)$. ξ is a white Gaussian random field, i.e. it has the covariance matrix $\mathbb{1}$, and $A_\tau = \mathbb{P}^\dagger \sqrt{e^\tau}$, i.e. it contains all information coming from the power spectrum.

With the above notation it is now possible to write down all Hamiltonians we need for the reconstruction. The Hamiltonian which is to be minimized for the ξ reconstruction is computed analogously to (3):

$$\mathcal{H}(\xi, d|\tau,\eta) \simeq \tfrac{1}{2}(d - Re^{\mathcal{F}(A_\tau\xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{\mathcal{F}(A_\tau\xi)}) + \tfrac{1}{2}\xi^\dagger\xi.$$

Since it will be needed later, the curvature of the above Hamiltonian is to be computed:

$$\begin{aligned}\Xi := \frac{\delta^2\mathcal{H}(\xi,d|\tau,\eta)}{\delta\xi\,\delta\xi^\dagger} &= A_\tau^\dagger (e^s)^\dagger R^\dagger N^{-1} R e^s A_\tau + \mathbb{1}\\ &\quad - (d - Re^s)^\dagger N^{-1} R e^s A_\tau A_\tau.\end{aligned}$$

The last term is not necessarily positive definite which is not allowed for a covariance operator[6]. However, this term is small in the vicinity of the minimum because it contains the residual $d - Re^s$. Therefore, it is dropped right from the beginning.

[6]Note that the curvature of the information Hamiltonian is at the same time used as an approximative covariance of the posterior.

ISBN 978-90-827970-1-5

The Hamiltonian for the power spectrum reconstruction has a very similar structure: The likelihood is accompanied by the prior. Here, we choose a smoothness prior on double-logarithmic scale. Δ is the Laplace operator acting on logarithmic scale $y = \log k$:

$$\mathcal{H}(\tau, d|\xi, \eta) \simeq \frac{1}{2}(d - Re^{\mathcal{F}(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{\mathcal{F}(A_\tau \xi)}) + \frac{1}{2\sigma^2}\tau^\dagger \Delta^\dagger \Delta \tau.$$

The parameter σ controls the strength of the smoothness prior.

The Hamiltonian for the noise covariance estimation has again the same structure except for the prior: Here, an Inverse-Gamma prior is employed:

$$\mathcal{H}(\eta, d|\xi, \tau) \simeq \frac{1}{2}(d - Re^{\mathcal{F}(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{\mathcal{F}(A_\tau \xi)}) + \eta^\dagger(\alpha - 1) + q^\dagger e^{-\eta} + \frac{1}{2}\mathbb{1}^\dagger \eta.$$

Note that the last term originates from the term $-\frac{1}{2}\log|2\pi N|$ in (3).

In order to compute an estimate for the posterior τ^* and η^*, the deviation between the correct posterior probability and the approximate one needs to be minimized. The metric of choice to compare probability distributions is the Kullbach-Leibler divergence:

$$\mathcal{D}_{\mathrm{KL}}(\tilde{\mathcal{P}}(\xi, \tau, \eta|d) \,\|\, \mathcal{P}(\xi, \tau, \eta|d)) = \int \mathcal{D}\xi\, \mathcal{D}\tau\, \mathcal{D}\eta\, \tilde{\mathcal{P}} \log \frac{\tilde{\mathcal{P}}}{\mathcal{P}}.$$

The posterior shall be approximated by the distribution:

$$\tilde{\mathcal{P}}(s, \tau, \eta|d) = \mathscr{G}(\xi - t, \Xi)\, \delta(\tau - \tau^*)\, \delta(\eta - \eta^*).$$

The integrals over τ and η simply collapse due to the δ-distributions. What remains are two objective function, one for the power spectrum and one for the noise covariance estimation:

$$\mathcal{D}_{\mathrm{KL},\tau} = \left\langle \frac{1}{2}(d - Re^{\mathcal{F}(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{\mathcal{F}(A_\tau \xi)}) \right\rangle_{\mathscr{G}(\xi - t, \Xi)} + \frac{1}{2\sigma^2}\tau^\dagger \Delta^\dagger \Delta \tau,$$

$$\mathcal{D}_{\mathrm{KL},\eta} = \left\langle \frac{1}{2}(d - Re^{\mathcal{F}(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{\mathcal{F}(A_\tau \xi)}) \right\rangle_{\mathscr{G}(\xi - t, \Xi)} + (\alpha - 1)^\dagger \eta + q^\dagger e^{-\eta} + \frac{1}{2}\mathbb{1}^\dagger \eta.$$

The expectation value $\langle \dots \rangle_{\mathscr{G}(\xi - t, \Xi)}$ can be computed by sampling from $\mathscr{G}(\xi - t, \Xi)$. For details on that refer to [10].

All in all, the complete inference algorithm for applying IFT to radio interferometric data has been derived. The free parameters of the machinery are: the strength of the smoothness prior on the power spectrum σ and the shape of the Inverse-Gamma prior on the noise covariance estimation α and q.

V. Application

Finally, let us apply the above derived Bayesian inference algorithm to real data. To this end, let us take a VLA measurement set of Cygnus A from 2003. It has a total integration time of 49100 seconds. Since we deal only with single-band imaging in this paper, let us take one channel centered at 327.5 MHz with a bandwidth of 2.8 Mhz. As prior settings we choose an uninformative flat Inverse-Gamma prior for the noise ($q = 10^{-5}$, $\alpha = 2$) and $\sigma = 1$ for the smoothness prior on the power spectrum.

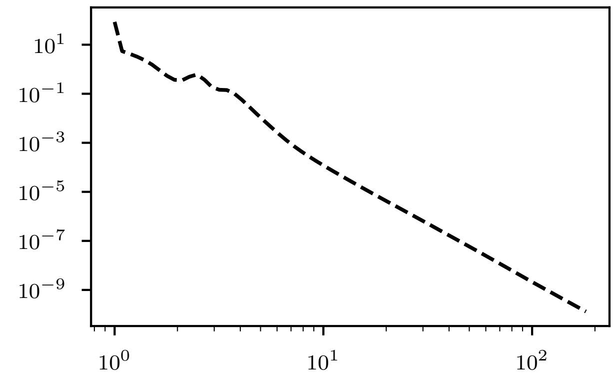

Fig. 2: Power spectrum of Cygnus A reconstruction.

The main result is presented in Figure 1. It shows the mean of the Gaussian which approximates the sky part of the posterior: $\mathscr{G}(s - m, D)$. Note that the figure shows the logarithmic flux. What singles out RESOLVE from many other imaging algorithms is its ability to provide an uncertainty map. It is depicted on the right-hand side on Figure 1. Additional to the sky model the algorithm learns the power spectrum e^τ as well. It is shown in Figure 2. Note that it does not possess much curvature on log-log scale as was expected by the Laplace prior on τ.

Finally, RESOLVE provides errorbars on the data points (see Figure 3). It is apparent the RESOLVE's error bars are five orders of magnitude bigger than the errorbars which are provided by the telescope.

The reconstruction was run on an Intel Core i5-4258U CPU using 300 MB main memory. The resolution of the reconstruction is 256^2 pixels for the sky model and 32 pixels in the power spectrum. The response operator R which incorporates a nonequispaced fast Fourier transform was implemented by employing the NFFT library which provides OpenMP parallelization [11].

The reconstruction including the analysis of the posterior statistics took approximately two hours of wall time.

VI. Conclusion

In this paper RESOLVE in its new incarnation was presented for the first time. Minimizing the Hamiltonian with respect to the map and the KL-divergence with respect to the power spectrum and the noise level provide a major speed-up. Also, the noise level of each data point was learned simultaneously with the map reconstruction for the first time. The main insights are:

- RESOLVE's noise estimation suggests a much higher noise level compared to the noise level which comes with the data set. This might be rooted in calibration artifacts which RESOLVE detects and puts into the noise.

ISBN 978-90-827970-1-5 © EURASIP 2018

(a) Error bars provided by telescope.

(b) Error bars provided by RESOVLE.

Fig. 3: Comparison of errorbars provided by the telescope and by RESOLVE. In the figures the logarithm of the variance of the data points is depicted.

- The migration from a simple fix-point iteration to minimization of Hamiltonian and KL-divergences was successful and is a big step forward towards an easy-to-use version of RESOLVE which can be shipped to a broad range of end-users.

The apparent next step towards a fully-integrated IFT radio data reconstruction pipeline is to include the calibration into the IFT inference. Other possible future work is to develop a fancier radio response function which can deal with wide-field images and to include point source reconstructions in the spirit of [12].

ACKNOWLEDGMENT

The authors would like to thank Rick Perley for the calibrated Cygnus A data and Landman Bester, Philipp Frank, Reimar Leike, Martin Reinecke, Oleg Smirnov and Rüdiger Westermann for numerous helpful discussions.

We acknowledge financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT).

REFERENCES

[1] T. Enßlin. “Information field theory”. In: *American Institute of Physics Conference Series*. Ed. by U. von Toussaint. Vol. 1553. American Institute of Physics Conference Series. Aug. 2013, pp. 184–191. DOI: 10.1063/1.4819999. arXiv: 1301.2556 [astro-ph.IM].

[2] H. Junklewitz et al. “RESOLVE: A new algorithm for aperture synthesis imaging of extended emission in radio astronomy”. In: *A&A* 586, A76 (Feb. 2016), A76. DOI: 10.1051/0004-6361/201323094. arXiv: 1311.5282 [astro-ph.IM].

[3] Oleg M. Smirnov. “Revisiting the radio interferometer measurement equation. I. A full-sky Jones formalism”. In: *Astron. Astrophys.* 527 (2011), A106. DOI: 10.1051/0004-6361/201016082. arXiv: 1101.1764 [astro-ph.IM].

[4] Information Field Theory. *Information Field Theory — Wikipedia, The Free Encyclopedia*. [Online; accessed 26-Feb-2018]. 2018. URL: http://en.wikipedia.org/w/index.php?title=Estimation_lemma&oldid=375747928.

[5] Norbert Wiener et al. “Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications”. In: *Bull. Amer. Math. Soc.* (1949). DOI: 10.1090/S0002-9904-1950-09416-6.

[6] Martin Reinecke, Marco, Selig, Theo Steininger. *NIFTy – Numerical Information Field TheorY*. Version nifty4. Feb. 5, 2018. URL: https://gitlab.mpcdf.mpg.de/ift/NIFTy.

[7] M. Selig et al. “NIFTY - Numerical Information Field Theory. A versatile PYTHON library for signal inference”. In: *A&A* 554, A26 (June 2013), A26. DOI: 10.1051/0004-6361/201321236. arXiv: 1301.4499 [astro-ph.IM].

[8] T. Steininger et al. “NIFTy 3 - Numerical Information Field Theory - A Python framework for multicomponent signal inference on HPC clusters”. In: *ArXiv e-prints* (Aug. 2017). arXiv: 1708.01073 [astro-ph.IM].

[9] Norbert Wiener. “Generalized harmonic analysis”. In: *Acta mathematica* 55.1 (1930), pp. 117–258.

[10] J. Knollmüller, T. Steininger, and T. A. Enßlin. “Inference of signals with unknown correlation structure from non-linear measurements”. In: *ArXiv e-prints* (Nov. 2017). arXiv: 1711.02955 [stat.ME].

[11] Jens Keiner, Stefan Kunis, and Daniel Potts. “Using NFFT 3—a software library for various nonequispaced fast Fourier transforms”. In: *ACM Transactions on Mathematical Software (TOMS)* 36.4 (2009), p. 19.

[12] D. Pumpe, M. Reinecke, and T. A. Enßlin. “Denoising, Deconvolving and Decomposing multi-Dimensional Photon Observations- The D4PO Algorithm”. In: *ArXiv e-prints* (Feb. 2018). arXiv: 1802.02013 [astro-ph.IM].

ISBN 978-90-827970-1-5