

# Signal Processing Tools for Radio Astronomy

Alle-Jan van der Veen and Stefan J. Wijnholds

**Abstract** Radio astronomy is known for its very large telescope dishes, but is currently making a transition towards the use of large numbers of small elements. For example, the Low Frequency Array, commissioned in 2010, uses about 50 stations, each consisting of at least 96 low band antennas and 768 high band antennas. For the Square Kilometre Array, planned for 2024, the numbers will be even larger. These instruments pose interesting array signal processing challenges. To present some aspects, we start by describing how the measured correlation data is traditionally converted into an image, and translate this into an array signal processing framework. This paves the way for a number of alternative image reconstruction techniques, such as a Weighted Least Squares approach. Self-calibration of the instrument is required to handle instrumental effects such as the unknown, possibly direction dependent, response of the receiving elements, as well as unknown propagation conditions through the Earth's troposphere and ionosphere. Array signal processing techniques seem well suited to handle these challenges. The fact that the noise power at each antenna element may be different motivates the use of Factor Analysis, as a more appropriate alternative to the eigenvalue decomposition that is commonly used in array processing. Factor Analysis also proves to be very useful for interference mitigation. Interestingly, image reconstruction, calibration and interference mitigation are often intertwined in radio astronomy, turning this into an area with very challenging signal processing problems.

---

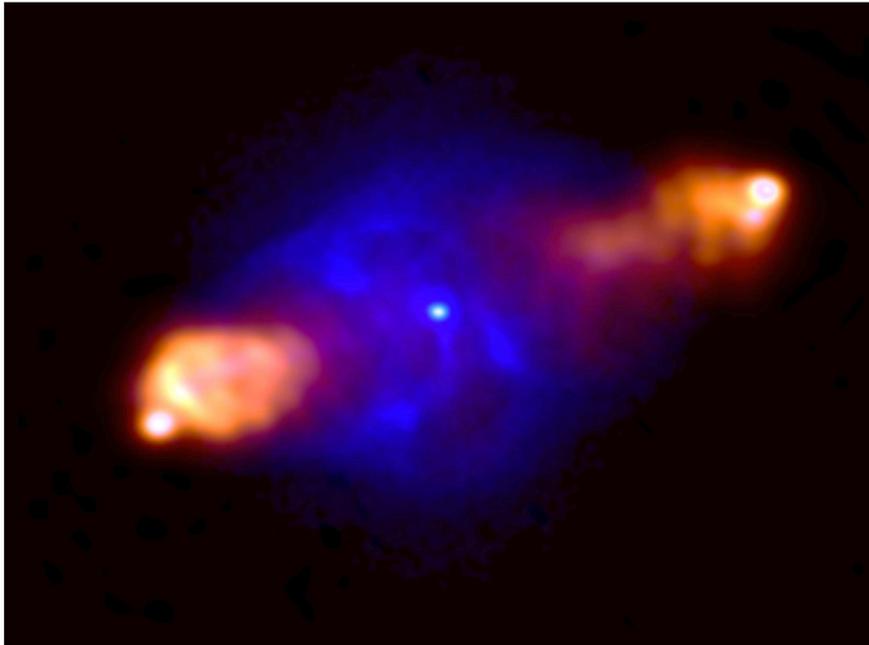
A.J. van der Veen  
TU Delft, Fac. EEMCS, Mekelweg 4, 2628 CD Delft, The Netherlands  
e-mail: a.j.vanderveen@tudelft.nl

S.J. Wijnholds  
Netherlands Institute for Radio Astronomy (ASTRON), Oude Hoogeveensedijk 4,  
7991 PD Dwingeloo, The Netherlands  
e-mail: wijnholds@astron.nl

## 1 Introduction

Astronomical instruments measure cosmic particles or electromagnetic waves impinging on the Earth. Astronomers use the data generated by these instruments to study physical phenomena outside the Earth's atmosphere. In recent years, astronomy has transformed into a multi-modal science in which observations at multiple wavelengths are combined. Figure 1 provides a nice example showing the lobed structure of the famous radio source Cygnus A as observed at 240 MHz with the Low Frequency Array (LOFAR) overlaid by an X-Ray image observed by the Chandra satellite, which shows a much more compact source.

Such images are only possible if the instruments used to observe different parts of the electromagnetic spectrum provide similar resolution. Since the resolution is determined by the ratio of observed wavelength and aperture diameter, the aperture of a radio telescope has to be 5 to 6 orders of magnitude larger than that of an optical telescope to provide the same resolution. This implies that the aperture of a radio telescope should have a diameter of several hundreds of kilometers. Most current and future radio telescopes



**Fig. 1** Radio image of Cygnus A observed at 240 MHz with the Low Frequency Array (showing mostly the lobes left and right), overlaid over an X-Ray image of the same source observed by the Chandra satellite (the fainter central cloud). (*Courtesy of Michael Wise and John McKean.*)

therefore exploit interferometry to synthesize a large aperture from a number of relatively small receiving elements.

An interferometer measures the correlation of the signals received by two antennas spaced at a certain distance. After a number of successful experiments in the 1950s and 1960s, two arrays of 25-m dishes were built in the 1970s: the 3 km Westerbork Synthesis Radio Telescope (WSRT, 14 dishes) in Westerbork, The Netherlands and the 36 km Very Large Array (VLA, 27 movable dishes) in Socorro, New Mexico, USA. These telescopes use Earth rotation to obtain a sequence of correlations for varying antenna baselines, resulting in high-resolution images via *synthesis mapping*. A more extensive historical overview is presented in [37].

The radio astronomy community is currently commissioning a new generation of radio telescopes for low frequency observations, including the Murchison Widefield Array (MWA) [24] in Western Australia and the Low Frequency Array (LOFAR) [42] in Europe. These telescopes exploit phased array technology to form a large collecting area with  $\sim 1000$  to  $\sim 50,000$  receiving elements. The community is also making detailed plans for the Square Kilometre Array (SKA), a future radio telescope that should be one to two orders of magnitude more sensitive than any radio telescope built to date [12]. This will require millions of elements to provide the desired collecting area of order one square kilometer.

The individual antennas in a phased array telescope have an extremely wide field-of-view, often the entire visible sky. This poses a number of signal processing challenges, because certain assumptions that work well for small fields-of-view (celestial sphere approximated by a plane, homogenous propagation conditions over the field-of-view), are no longer valid. Furthermore, the data volumes generated by these new instruments will be huge and will have to be reduced to manageable proportions by a real-time automated data processing pipeline. This combination of challenges led to a flurry of research activity in the area of array calibration, imaging and RFI mitigation, which are often intertwined in the astronomical data reduction.

The goal of calibration is to find the unknown instrumental, atmospheric and ionospheric disturbances. The imaging procedure should be able to apply appropriate corrections based on the outcome of the calibration process to produce a proper image of the sky. In this chapter, we review some of the array processing techniques that have been proposed for use in the data reduction pipelines, some of which are now being used in the LOFAR data reduction pipelines.

## 2 Notation

Matrices and vectors will be denoted by boldface upper-case and lower-case symbols, respectively. Entries of a matrix  $\mathbf{A}$  are denoted by  $a_{ij}$ , and its

columns by  $\mathbf{a}_i$ . Overbar  $\overline{(\cdot)}$  denotes complex conjugation. The transpose operator is denoted by  $T$ , the complex conjugate (Hermitian) transpose by  $H$  and the Moore-Penrose pseudo-inverse by  $\dagger$ . For matrices  $\mathbf{A}$  of full column rank, i.e.,  $\mathbf{A}^H \mathbf{A}$  invertible, this is equal to the left inverse:

$$\mathbf{A}^\dagger = (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H. \quad (1)$$

The expectation operator is denoted by  $E\{\cdot\}$ .

We will multiply matrices in many different ways. Apart from the usual multiplication  $\mathbf{A}\mathbf{B}$ , we will use  $\mathbf{A} \odot \mathbf{B}$  to denote the element-wise matrix multiplication (Hadamard product), and  $\mathbf{A} \otimes \mathbf{B}$  to denote the Kronecker product,

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

We will also use the Khatri-Rao or column-wise Kronecker product of two matrices: let  $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots]$  and  $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots]$ , then

$$\mathbf{A} \circ \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1, \mathbf{a}_2 \otimes \mathbf{b}_2, \cdots].$$

Depending on the context,  $\text{diag}(\cdot)$  converts a vector to a diagonal matrix with the elements of the vector placed on the main diagonal, or converts a general matrix to a diagonal matrix by selecting its main diagonal. Further,  $\text{vec}(\cdot)$  converts a matrix to a vector by stacking the columns of the matrix.

Properties of Kronecker products are listed in, e.g., [29]. We will frequently use the following properties:

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD} \quad (2)$$

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{vec}(\mathbf{B}) \quad (3)$$

$$\text{vec}(\mathbf{A} \text{diag}(\mathbf{b}) \mathbf{C}) = (\mathbf{C}^T \circ \mathbf{A})\mathbf{b} \quad (4)$$

Property (3) is used to move a matrix  $\mathbf{B}$  from the middle of an equation to the right of it, exploiting the linearity of the product. Property (4) is a special case of it, to be used if  $\mathbf{B}$  is a diagonal matrix: in that case  $\text{vec}(\mathbf{B})$  has many zero entries, and we can omit the corresponding columns of  $\mathbf{C}^T \otimes \mathbf{A}$ , leaving only the columns of the Khatri-Rao product  $\mathbf{C}^T \circ \mathbf{A}$ .

A special case of (3) is

$$\text{vec}(\mathbf{a}\mathbf{a}^H) = \bar{\mathbf{a}} \otimes \mathbf{a} \quad (5)$$

which shows how a rank-1 matrix  $\mathbf{a}\mathbf{a}^H$  is related to a vector with a specific ‘‘Kronecker structure’’.

### 3 Basic concepts of interferometry and image formation

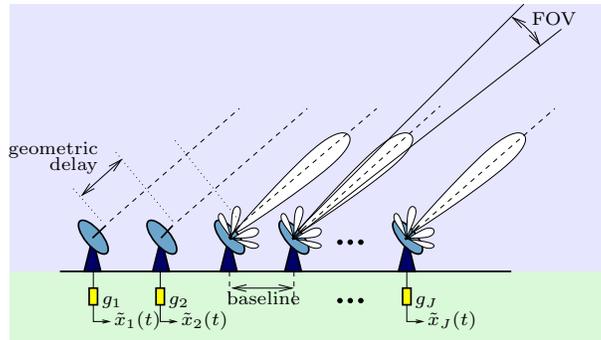
The concept of interferometry is illustrated in figure 2. An interferometer measures the spatial coherency of the incoming electromagnetic field. This is done by correlating the signals from the individual receivers with each other. The correlation of each pair of receiver outputs provides the amplitude and phase of the spatial coherence function for the *baseline* defined by the vector pointing from the first to the second receiver in a pair. In radio astronomy, these correlations are called the *visibilities*.

#### 3.1 Data acquisition

Mathematically, the correlation process is described as follows. Assume that there are  $J$  array elements (telescopes). The RF signal  $\tilde{x}_j(t)$  from the  $j$ th telescope is first moved to baseband where it is denoted by  $x_j(t)$ , then sampled and split into narrow subbands, e.g., of 100 kHz each, such that the “narrowband condition” holds. This condition states that the maximal geometrical delay across the array should be fairly representable by a phase shift of the complex baseband signal, and this property is discussed in more detail in the next subsection.

The resulting signal is called  $x_j(n, k)$ , for the  $j$ th telescope,  $n$ th time bin, and for the subband frequency centered at RF frequency  $f_k$ . The  $J$  signals are stacked into a  $J \times 1$  vector  $\mathbf{x}(n, k)$ .

A single correlation matrix is formed by “integrating” (summing) the crosscorrelation products  $\mathbf{x}(n, k)\mathbf{x}^H(n, k)$  over  $N$  subsequent samples,



**Fig. 2** Schematic overview of a radio interferometer.

$$\hat{\mathbf{R}}_{m,k} = \frac{1}{N} \sum_{n=(m-1)N}^{mN-1} \mathbf{x}(n,k) \mathbf{x}^H(n,k), \quad (6)$$

where  $m$  is the index of the corresponding ‘‘short-term interval’’ (STI) over which is correlated. The processing chain is summarized in figure 3.

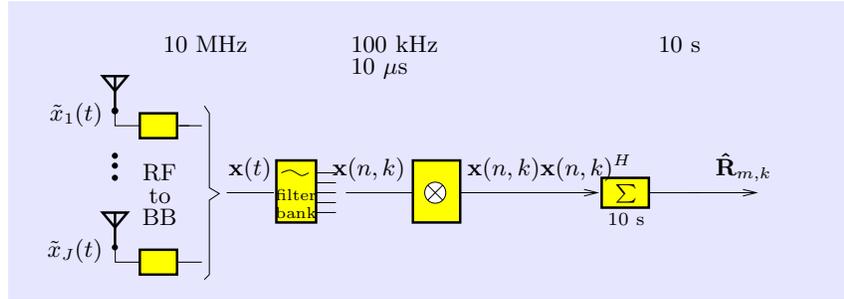
The duration of a STI depends on the stationarity of the data, which is limited by factors like Earth rotation and the diameter of the array. For the Westerbork array, a typical value for the STI is 10 to 30 s; the total observation can last for up to 12 hours. The resulting number of samples  $N$  in a snapshot observation is equal to the product of bandwidth and integration time and typically ranges from  $10^3$  (1 s, 1 kHz) to  $10^6$  (10 s, 100 kHz) in radio astronomical applications.

### 3.2 Complex baseband signal representation

Before we can derive a data model, we need to include some more details on the RF to baseband conversion. In signal processing, signals are usually represented by their low pass equivalents, which is a suitable representation for narrowband signals in a digital communication system, and also applicable in the radio astronomy context. A real valued bandpass signal with center frequency  $f_c$  may be written as

$$\tilde{s}(t) = \text{real}\{s(t)e^{j2\pi f_c t}\} = x(t) \cos 2\pi f_c t - y(t) \sin 2\pi f_c t \quad (7)$$

where  $s(t) = x(t) + jy(t)$  is the *complex envelope* of the RF signal  $\tilde{s}(t)$ , also called the *complex baseband signal*. The real and imaginary parts,  $x(t)$  and  $y(t)$ , are called the in-phase and quadrature components of the signal  $\tilde{s}(t)$ . In practice, they are generated by multiplying the received signal with  $\cos 2\pi f_c t$  and  $\sin 2\pi f_c t$  followed by low-pass filtering.



**Fig. 3** The processing chain to obtain covariance data.

Suppose that the bandpass signal  $\tilde{s}(t)$  is delayed by a time  $\tau$ . This can be written as

$$\tilde{s}_\tau(t) := \tilde{s}(t - \tau) = \text{real}\{s(t - \tau)e^{j2\pi f_c(t - \tau)}\} = \text{real}\{s(t - \tau)e^{-j2\pi f_c\tau}e^{j2\pi f_c t}\}.$$

The complex envelope of the delayed signal is thus  $s_\tau(t) = s(t - \tau)e^{-j2\pi f_c\tau}$ . Let  $W$  be the bandwidth of the complex envelope (the baseband signal) and let  $S(f)$  be its Fourier transform. We then have

$$s(t - \tau) = \int_{-W/2}^{W/2} S(f)e^{-j2\pi f\tau}e^{j2\pi ft}df \approx \int_{-W/2}^{W/2} S(f)e^{j2\pi ft}df = s(t)$$

where the approximation  $e^{-j2\pi f\tau} \approx 1$  is valid if  $|2\pi f\tau| \ll 1$  for all frequencies  $|f| \leq \frac{W}{2}$ . Ignoring a factor  $\pi$ , the resulting condition  $W\tau \ll 1$  is called the narrowband condition. Under this condition, we have for the complex envelope  $s_\tau(t)$  of the delayed bandpass signal  $\tilde{s}_\tau(t)$  that

$$s_\tau(t) \approx s(t)e^{-j2\pi f_c\tau} \quad \text{for } W\tau \ll 1.$$

The conclusion is that, for narrowband signals, time delays smaller than the inverse bandwidth may be represented as phase shifts of the complex envelope. Phased array processing heavily depends on this step. For radio astronomy, the maximal delay  $\tau$  is equal to the maximal geometric delay, which can be related to the diameter of the array. The bandwidth  $W$  is the bandwidth of each subband  $f_k$  in the RF processing chain that we discussed in the previous subsection.

### 3.3 Basic data model

We return to the radio astronomy context. For our purposes, it is convenient to model the sky as consisting of a collection of  $Q$  spatially discrete point sources, with  $s_q(n, k)$  the signal of the  $q$ th source at time sample  $n$  and frequency  $f_k$ .

In the simplest case, the signal received at the first antenna is a direct sum of these source signals, and the signal at the  $j$ th antenna is a sum of delayed signals, where the delays are geometric delays that depend on the direction under which each of the signals are observed. In the previous subsection, we saw that under the narrowband condition a delay of a narrowband signal  $s(t, k)$  by  $\tau$  can be represented by a phase shift:

$$s_\tau(t, k) = e^{-j2\pi f_k\tau} s(t, k)$$

which takes the form of a multiplication of  $s(t, k)$  by a complex number. Let  $[x_j, y_j, z_j]^T$  be the location of the  $j$ th antenna, with respect to the first

antenna. Further, let  $\mathbf{p}_q$  be a unit-length direction vector pointing into the direction of the  $q$ th source.

The geometrical delay  $\tau$  at antenna  $j$  for a signal coming from direction  $\mathbf{p}_q$  can be computed as follows. For a signal traveling directly from antenna 1 to antenna  $j$ , the delay is the distance between both antennas, divided by  $c$ , the speed of light. For any other direction, the delay depends on the cosine of the angle of incidence (compared to the baseline vector), and is thus described by the inner product of the location vector with the direction vector,

$$\tau = \frac{[x_j, y_j, z_j] \mathbf{p}_q}{c}.$$

Overall, the phase factor representing the geometric delay is

$$e^{-j2\pi f_k \tau} = e^{-j\mathbf{z}_j^T \mathbf{p}_q}, \quad \mathbf{z}_j = \frac{2\pi f_k}{c} \begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix}.$$

Here, we have introduced  $\mathbf{z}_j$  as a normalized location vector, assuming  $\mathbf{z}_1 = \mathbf{0}$  is the location of the first antenna (i.e., the phase reference). As the Earth rotates, the relative locations of the telescopes are also moving, hence  $\mathbf{z}_j$  is a function of sample time  $n$  (and of  $f_k$ , due to the normalization), and we write it as  $\mathbf{z}_j(n, k)$ . The coordinates of source direction vectors  $\mathbf{p}_q$  are expressed as<sup>1</sup>  $(\ell, m, n)$ , where  $\ell, m$  are direction cosines, and  $n = \sqrt{1 - \ell^2 - m^2}$  due to the normalization. There are several conventions and details regarding coordinate systems [37], but they are not of concern for us here.

Taking the phase factors into account, we can model the received signal vector  $\mathbf{x}(n, k)$  as

$$\mathbf{x}(n, k) = \sum_{q=1}^Q \mathbf{a}_q(n, k) s_q(n, k) + \mathbf{n}(n, k) \quad (8)$$

where  $\mathbf{a}_q(n, k)$  is called the ‘‘array response vector’’ for the  $q$ th source, consisting of the phase multiplication factors, and  $\mathbf{n}(n, k)$  is an additive noise vector, due to thermal noise at the receiver. We will model  $s_q(n, k)$  and  $\mathbf{n}(n, k)$  as baseband complex envelope representations of zero mean wide sense stationary white Gaussian random processes sampled at the Nyquist rate.

With the above discussion, the array response vector is modeled (for an ideal receiver) as

$$\mathbf{a}_q(n, k) = e^{-j\mathbf{Z}(n, k)^T \mathbf{p}_q}, \quad \mathbf{Z}(n, k) = [\mathbf{z}_1(n, k), \dots, \mathbf{z}_J(n, k)]. \quad (9)$$

For convenience of notation, we will in future usually drop the dependence on the frequency  $f_k$  (index  $k$ ) from the notation.

---

<sup>1</sup> with abuse of notation, as  $m, n$  are not related to the time variables used earlier.

Previously, in (6), we defined correlation estimates  $\hat{\mathbf{R}}_m$  as the output of the data acquisition process, where the time index  $m$  corresponds to the  $m$ th short term integration interval (STI), such that  $(m-1)N \leq n \leq mN$ . Due to Earth rotation, the vector  $\mathbf{a}_q(n)$  changes slowly with time, but we assume that within an STI it can be considered constant and can be represented, with some abuse of notation, by  $\mathbf{a}_q(m)$ . In that case,  $\mathbf{x}(n)$  is wide sense stationary over the STI, and a single STI autocovariance is defined as

$$\mathbf{R}_m = E\{\mathbf{x}(n)\mathbf{x}^H(n)\}, \quad m = \lceil \frac{n}{N} \rceil \quad (10)$$

where  $\mathbf{R}_m$  has size  $J \times J$ . Each element of  $\mathbf{R}_m$  represents the interferometric correlation along the baseline vector between the two corresponding receiving elements. It is estimated by STI sample covariance matrices  $\hat{\mathbf{R}}_m$  defined in (6), and our stationarity assumptions imply  $E\{\hat{\mathbf{R}}_m\} = \mathbf{R}_m$ .

If we consider only a single signal from direction  $\mathbf{p}_q$  and look at entry  $(i, j)$  of  $\mathbf{R}_m$ , then its value is

$$(\mathbf{R}_m)_{i,j} = E\{x_i(n)\bar{x}_j(n)\} = \sigma_q^2 e^{-j(\mathbf{z}_i(m) - \mathbf{z}_j(m))^T \mathbf{p}_q}. \quad (11)$$

The vector  $\mathbf{z}_i(m) - \mathbf{z}_j(m)$  is the *baseline*: the (normalized) vector pointing from telescope  $i$  to telescope  $j$ . In radio astronomy, it is usually expressed in coordinates denoted by  $(u, v, w)$ . The objective in telescope design is often to have as many different baselines as possible. In that case the entries of  $\mathbf{R}_m$  are different and non-redundant. As the Earth turns, the baselines also turn, thus giving rise to new baseline directions. We will see later that the set of baselines during an observation determines the spatial sampling function by which the incoming wave field is sampled, with important implications on the quality of the resulting image.

If we generalize now (11) to  $Q$  sources and add zero mean noise, uncorrelated from antenna to antenna, as in the signal model (8), we obtain what is known as the *measurement equation*, or covariance data model,

$$\mathbf{R}_m = \mathbf{A}_m \boldsymbol{\Sigma}_s \mathbf{A}_m^H + \boldsymbol{\Sigma}_n, \quad (12)$$

$$\begin{aligned} \text{where } \mathbf{A}_m &= [\mathbf{a}_1(m), \dots, \mathbf{a}_Q(m)] \\ \boldsymbol{\Sigma}_s &= \text{diag}\{[\sigma_1^2, \dots, \sigma_Q^2]\} \\ \boldsymbol{\Sigma}_n &= E\{\mathbf{n}(n)\mathbf{n}^H(n)\} = \text{diag}\{[\sigma_{n,1}^2, \dots, \sigma_{n,J}^2]\}. \end{aligned}$$

Here,  $\sigma_q^2 = E\{|s_q(n, k)|^2\}$  is the variance of the  $q$ th source,  $\boldsymbol{\Sigma}_s$  is the corresponding signal covariance matrix, and  $\boldsymbol{\Sigma}_n$  is the noise covariance matrix. (With abuse of notation, subscript  $n$  is now used to signify “noise”.) Noise is assumed to be independent but not evenly distributed across the array. The noise variances  $\sigma_{n,j}^2$  are considered unknown.

Under ideal circumstances, the array response matrix  $\mathbf{A}_m$  is just a phase matrix: its columns are given by the vectors  $\mathbf{a}_q(m)$  in (9), and its entries express the phase shifts due to the geometrical delays associated with the array and source geometry. We will later generalize this and introduce directional disturbances due to non-isotropic antennas, unequal antenna gains, and disturbances due to atmospheric effects.

### 3.4 Image formation for the ideal data model

Ignoring the additive noise and using the ideal array response matrix  $\mathbf{A}_m$ , the measurement equation (12), in its simplest form, can be written as

$$(\mathbf{R}_m)_{i,j} = \sum_{q=1}^Q I(\mathbf{p}_q) e^{-j(\mathbf{z}_i(m) - \mathbf{z}_j(m))^T \mathbf{p}_q} \quad (13)$$

where  $(\mathbf{R}_m)_{i,j}$  is the correlation between antennas  $i$  and  $j$  at STI interval  $m$ ,  $I(\mathbf{p}_q) = \sigma_q^2$  is the brightness (power) of the source in direction  $\mathbf{p}_q$ ,  $\mathbf{z}_i(m)$  is the normalized location vector of the  $i$ th antenna at STI  $m$ , and  $\mathbf{p}_q$  is the unit direction vector (position) of the  $q$ th source.

The function  $I(\mathbf{p})$  is the brightness image (or *map*) of interest: it is this function that is shown when we refer to a radio-astronomical image like figure 1. It is a function of the direction vector  $\mathbf{p}$ : this is a 3D vector, but due to its normalization it depends on only two parameters. We could e.g., show  $I(\cdot)$  as function of the direction cosines  $(\ell, m)$ , or of the corresponding angles.

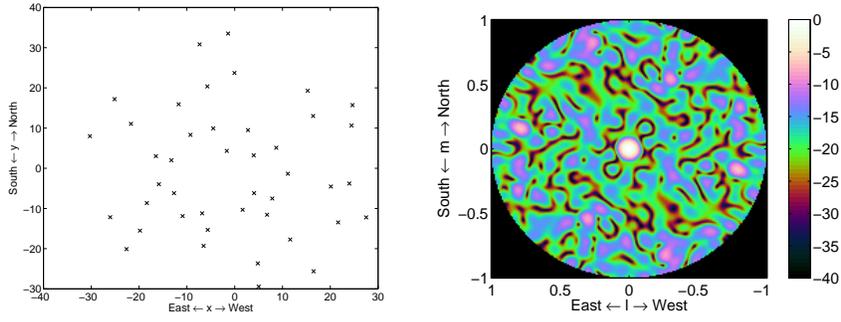
For our discrete point-source model, the brightness image is

$$I(\mathbf{p}) = \sum_{q=1}^Q \sigma_q^2 \delta(\mathbf{p} - \mathbf{p}_q) \quad (14)$$

where  $\delta(\cdot)$  is a Kronecker delta, and the direction vector  $\mathbf{p}$  is mapped to the location of “pixels” in the image (various transformations are possible). Only the pixels  $\mathbf{p}_q$  are nonzero, and have value equal to the source variance  $\sigma_q^2$ .

Equation (13) describes the relation between the visibility model and the desired image, and it has the form of a Fourier transform; it is known in radio astronomy as the Van Cittert-Zernike theorem [33, 37]. Image formation (*map making*) is essentially the inversion of this relation. Unfortunately, we have only a finite set of observations, therefore we can only obtain a *dirty image*: if we apply the inverse Fourier transformation to the measured correlation data, we obtain

$$\hat{I}_D(\mathbf{p}) := \sum_{i,j,m} (\hat{\mathbf{R}}_m)_{ij} e^{j(\mathbf{z}_i(m) - \mathbf{z}_j(m))^T \mathbf{p}}. \quad (15)$$



**Fig. 4** (a) Coordinates of the antennas in a LOFAR station, which defines the spatial sampling function, and (b) the resulting *dirty beam*.

In terms of the measurement data model (13), the “expected value” of the image is obtained by replacing  $\hat{\mathbf{R}}_m$  by  $\mathbf{R}_m$ , or

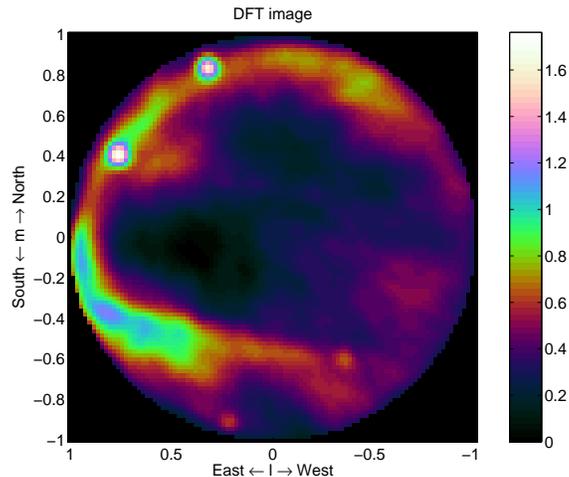
$$\begin{aligned}
 I_D(\mathbf{p}) &:= \sum_{i,j,m} (\mathbf{R}_m)_{i,j} e^{j(\mathbf{z}_i(m) - \mathbf{z}_j(m))^T \mathbf{p}} \\
 &= \sum_{i,j,m} \sum_q \sigma_q^2 e^{j(\mathbf{z}_i(m) - \mathbf{z}_j(m))^T (\mathbf{p} - \mathbf{p}_q)} \\
 &= \sum_q I(\mathbf{p}_q) B(\mathbf{p} - \mathbf{p}_q) \\
 &= I(\mathbf{p}) * B(\mathbf{p})
 \end{aligned} \tag{16}$$

where the *dirty beam* is given by

$$B(\mathbf{p}) := \sum_{i,j,m} e^{j(\mathbf{z}_i(m) - \mathbf{z}_j(m))^T \mathbf{p}}. \tag{17}$$

The dirty image  $I_D(\mathbf{p})$  is the desired image  $I(\mathbf{p})$  convolved with the dirty beam  $B(\mathbf{p})$ : every point source excites a beam  $B(\mathbf{p} - \mathbf{p}_q)$  centered at its location  $\mathbf{p}_q$ . Note that  $B(\mathbf{p})$  is a known function: it only depends on the locations of the telescopes, or rather the set of telescope baselines  $\mathbf{z}_i(m) - \mathbf{z}_j(m)$ .

An example of a set of antenna coordinates and the corresponding dirty beam is shown in figure 4. This is for a single low-band LOFAR station and a single STI and frequency bin. The dirty beam has heavy sidelobes as high as  $-10$  dB. A resulting dirty image is shown in figure 5. In this image, we see the complete sky, in  $(\ell, m)$  coordinates, where the reference direction is pointing towards zenith. The strong visible sources are Cassiopeia A and Cygnus A, also visible is the milky way, ending in the north polar spur (NPS) and, weaker, Virgo A. In the South, the Sun is visible as well. The image was



**Fig. 5** Dirty image following (16), using LOFAR station data.

obtained by averaging 25 STIs, each consisting of 10 s data in 25 frequency channels of 156 kHz wide taken from the band 45–67 MHz, avoiding the locally present radio interference. As this shows data from a single LOFAR station, with a relatively small maximal baseline (65 m), the resolution is limited and certainly not representative of the capabilities of the full LOFAR array.

The dirty beam is essentially a non-ideal point spread function due to finite and non-uniform spatial sampling: we only have a limited set of baselines. The dirty beam usually has a main lobe centered at  $\mathbf{p} = \mathbf{0}$ , and many side lobes. If we would have a large number of telescopes positioned in a uniform rectangular grid, the dirty beam would be a 2-D sinc-function (similar to a boxcar taper in time-domain sampling theory). The resulting beam size is inversely proportional to the aperture (diameter) of the array. This determines the *resolution* in the dirty image. The sidelobes of the beam give rise to confusion between sources: it is unclear whether a small peak in the image is caused by the main lobe of a weak source, or the sidelobe of a strong source. Therefore, attempts are made to design the array such that the sidelobes are low. It is also possible to introduce weighting coefficients (“tapers”) in (16) to obtain an acceptable beamshape.

Another aspect is the summation over  $m$  (STI intervals) in (17), where the rotation of the Earth is used to obtain essentially many more antenna baselines. The effect of this is that the sidelobes tend to get averaged out, to some extent. Many images are also formed by averaging over a small number of frequency bins (assuming the  $\sigma_q^2$  are constant over these frequency bins),

which enters into the equations in exactly the same way: Replace  $\mathbf{z}_i(m)$  by  $\mathbf{z}_i(m, k)$  and also sum over the frequency index  $k$ .

## 4 Deconvolution algorithms for image formation

*Deconvolution* is the process of recovering  $I(\cdot)$  from  $I_D(\cdot)$  using knowledge of the dirty beam, and thus to obtain the original high-resolution, “clean” image. A standard algorithm for doing this is CLEAN [15] and variants; however, many other algorithms are possible, depending on the underlying model assumptions and in a trade-off between accuracy and numerical complexity.

After a telescope has been designed and built, deconvolution is the most important step in image formation. It can increase the dynamic range (ratio between powers of the strongest and the weakest features in the image) by several orders of magnitude. However, the numerical complexity is often large, and high-resolution images require dedicated hardware solutions and sometimes even supercomputers. This section will describe some of the algorithms for deconvolution. Additional overviews are available in [8, 9, 19, 22], as well as in the books [37, 2].

### 4.1 The CLEAN algorithm

A popular method for deconvolution is the CLEAN algorithm [15]. From the dirty image  $I_D(\mathbf{p})$  and the known dirty beam  $B(\mathbf{p})$ , the desired image  $I(\mathbf{p})$  is obtained via a sequential Least Squares fitting method. The algorithm is based on the assumption that the sky is mostly empty, and consists of a set of discrete point sources. The brightest source is estimated first, its contribution is subtracted from the dirty image, then the next brightest source is subtracted, etc.

The algorithm further assumes that  $B(\mathbf{p})$  has its peak at the origin. Inside the loop, a candidate location  $\mathbf{p}_q$  is selected as the location of the largest peak in  $I_D(\mathbf{p})$ , the corresponding power  $\hat{\sigma}_q^2$  is estimated, and subsequently a small multiple of  $\hat{\sigma}_q^2 B(\mathbf{p} - \mathbf{p}_q)$  is subtracted from  $I_D(\mathbf{p})$ . The objective is to minimize the residual, until it converges to the noise level:

$$\begin{aligned}
 & q = 0 \\
 & \text{while } I_D(\mathbf{p}) \text{ is not noise-like:} \\
 & \quad \left[ \begin{array}{l} q = q + 1 \\ \mathbf{p}_q = \arg \max_{\mathbf{p}} I_D(\mathbf{p}) \\ \hat{\sigma}_q^2 = I_D(\mathbf{p}_q) / B(\mathbf{0}) \\ I_D(\mathbf{p}) := I_D(\mathbf{p}) - \gamma \hat{\sigma}_q^2 B(\mathbf{p} - \mathbf{p}_q), \quad \forall \mathbf{p} \\ I_{\text{clean}}(\mathbf{p}) = I_D(\mathbf{p}) + \sum_q \gamma \hat{\sigma}_q^2 B_{\text{synth}}(\mathbf{p} - \mathbf{p}_q), \quad \forall \mathbf{p}. \end{array} \right.
 \end{aligned}$$

The scaling parameter  $\gamma \leq 1$  is called the loop gain; for accurate convergence it should be small because the estimated location of the peak is at a grid point, whereas the true location of the peak may be in between grid points.  $B_{synth}(\mathbf{p})$  is a “synthetic beam”, usually a Gaussian bell-shape with about the same beam width as the main lobe of the dirty beam; it is introduced to mask the otherwise high artificial resolution of the image.

In current imaging systems, instead of the subtractions on the dirty image, it is considered more accurate to do the subtractions on the covariance data  $\mathbf{R}_m$  instead,

$$\hat{\mathbf{R}}_m := \hat{\mathbf{R}}_m - \gamma \hat{\sigma}_q^2 \mathbf{a}_q(m) \mathbf{a}_q(m)^H$$

and then to recompute the dirty image. For efficiency, usually a number of peaks are estimated from the dirty image together, the covariance is updated for this ensemble, and then the residual image is recomputed.

## 4.2 Imaging using a beamforming formulation

We will investigate some alternative deconvolution algorithms. For simplicity of notation, we assume from now on that only a single STI snapshot is used in the imaging, hence we also drop the time index  $m$  from the equations. The results can easily be extended.

The imaging process transforms the covariances of the received signals to an image of the source structure within the field-of-view of the receivers. In array processing terms, it can be described as follows [19]. Assume a data model as in (8), and recall the definition of the array response vector  $\mathbf{a}(\mathbf{p})$  in (9). There are  $J$  antennas. To determine the power of a signal arriving from a particular direction  $\mathbf{p}$ , a weight vector

$$\mathbf{w}(\mathbf{p}) = \frac{1}{J} \mathbf{a}(\mathbf{p}) = \frac{1}{J} e^{-j\mathbf{z}^T \mathbf{p}} \quad (18)$$

is applied to the array signal vector  $\mathbf{x}(n)$ . The operation  $y(n) = \mathbf{w}^H \mathbf{x}(n)$  is generally called beamforming. The choice  $\mathbf{w} = \mathbf{a}$  precisely compensates the geometric phase delays so that the antenna signals are added in phase. This can be regarded as a spatially matched filter, or *conjugate field match*. The (often omitted) scaling by  $1/J$  ensures the correct scaling of the output power. Indeed, the output power of a beamformer is, generally,

$$E\{|y|^2\} = \mathbf{w}^H E\{\mathbf{x}\mathbf{x}^H\} \mathbf{w} = \mathbf{w}^H \mathbf{R} \mathbf{w}.$$

For a data model consisting of a single source with power  $\sigma^2$  arriving from direction  $\mathbf{a}(\mathbf{p})$ , i.e.,  $\mathbf{x}(n) = \mathbf{a}(\mathbf{p})s(n)$ , we have, with  $\mathbf{w} = \frac{1}{J} \mathbf{a}(\mathbf{p})$ ,

$$E\{|y|^2\} = \mathbf{w}^H (\mathbf{a} \sigma^2 \mathbf{a}^H) \mathbf{w} = \sigma^2 \frac{\mathbf{a}^H \mathbf{a}}{J} \frac{\mathbf{a}^H \mathbf{a}}{J} = \sigma^2. \quad (19)$$

Thus, the matched beamformer corrects precisely the signal delays (phase shifts) present in  $\mathbf{a}(\mathbf{p})$ , when  $\mathbf{w}$  matches  $\mathbf{a}(\mathbf{p})$ , i.e. the beamformer is pointed into the same direction as the source. If the beamformer is pointed into other directions, the response is usually much smaller.

Using the beamformer to scan over all pixels  $\mathbf{p}$  in an image, we can create an image via beamforming as

$$\hat{I}_{BF}(\mathbf{p}) = \mathbf{w}(\mathbf{p})^H \hat{\mathbf{R}} \mathbf{w}(\mathbf{p}) \quad (20)$$

and the corresponding model for this image is

$$I_{BF}(\mathbf{p}) = \mathbf{w}(\mathbf{p})^H \mathbf{R} \mathbf{w}(\mathbf{p}). \quad (21)$$

The matched filter corresponds to weights  $\mathbf{w}(\mathbf{p})$  defined as in (18). Except for a factor  $J^2$ , the image  $I_{BF}(\mathbf{p})$  is for this choice identical to the dirty image  $I_D(\mathbf{p})$  defined in (16)! Indeed, starting from (16), we can write

$$I_D(\mathbf{p}) = \sum_{i,j} \mathbf{R}_{ij} e^{j(\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{p}} = \sum_{i,j} e^{j\mathbf{z}_i^T \mathbf{p}} \mathbf{R}_{ij} e^{-j\mathbf{z}_j^T \mathbf{p}} = \mathbf{a}(\mathbf{p})^H \mathbf{R} \mathbf{a}(\mathbf{p})$$

which is the beamforming image obtained using  $\mathbf{w}(\mathbf{p}) = \mathbf{a}(\mathbf{p})$ . The response to a single source at the origin is

$$\begin{aligned} B(\mathbf{p}) &= \mathbf{a}(\mathbf{p})^H \mathbf{a}(\mathbf{0}) \mathbf{a}(\mathbf{0})^H \mathbf{a}(\mathbf{p}) \\ &= \mathbf{a}(\mathbf{p})^H \mathbf{1} \mathbf{1}^H \mathbf{a}(\mathbf{p}) \\ &= \mathbf{1}^H [\mathbf{a}(\mathbf{p}) \mathbf{a}(\mathbf{p})^H] \mathbf{1} \\ &= \sum_{i,j} e^{-j(\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{p}} \end{aligned}$$

which is the dirty beam defined in (17), now written in beamforming notation. It typically has a spike at  $\mathbf{p} = \mathbf{0}$ , and many sidelobes, depending on the spatial sampling function. We have already seen that these sidelobes limit the resolution, as they can be confused with (or mask) other sources.

So far, we looked at the response to a source, but ignored the effect of the noise on an image. In the beamforming formulation, the response to a data set which only consists of noise, or  $\mathbf{R} = \mathbf{R}_n$  is

$$I(\mathbf{p}) = \mathbf{w}(\mathbf{p})^H \mathbf{R}_n \mathbf{w}(\mathbf{p}).$$

Suppose that the noise is spatially white,  $\mathbf{R}_n = \sigma_n^2 \mathbf{I}$ , and that we use the matched beamformer (18), we obtain

$$I(\mathbf{p}) = \sigma_n^2 \frac{\mathbf{a}(\mathbf{p})^H}{J} \frac{\mathbf{a}(\mathbf{p})}{J} = \sigma_n^2 \frac{\|\mathbf{a}(\mathbf{p})\|^2}{J^2} = \frac{\sigma_n^2}{J} \quad (22)$$

since all entries of  $\mathbf{a}(\mathbf{p})$  have unit magnitude. As this is a constant, the image will be “flat”. For a general data set  $\mathbf{R} = \mathbf{R}_s + \mathbf{R}_n$ , the responses to the sources and to the noise will be added. Comparing (19) to (22), we see that the noise is suppressed by a factor  $J$  compared to a point source signal coming from a specific direction. This is the *array gain*. If we use multiple STIs and/or frequencies  $f_k$ , the array gain can be larger than  $J$ .

### 4.3 Imaging via adaptive beamforming: MVDR and AAR

Now that we have made the connection of imaging to beamforming, we can apply a range of other beamforming techniques instead of the matched filter, such as the class of spatially adaptive beamformers. In fact, these can be considered as 2D spatial-domain versions of (now classical) spectrum estimation techniques for estimating the power spectral density of a random process (viz. [14]), and the general idea is that we can obtain a higher resolution if the sidelobes generated by strong sources are made small.

As an example, the “minimum variance distortionless response” (MVDR) beamformer is defined such that, for pixel  $\mathbf{p}$ , the response towards the direction of interest  $\mathbf{p}$  is unity, but signals from other directions are suppressed as much as possible, i.e.,

$$\mathbf{w}(\mathbf{p}) = \arg \min_{\mathbf{w}} \mathbf{w}^H \mathbf{R} \mathbf{w}, \quad \text{such that } \mathbf{w}^H \mathbf{a}(\mathbf{p}) = 1.$$

This problem can be solved in various ways. For example, after making a transformation  $\mathbf{w}' := \mathbf{R}^{1/2} \mathbf{w}$ ,  $\mathbf{a}' := \mathbf{R}^{-1/2} \mathbf{a}$ , the problem becomes

$$\mathbf{w}'(\mathbf{p}) = \arg \min_{\mathbf{w}'} \|\mathbf{w}'\|^2, \quad \text{such that } \mathbf{w}'^H \mathbf{a}'(\mathbf{p}) = 1.$$

To minimize the norm of  $\mathbf{w}'$ , it should be aligned to  $\mathbf{a}'$ , i.e.,  $\mathbf{w}' = \alpha \mathbf{a}'$ , and the solution is  $\mathbf{w}' = \mathbf{a}' / (\mathbf{a}'^H \mathbf{a}')$ . In terms of the original variables, the solution is then

$$\mathbf{w}(\mathbf{p}) = \frac{\mathbf{R}^{-1} \mathbf{a}(\mathbf{p})}{\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-1} \mathbf{a}(\mathbf{p})},$$

and the resulting image can thus be described as

$$I_{MVDR}(\mathbf{p}) = \mathbf{w}(\mathbf{p})^H \mathbf{R} \mathbf{w}(\mathbf{p}) = \frac{1}{\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-1} \mathbf{a}(\mathbf{p})}.$$

For a point-source model, this image will have a high resolution: two sources that are closely spaced will be resolved. The corresponding beam responses to different sources will in general be different: the beamshape is spatially varying.

The MVDR image is to be used instead of the dirty image  $I_D(\mathbf{p})$  in the CLEAN loop. Due to its high resolution, the location of sources is better estimated than using the original dirty image (and the location estimate can be further improved by searching for the true peak on a smaller grid in the vicinity of the location of the maximum). A second modification to the CLEAN loop is also helpful: Suppose that the location of the brightest source is  $\mathbf{p}_q$ , then the corresponding power  $\alpha_q$  should be estimated by minimizing the residual  $\|\mathbf{R} - \alpha\mathbf{a}(\mathbf{p}_q)\mathbf{a}(\mathbf{p}_q)^H\|^2$ . This can be done in closed form: using (5) we find

$$\|\mathbf{R} - \alpha\mathbf{a}(\mathbf{p}_q)\mathbf{a}(\mathbf{p}_q)^H\| = \|\text{vec}(\mathbf{R}) - \alpha[\bar{\mathbf{a}}(\mathbf{p}_q) \otimes \mathbf{a}(\mathbf{p}_q)]\|.$$

The optimal least squares solution for  $\alpha$  is, using (1), (3) and (2) in turn,

$$\begin{aligned} \alpha_q &= [\bar{\mathbf{a}}(\mathbf{p}_q) \otimes \mathbf{a}(\mathbf{p}_q)]^\dagger \text{vec}(\mathbf{R}) \\ &= \frac{[\bar{\mathbf{a}}(\mathbf{p}_q) \otimes \mathbf{a}(\mathbf{p}_q)]^H \text{vec}(\mathbf{R})}{[\bar{\mathbf{a}}(\mathbf{p}_q) \otimes \mathbf{a}(\mathbf{p}_q)]^H [\bar{\mathbf{a}}(\mathbf{p}_q) \otimes \mathbf{a}(\mathbf{p}_q)]} \\ &= \frac{\mathbf{a}(\mathbf{p}_q)^H \mathbf{R} \mathbf{a}(\mathbf{p}_q)}{[\mathbf{a}(\mathbf{p}_q)^H \mathbf{a}(\mathbf{p}_q)]^2} \\ &= \frac{\mathbf{a}(\mathbf{p}_q)^H \mathbf{R} \mathbf{a}(\mathbf{p}_q)}{J^2} \end{aligned}$$

which is the power estimate of the matched filter. In the CLEAN loop,  $\mathbf{R}$  should be replaced by its estimate  $\hat{\mathbf{R}}$  minus the estimated components until  $q$ , and also a constraint that  $\alpha_q$  is to be positive should be included. This method and a number of refinements are proposed in [1].

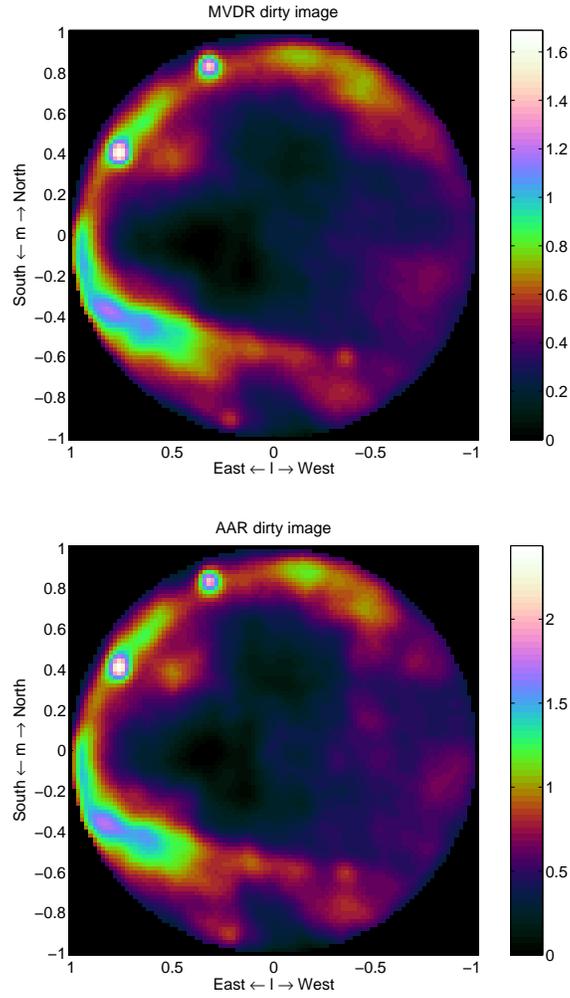
A problem with the MVDR image and other adaptive beamformers is that the output noise power is not spatially uniform. Consider the data model  $\mathbf{R} = \mathbf{A}\Sigma_s\mathbf{A}^H + \Sigma_n$ , where  $\Sigma_n = \sigma_n^2\mathbf{I}$  is the noise covariance matrix, then at the output of the beamformer the noise power is

$$\begin{aligned} \sigma_y^2(\mathbf{p}) &= \mathbf{w}(\mathbf{p})^H \mathbf{R}_n \mathbf{w}(\mathbf{p}) \\ &= \frac{\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-1} (\sigma_n^2 \mathbf{I}) \mathbf{R}^{-1} \mathbf{a}(\mathbf{p})}{[\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-1} \mathbf{a}(\mathbf{p})]^2} \\ &= \sigma_n^2 \frac{\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-2} \mathbf{a}(\mathbf{p})}{[\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-1} \mathbf{a}(\mathbf{p})]^2}. \end{aligned}$$

Thus, the output noise power is direction dependent.

As a remedy to this, a related beamformer which satisfies the constraint  $\mathbf{w}(\mathbf{p})^H \mathbf{w}(\mathbf{p}) = 1$  (and therefore has spatially uniform output noise) is obtained by using a different scaling of the MVDR beamformer:

$$\mathbf{w}(\mathbf{p}) = \mu \mathbf{R}^{-1} \mathbf{a}(\mathbf{p}), \quad \mu = \frac{1}{\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-2} \mathbf{a}(\mathbf{p})}.$$



**Fig. 6** Dirty images corresponding to the (a) MVDR and (b) AAR beamformers.

This beamformer is known as the “Adapted Angular Response” (AAR) [6]. The resulting image is

$$I_{AAR}(\mathbf{p}) = \mathbf{w}(\mathbf{p})^H \mathbf{R} \mathbf{w}(\mathbf{p}) = \frac{\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-1} \mathbf{a}(\mathbf{p})}{[\mathbf{a}(\mathbf{p})^H \mathbf{R}^{-2} \mathbf{a}(\mathbf{p})]^2}.$$

It has a high resolution and suppresses sidelobe interference under the white noise constraint. It was proposed for use in the CLEAN loop in [1], the resulting CLEANed image was called LS-MVI.

Example MVDR and AAR dirty images using the same LOFAR station as before are shown in figure 6. At first sight, the performance of these images is quite similar to that of the original dirty image in figure 5. There may be situations where the differences are more pronounced (see [1] for examples), e.g., the resolution for closely spaced point sources is expected to be significantly improved.

## 5 Least Squares imaging

In the previous section, we discussed various algorithms based on the CLEAN algorithm. This algorithm uses a successive approximation of the dirty image using a point source model. In this section, we take a model-based approach. The imaging problem is formulated as a parametric estimation problem where certain parameters (source locations, powers, noise variance) are unknown and need to be estimated. Although we start from a Maximum Likelihood formulation, we will quickly arrive at a more feasible Least Squares approach. The discussion follows to some extent [31], which is a general array processing approach to a very similar problem and can be read for further details.

### 5.1 Matrix formulation of the data model

Let us start again from the data model (8). For simplicity, we consider only a single frequency bin and STI interval, but all results can be generalized straightforwardly. The model for the signals arriving at the antenna array is thus

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \mathbf{n}(n)$$

and the covariance of  $\mathbf{x}$  is (viz. (12))

$$\mathbf{R} = \mathbf{A}\boldsymbol{\Sigma}_s\mathbf{A}^H + \boldsymbol{\Sigma}_n.$$

We have available a sample covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N} \sum_n \mathbf{x}(n)\mathbf{x}(n)^H$$

which serves as the input data for the imaging step. Let us now vectorize this data model by defining

$$\hat{\mathbf{r}} = \text{vec}(\hat{\mathbf{R}}), \quad \mathbf{r} = \text{vec}(\mathbf{R})$$

where  $\mathbf{r}$  has the data model (using (4))

$$\mathbf{r} = (\bar{\mathbf{A}} \circ \mathbf{A})\boldsymbol{\sigma}_s + \text{vec}(\boldsymbol{\Sigma}_n).$$

If  $\boldsymbol{\Sigma}_n$  is diagonal, we can write  $\text{vec}(\boldsymbol{\Sigma}_n) = (\mathbf{I} \circ \mathbf{I})\boldsymbol{\sigma}_n$ , where  $\boldsymbol{\sigma}_n$  is a vector containing the diagonal entries of  $\boldsymbol{\Sigma}_n$ . Define  $\mathbf{M}_s = \bar{\mathbf{A}} \circ \mathbf{A}$  and  $\mathbf{M}_n = \mathbf{I} \circ \mathbf{I}$ . Then

$$\mathbf{r} = \mathbf{M}_s\boldsymbol{\sigma}_s + \mathbf{M}_n\boldsymbol{\sigma}_n = [\mathbf{M}_s \ \mathbf{M}_n] \begin{bmatrix} \boldsymbol{\sigma}_s \\ \boldsymbol{\sigma}_n \end{bmatrix} = \mathbf{M}\boldsymbol{\sigma}.$$

In this formulation, several modifications can be introduced. E.g., a nondiagonal noise covariance matrix  $\boldsymbol{\Sigma}_n$  will lead to a more general  $\mathbf{M}_n$ , and if  $\boldsymbol{\Sigma}_n = \sigma_n^2\mathbf{I}$ , we have  $\mathbf{M}_n = \text{vec}(\mathbf{I})$  and  $\boldsymbol{\sigma}_n = \sigma_n^2$ . Some other options are discussed in [31].

We can further write

$$\hat{\mathbf{r}} = \mathbf{r} + \mathbf{w} = \mathbf{M}\boldsymbol{\sigma} + \mathbf{w}, \quad (23)$$

where  $\hat{\mathbf{r}}$  is the available ‘‘measurement data’’,  $\mathbf{r}$  is its mean (expected value), and  $\mathbf{w}$  is zero mean additive noise. It is not hard to derive that (for Gaussian signals) the covariance of this noise is [31]

$$\mathbf{C}_w = E(\hat{\mathbf{r}} - \mathbf{r})(\hat{\mathbf{r}} - \mathbf{r})^H = \frac{1}{N}(\bar{\mathbf{R}} \otimes \mathbf{R})$$

where  $N$  is the number of samples on which  $\hat{\mathbf{R}}$  is based. We have thus written our original data model on  $\mathbf{x}$  as a similar data model on  $\hat{\mathbf{r}}$ . Many estimation techniques from the literature that are usually applied to data models for  $\mathbf{x}$  can be applied to the data model for  $\mathbf{r}$ . Furthermore, it is straightforward to extend this vectorized formulation to include multiple snapshots over time and frequency to increase the amount of measurement data and thus to improve the imaging result: Simply stack the covariance data in  $\hat{\mathbf{r}}$  and include the model structure in  $\mathbf{M}$ ; note that  $\boldsymbol{\sigma}$  remains unchanged.

The unknown parameters in the data model are, first of all, the powers  $\boldsymbol{\sigma}$ . These appear linear in the model. Regarding the positions of the sources, we can consider two cases:

1. We consider a point source model with a ‘‘small’’ number of sources. In that case,  $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$  and  $\mathbf{M} = \mathbf{M}(\boldsymbol{\theta})$ , where  $\boldsymbol{\theta}$  is some parametrization of the unknown locations of the sources (the position vectors  $\mathbf{p}_q$  for each source). These enter in a nonlinear way into the model  $\mathbf{M}(\boldsymbol{\theta})$ . The image  $I(\mathbf{p})$  is constructed following (14), usually convolved with a synthetic beam  $B_{synth}(\mathbf{p})$  to make the image look nicer.
2. Alternatively, we consider a model where for each pixel in the image, we assume a corresponding point source: the source positions  $\mathbf{p}_q$  directly correspond to the pixels in the image. This can lead to a large number of sources. With the locations of the pixels predetermined,  $\mathbf{M}$  is a priori known and not a function of  $\boldsymbol{\theta}$ , but  $\mathbf{M}$  will have many columns (one for

each pixel-source). The image  $I(\mathbf{p})$  has a one-to-one relation to the source power vector  $\boldsymbol{\sigma}_s$ , we can thus regard  $\boldsymbol{\sigma}_s$  as the image in this case.

We need to pose several requirements on  $\mathbf{M}$  or  $\mathbf{M}(\boldsymbol{\theta})$  to ensure identifiability. First of all, in the first case we must have  $\mathbf{M}(\boldsymbol{\theta}) = \mathbf{M}(\boldsymbol{\theta}') \rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}'$ , otherwise we cannot uniquely find  $\boldsymbol{\theta}$  from  $\mathbf{M}$ . Furthermore, for both cases we will require that  $\mathbf{M}$  is a tall matrix (more rows than columns) and has full column rank, so that it has a left inverse (this will allow to estimate  $\boldsymbol{\sigma}$ ). This puts a limit on the number of sources in the image (number of columns of  $\mathbf{M}$ ) in relation to the number of observations (rows). If more snapshots (STIs) and/or multiple frequencies are available, as is the case in practice, then  $\mathbf{M}$  will become taller, and more sources can be estimated thus increasing the resolution. If  $\mathbf{M}$  is not tall, then there are some ways to generalize this, e.g. via the context of compressive sampling where we can have  $\mathbf{M}$  wide as long as  $\boldsymbol{\sigma}$  is sparse [43], which we will briefly discuss in subsection 5.6.

For the moment, we will continue with the second formulation: one source per pixel, fewer pixels than available correlation data.

## 5.2 Matrix formulation of imaging via beamforming

Let us now again interpret the “beamforming image” (20) as a linear transformation on the covariance data  $\hat{\mathbf{r}}$ . We can stack all image values  $I(\mathbf{p})$  over all pixels  $\mathbf{p}$  into a single vector  $\hat{\mathbf{i}}$ , and similarly, we can collect the weights  $\mathbf{w}(\mathbf{p})$  over all pixels into a single matrix  $\mathbf{W} = [\mathbf{w}(\mathbf{p}_1), \mathbf{w}(\mathbf{p}_2), \dots]$ . From (3), we know that  $\mathbf{w}^H \mathbf{R} \mathbf{w} = (\overline{\mathbf{w}} \otimes \mathbf{w})^H \text{vec}(\hat{\mathbf{R}})$ , so that we can write

$$\hat{\mathbf{i}}_{BF} = (\overline{\mathbf{W}} \circ \mathbf{W})^H \hat{\mathbf{r}}. \quad (24)$$

We saw before that the dirty image is obtained if we use the matched filter. In this case, we have  $\mathbf{W} = \frac{1}{J} \mathbf{A}$ , where  $\mathbf{A}$  contains the array response vectors  $\mathbf{a}(\mathbf{p})$  for every pixel  $\mathbf{p}$  of interest. In this case, the image is

$$\hat{\mathbf{i}}_D = \frac{1}{J^2} (\overline{\mathbf{A}} \circ \mathbf{A})^H \hat{\mathbf{r}} = \frac{1}{J^2} \mathbf{M}_s^H \hat{\mathbf{r}}. \quad (25)$$

The expected value of the image is obtained by using  $\mathbf{r} = \mathbf{M}\boldsymbol{\sigma}$ :

$$\mathbf{i}_D = \frac{1}{J^2} \mathbf{M}_s^H \mathbf{M} \boldsymbol{\sigma} = \frac{1}{J^2} (\mathbf{M}_s^H \mathbf{M}_s) \boldsymbol{\sigma}_s + \frac{1}{J^2} (\mathbf{M}_s^H \mathbf{M}_n) \boldsymbol{\sigma}_n.$$

The quality or “performance” of the image, or how close  $\hat{\mathbf{i}}_D$  is to  $\mathbf{i}_D$ , is related to its covariance,

$$\text{cov}(\hat{\mathbf{i}}_D) = E\{(\hat{\mathbf{i}}_D - \mathbf{i}_D)(\hat{\mathbf{i}}_D - \mathbf{i}_D)^H\} = \frac{1}{J^4} \mathbf{M}_s^H \mathbf{C}_w \mathbf{M}_s$$

where  $\mathbf{C}_w = \frac{1}{N}(\bar{\mathbf{R}} \otimes \mathbf{R})$  is the covariance of the noise on the covariance data. Since usually the astronomical sources are much weaker than the noise (often at least by a factor 100), we can approximate  $\mathbf{R} \approx \boldsymbol{\Sigma}_n$ . If the noise is spatially white,  $\boldsymbol{\Sigma}_n = \sigma_n^2 \mathbf{I}$ , we obtain for the covariance of  $\hat{\mathbf{i}}_D$

$$\text{cov}(\hat{\mathbf{i}}_D) \approx \frac{\sigma_n^4}{J^4 N} \mathbf{M}_s^H \mathbf{M}_s.$$

The variance in the image is given by the diagonal of this expression. From this and the structure of  $\mathbf{M}_s = (\bar{\mathbf{A}} \circ \mathbf{A})$  and the structure of  $\mathbf{A}$ , we can see that the variance on each pixel in the dirty image is constant,  $\sigma_n^4/(J^2 N)$ , but that the noise on the image is correlated, possibly leading to visible structures in the image. This is a general phenomenon.

Similar equations can be derived for the MVDR image and the AAR image.

### 5.3 Weighted Least Squares imaging

At this point, the deconvolution problem can be formulated as a maximum likelihood (ML) estimation problem, and solving this problem should provide a statistically efficient estimate of the parameters. Since all signals are assumed to be i.i.d. Gaussian signals, the derivation is standard and the ML estimates are obtained by minimizing the negative log-likelihood function [31]

$$\{\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\theta}}\} = \arg \min_{\boldsymbol{\sigma}, \boldsymbol{\theta}} \left( \ln |\mathbf{R}(\boldsymbol{\sigma}, \boldsymbol{\theta})| + \text{tr} \left( \mathbf{R}^{-1}(\boldsymbol{\sigma}, \boldsymbol{\theta}) \hat{\mathbf{R}} \right) \right). \quad (26)$$

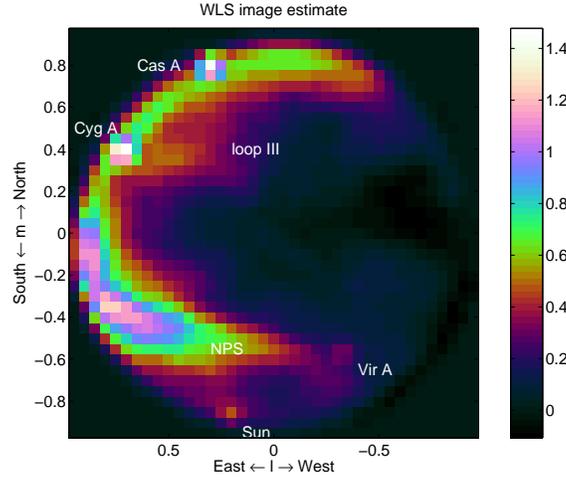
where  $|\cdot|$  denotes the determinant.  $\mathbf{R}(\boldsymbol{\sigma}, \boldsymbol{\theta})$  is the model, i.e.,  $\text{vec}(\mathbf{R}(\boldsymbol{\sigma}, \boldsymbol{\theta})) = \mathbf{r} = \mathbf{M}(\boldsymbol{\theta})\boldsymbol{\sigma}$ .

In this subsection, we will consider the overparametrized case, where each pixel in the image corresponds to a source. In this case,  $\mathbf{M}$  is a priori known, the model is linear, and the ML problem reduces to a Weighted Least Squares (WLS) problem to match  $\hat{\mathbf{r}}$  to the model  $\mathbf{r}$ :

$$\hat{\boldsymbol{\sigma}} = \arg \min_{\boldsymbol{\sigma}} \|\mathbf{C}_w^{-1/2}(\hat{\mathbf{r}} - \mathbf{r})\|_2^2 = \arg \min_{\boldsymbol{\sigma}} (\hat{\mathbf{r}} - \mathbf{M}\boldsymbol{\sigma})^H \mathbf{C}_w^{-1} (\hat{\mathbf{r}} - \mathbf{M}\boldsymbol{\sigma}) \quad (27)$$

where we fit the “data”  $\hat{\mathbf{r}}$  to the model  $\mathbf{r} = \mathbf{M}\boldsymbol{\sigma}$ . The correct weighting is the inverse of the covariance of the residual,  $\mathbf{w} = \hat{\mathbf{r}} - \mathbf{r}$ , i.e., the noise covariance matrix  $\mathbf{C}_w = \frac{1}{N}(\bar{\mathbf{R}} \otimes \mathbf{R})$ . For this, we may also use the estimate  $\hat{\mathbf{C}}_w$  obtained by using  $\hat{\mathbf{R}}$  instead of  $\mathbf{R}$ . Using the assumption that the astronomical sources are much weaker than the noise we could contemplate to use  $\mathbf{R} \approx \boldsymbol{\Sigma}_n$  for the weighting. If the noise is spatially white,  $\boldsymbol{\Sigma}_n = \sigma_n^2 \mathbf{I}$ , the weighting can then even be omitted.

The solution of (27) is obtained by applying the pseudo-inverse,



**Fig. 7** Image corresponding to the WLS formulation (28).

$$\hat{\boldsymbol{\sigma}} = [\mathbf{C}_w^{-1/2} \mathbf{M}]^\dagger \mathbf{C}_w^{-1/2} \hat{\mathbf{r}} = (\mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M})^{-1} \mathbf{M}^H \mathbf{C}_w^{-1} \hat{\mathbf{r}} =: \mathbf{M}_d^{-1} \hat{\boldsymbol{\sigma}}_d \quad (28)$$

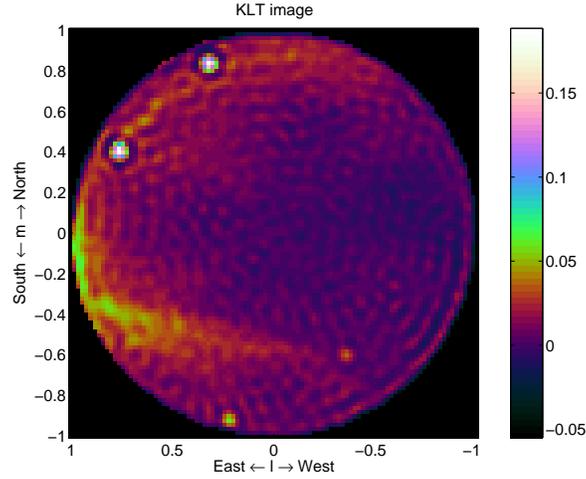
where

$$\mathbf{M}_d := \mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M}, \quad \hat{\boldsymbol{\sigma}}_d := \mathbf{M}^H \mathbf{C}_w^{-1} \hat{\mathbf{r}}.$$

Here, we can consider the term  $\hat{\boldsymbol{\sigma}}_d = \mathbf{M}^H \mathbf{C}_w^{-1} \hat{\mathbf{r}}$  as a “dirty image”: it is comparable to (25), although we have introduced a weighting by  $\mathbf{C}_w^{-1}$  and estimate the noise covariance parameters  $\boldsymbol{\sigma}_n$  as well as the source powers in  $\boldsymbol{\sigma}_s$  (the actual image). The factor  $1/J^2$  in (25) can be seen as a crude approximation of  $\mathbf{M}_d^{-1}$ .

Figure 7 shows an example WLS image for the same LOFAR data set as before. The resolution (number of pixels) in this image is kept limited (about 1000) for reasons discussed below.

The term  $\mathbf{M}_d^{-1} = (\mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M})^{-1}$  is a deconvolution operation. This inversion can only be carried out if the deconvolution matrix  $\mathbf{M}_d = \mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M}$  is not rank deficient. This requires at least that  $\mathbf{M}$  is a tall matrix (“less pixels than observations” in case we take one source per pixel). Thus, high resolution WLS imaging is only possible if a limited number of sources is present. The condition number of  $\mathbf{M}_d$ , i.e., the ratio of the largest to the smallest eigenvalue of  $\mathbf{M}_d$ , gives important information on our ability to compute its inverse: LS theory tells us that the noise on  $\hat{\boldsymbol{\sigma}}_d$  could, in the worst case, be magnified by this factor. The optimal (smallest) condition number of any matrix is 1, which is achieved if  $\mathbf{M}_d$  is a scaling of the identity matrix, or if the columns of  $\mathbf{C}_w^{-1/2} \mathbf{M}$  are all orthogonal to each other. If the size of  $\mathbf{M}$  becomes less tall, then the condition number of  $\mathbf{M}_d$  becomes larger (worse),



**Fig. 8** Image corresponding to the KLT solution (29).

and once it is a wide matrix,  $\mathbf{M}$  is singular and the condition number will be infinite. Thus, we have a trade-off between the resolution (number of pixels in the image) and the noise enhancement.

The definition of  $\mathbf{M}_d$  shows that it is not data dependent, and it can be precomputed for a given telescope configuration and observation interval. It is thus possible to explore this trade-off beforehand. To avoid numerical instabilities (noise enhancement), we would usually compute a regularized inverse or pseudo-inverse of this matrix, e.g., by first computing the eigenvalue decomposition

$$\mathbf{M}_d = \mathbf{U}\mathbf{A}\mathbf{U}^H$$

where  $\mathbf{U}$  contains the (orthonormal) eigenvectors and  $\mathbf{A}$  is a diagonal matrix containing the eigenvalues, sorted from large to small. Given a threshold  $\epsilon$  on the eigenvalues, we can define  $\tilde{\mathbf{A}}$  to be a diagonal matrix containing only the eigenvalues larger than  $\epsilon$ , and  $\tilde{\mathbf{U}}$  a matrix containing the corresponding eigenvectors. The  $\epsilon$ -threshold pseudo-inverse is then given by

$$\mathbf{M}_d^\dagger := \tilde{\mathbf{U}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{U}}^H$$

and the resulting image is

$$\boldsymbol{\sigma} = \tilde{\mathbf{U}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{U}}^H\boldsymbol{\sigma}_d. \quad (29)$$

This can be called the “Karhunen-Loève” image, as the rank reduction is related to the Karhunen-Loève transform (KLT). It corresponds to selecting

an optimal (Least Squares) set of basis vectors on which to project a certain data set, here  $\boldsymbol{\sigma}_d$ .

An example KLT image is shown in figure 8. In this image, the number of pixels is much larger than before in figure 7 (about 9000), but the rank of the matrix  $\mathbf{M}_d$  is truncated at 1/200 times the largest eigenvalue, leaving about 1300 out of 9000 image components. The result is not quite satisfactory: the truncation to a reduced basis results in annoying ripple artifacts in the image.

Computing the eigenvalue decomposition for large matrices is complex. A computationally simpler alternative is to compute a regularized inverse of  $\mathbf{M}_d$ , i.e., to take the inverse of  $\mathbf{M}_d + \epsilon \mathbf{I}$ . This should yield similar (although not identical) results.

If we use the alternative sky model where we assume a point source model with a “small” number of sources ( $\mathbf{M} = \mathbf{M}(\boldsymbol{\theta})$ ), then the conditioning of  $\mathbf{M}_d$ , and thus the performance of the deconvolution, is directly related to this number of sources.

The performance of the method is assessed by looking at the covariance of the resulting image (plus noise parameters)  $\boldsymbol{\sigma}$  in (28). It is given by

$$\begin{aligned} \mathbf{C}\boldsymbol{\sigma} &= (\mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M})^{-1} \mathbf{M}^H \mathbf{C}_w^{-1} (\mathbf{C}_w) \mathbf{C}_w^{-1} \mathbf{M} (\mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M})^{-1} \\ &= (\mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M})^{-1} = \mathbf{M}_d^{-1}. \end{aligned}$$

This again shows that the performance of the imaging method follows directly from the conditioning of the deconvolution matrix  $\mathbf{M}_d$ . If  $\mathbf{M}_d$  is sufficiently well conditioned, the noise on the image is limited, otherwise it may be large. The formulation also shows that the pixels in the image are correlated ( $\mathbf{M}_d$  is in general not diagonal), as we obtained before for the dirty image.

Similarly, if we use the pseudo-inverse  $\mathbf{M}_d^\dagger = \tilde{\mathbf{U}} \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{U}}^H$  for the deconvolution, then we obtain  $\mathbf{C}\boldsymbol{\sigma} = \mathbf{M}_d^\dagger$ . In this case, the noise enhancement depends on the chosen threshold  $\epsilon$ . Also, the rank of  $\mathbf{C}\boldsymbol{\sigma}$  depends on this threshold, and since it is not full rank, the number of independent components (sources) in the image is smaller than the number of shown pixels: the rank reduction defines a form of interpolation.

Using a rank truncation for radio astronomy imaging was already suggested in [7]. Unfortunately, if the number of pixels is large, this technique by itself is not sufficient to obtain good images, e.g., the resulting pixels may not all be positive, which is unphysical for an intensity image. Thus, the overparametrized case requires additional constraints; some options are discussed in subsection 5.6.

#### 5.4 Estimating the position of the sources

Let us now consider the use of the alternative formulation, where we write  $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$  and  $\mathbf{M} = \mathbf{M}(\boldsymbol{\theta})$ , where  $\boldsymbol{\theta}$  captures the positions of the limited

number of sources in the image. In this case, we have to estimate both  $\boldsymbol{\sigma}$  and  $\boldsymbol{\theta}$ . If we start again from the ML formulation (26), it does not seem feasible to solve this minimization problem in closed form. However, we can again resort to the WLS covariance matching problem and solve instead

$$\begin{aligned} \{\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\theta}}\} &= \arg \min_{\boldsymbol{\sigma}, \boldsymbol{\theta}} \|\mathbf{C}_w^{-1/2}(\hat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\sigma}, \boldsymbol{\theta}))\|^2 \\ &= \arg \min_{\boldsymbol{\sigma}, \boldsymbol{\theta}} (\hat{\mathbf{r}} - \mathbf{M}(\boldsymbol{\theta})\boldsymbol{\sigma})^H \mathbf{C}_w^{-1} (\hat{\mathbf{r}} - \mathbf{M}(\boldsymbol{\theta})\boldsymbol{\sigma}). \end{aligned} \quad (30)$$

It is known that the resulting estimates are, for a large number of samples, equivalent to ML estimates and therefore asymptotically efficient [31].

The WLS problem is separable: suppose that the optimal  $\boldsymbol{\theta}$  is known, so that  $\mathbf{M} = \mathbf{M}(\boldsymbol{\theta})$  is known, then the corresponding  $\boldsymbol{\sigma}$  will satisfy the solution which we found earlier:

$$\hat{\boldsymbol{\sigma}} = (\mathbf{M}^H \mathbf{C}_w^{-1} \mathbf{M})^{-1} \mathbf{M}^H \mathbf{C}_w^{-1} \hat{\mathbf{r}}.$$

Substituting this solution back into the problem, we obtain

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \min_{\boldsymbol{\theta}} \hat{\mathbf{r}}^H [\mathbf{I} - \mathbf{M}(\boldsymbol{\theta})(\mathbf{M}(\boldsymbol{\theta})^H \mathbf{C}_w^{-1} \mathbf{M}(\boldsymbol{\theta}))^{-1} \mathbf{M}(\boldsymbol{\theta})^H \mathbf{C}_w^{-1}]^H \cdot \\ &\quad \cdot \mathbf{C}_w^{-1} \cdot (\mathbf{I} - \mathbf{M}(\boldsymbol{\theta})(\mathbf{M}(\boldsymbol{\theta})^H \mathbf{C}_w^{-1} \mathbf{M}(\boldsymbol{\theta}))^{-1} \mathbf{M}(\boldsymbol{\theta})^H \mathbf{C}_w^{-1}) \hat{\mathbf{r}} \\ &= \arg \min_{\boldsymbol{\theta}} \hat{\mathbf{r}}^H \mathbf{C}_w^{-1/2} (\mathbf{I} - \boldsymbol{\Pi}(\boldsymbol{\theta})) \mathbf{C}_w^{-1/2} \hat{\mathbf{r}} \\ &= \arg \max_{\boldsymbol{\theta}} \hat{\mathbf{r}}^H \mathbf{C}_w^{-1/2} \boldsymbol{\Pi}(\boldsymbol{\theta}) \mathbf{C}_w^{-1/2} \hat{\mathbf{r}} \end{aligned}$$

where  $\boldsymbol{\Pi}(\boldsymbol{\theta}) = \mathbf{C}_w^{-1/2} \mathbf{M}(\boldsymbol{\theta})(\mathbf{M}(\boldsymbol{\theta})^H \mathbf{C}_w^{-1} \mathbf{M}(\boldsymbol{\theta}))^{-1} \mathbf{M}(\boldsymbol{\theta})^H \mathbf{C}_w^{-1/2}$ .

$\boldsymbol{\Pi}(\boldsymbol{\theta})$  is an orthogonal projection:  $\boldsymbol{\Pi}^2 = \boldsymbol{\Pi}$ ,  $\boldsymbol{\Pi}^H = \boldsymbol{\Pi}$ . The projection is onto the column span of  $\mathbf{M}'(\boldsymbol{\theta}) := \mathbf{C}_w^{-1/2} \mathbf{M}(\boldsymbol{\theta})$ . The estimation of the source positions  $\boldsymbol{\theta}$  is nonlinear. It could be obtained iteratively using a Newton iteration (cf. [31]). The sources can also be estimated sequentially [31], which provides an alternative to the CLEAN algorithm.

### 5.5 Two-step WLS solution

In the previous formulation (28), we estimated  $\boldsymbol{\sigma}$ , which contains both the source powers and the noise powers. However, the image is related only to the source powers,  $\boldsymbol{\sigma}_s$ . We can write more explicitly how these are estimated.

The optimization problem is again separable: given the optimal  $\boldsymbol{\sigma}_s$ , the “known” part of the LS problem (27) is  $\hat{\mathbf{r}} - \mathbf{M}_s \boldsymbol{\sigma}_s$ , and the corresponding estimate for  $\boldsymbol{\sigma}_n$  is

$$\hat{\boldsymbol{\sigma}}_n = (\mathbf{M}_n^H \mathbf{C}_w^{-1} \mathbf{M}_n)^{-1} \mathbf{M}_n \mathbf{C}_w^{-1} (\hat{\mathbf{r}} - \mathbf{M}_s \boldsymbol{\sigma}_s). \quad (31)$$

Plugging this solution back into the LS problem, we can first rewrite

$$\begin{aligned} & \mathbf{C}_w^{-1/2} (\hat{\mathbf{r}} - \mathbf{M}_s \boldsymbol{\sigma}_s - \mathbf{M}_n \hat{\boldsymbol{\sigma}}_n) \\ &= [\mathbf{C}_w^{-1/2} - \mathbf{C}_w^{-1/2} \mathbf{M}_n (\mathbf{M}_n^H \mathbf{C}_w^{-1} \mathbf{M}_n)^{-1} \mathbf{M}_n \mathbf{C}_w^{-1}] \hat{\mathbf{r}} - \\ & \quad - [\mathbf{C}_w^{-1/2} - \mathbf{C}_w^{-1/2} \mathbf{M}_n (\mathbf{M}_n^H \mathbf{C}_w^{-1} \mathbf{M}_n)^{-1} \mathbf{M}_n \mathbf{C}_w^{-1}] \mathbf{M}_s \boldsymbol{\sigma}_s \\ &= \mathbf{P}^\perp \mathbf{C}_w^{-1/2} (\hat{\mathbf{r}} - \mathbf{M}_s \boldsymbol{\sigma}_s) \end{aligned}$$

where

$$\mathbf{P}^\perp = \mathbf{I} - \mathbf{P}, \quad \mathbf{P} = \mathbf{C}_w^{-1/2} \mathbf{M}_n (\mathbf{M}_n^H \mathbf{C}_w^{-1} \mathbf{M}_n)^{-1} \mathbf{M}_n \mathbf{C}_w^{-1/2}.$$

Similar to  $\mathbf{II}$ , we can show that  $\mathbf{P}$  is an orthogonal projection. Hence, also  $\mathbf{P}^\perp = \mathbf{I} - \mathbf{P}$  is an orthogonal projection, onto the complement of the range of  $\mathbf{C}_w^{-1/2} \mathbf{M}_n$ , i.e., the weighted range of the noise matrix. If  $\boldsymbol{\Sigma}_n$  is diagonal, this is equivalent to “projecting out the diagonal”, thus omitting these entries in the WLS fitting. It is interesting to note that, in current telescopes, the autocorrelations (main diagonal of  $\mathbf{R}$ ) are usually not estimated. That fits in very well with this scheme, as the projection would project them out anyway!

The resulting “compressed” WLS problem is

$$\hat{\boldsymbol{\sigma}}_s = \arg \min_{\boldsymbol{\sigma}_s} \|\mathbf{C}_s^{-1/2} (\hat{\mathbf{r}} - \mathbf{M}_s \boldsymbol{\sigma}_s)\|^2$$

where

$$\mathbf{C}_s^{-1} := \mathbf{C}_w^{-1/2} \mathbf{P}^\perp \mathbf{C}_w^{-1/2}.$$

(This is with some abuse of notation:  $\mathbf{C}_s$  is singular due to the projection, hence not invertible. However, we will only need  $\mathbf{C}_s^{-1}$ , and will use the above definition for it.) The solution  $\hat{\boldsymbol{\sigma}}_s$  will be exactly the same as in the original WLS problem (27), but now it is obtained in two steps: first  $\hat{\boldsymbol{\sigma}}_s$  and then, if required,  $\hat{\boldsymbol{\sigma}}_n$  via (31).

As the expression for the compressed problem is very similar to the original WLS problem, we obtain similar results: the solution is

$$\hat{\boldsymbol{\sigma}}_s = (\mathbf{M}_s^H \mathbf{C}_s^{-1} \mathbf{M}_s)^{-1} \mathbf{M}_s^H \mathbf{C}_s^{-1} \hat{\mathbf{r}}$$

which can also be written as

$$\hat{\boldsymbol{\sigma}}_s = \mathbf{M}_{ds}^{-1} \hat{\boldsymbol{\sigma}}_{ds}, \quad \mathbf{M}_{ds} = \mathbf{M}_s^H \mathbf{C}_s^{-1} \mathbf{M}_s, \quad \hat{\boldsymbol{\sigma}}_{ds} = \mathbf{M}_s^H \mathbf{C}_s^{-1} \hat{\mathbf{r}}.$$

$\mathbf{M}_{ds}$  is the deconvolution matrix, and  $\hat{\boldsymbol{\sigma}}_{ds}$  is the WLS dirty image. This time, the dirty image is really a (vectorized) image, whereas in the previous discussion, the vector  $\hat{\boldsymbol{\sigma}}_d$  had an image component and a noise component.

The covariance of the image estimate  $\hat{\boldsymbol{\sigma}}_s$  is, using  $\mathbf{C}_s^{-1} \mathbf{C}_w \mathbf{C}_s^{-1} = \mathbf{C}_s^{-1}$ ,

$$\mathbf{C}\boldsymbol{\sigma}_s = (\mathbf{M}_s^H \mathbf{C}_s^{-1} \mathbf{M}_s)^{-1} \mathbf{M}_s^H \mathbf{C}_s^{-1} (\mathbf{C}_w) \mathbf{C}_s^{-1} \mathbf{M}_s (\mathbf{M}_s^H \mathbf{C}_s^{-1} \mathbf{M}_s)^{-1} = \mathbf{M}_{ds}^{-1}.$$

Thus, we obtain quite similar results as before when we estimated  $\boldsymbol{\sigma}$ , but now directly related to the image  $\boldsymbol{\sigma}_s$ , with the noise part  $\boldsymbol{\sigma}_n$  “projected out”.

## 5.6 Imaging using sparse reconstruction techniques

Compressive sampling/sensing (CS) is a “new” topic, currently drawing wide attention. It is connected to random sampling, and as such, it has been used in radio astronomy for a long time. In its basic formulation, we connect back to the measurement equation (23), or  $\hat{\mathbf{r}} = \mathbf{M}\boldsymbol{\sigma} + \mathbf{w}$ , and we consider the “overparametrized” formulation where each pixel in the image corresponds to a potential source in  $\boldsymbol{\sigma}$ , whereas  $\mathbf{M}$  is known. If the image is large, then the deconvolution problem (inversion of  $\mathbf{M}$ ) is ill conditioned. A direct inversion using (Weighted) Least Squares will give rise to unacceptable noise enhancement. We resorted to regularization by the KLT, which essentially projects the true image onto the selected basis, giving rise to artefacts. Without noise, any component orthogonal to the projection space can be added to the image without changing the modeling error: the image that fits the data is not unique. Additional constraints are needed. Examples are:

1. *Sparsity of the solution vector*, typically obtained by using an  $\ell_1$  norm (sum of absolute values), resulting in convex optimization problems like [21, 43]

$$\min_{\boldsymbol{\sigma}} \|\boldsymbol{\sigma}\|_1 \quad \text{subject to} \quad \|\hat{\mathbf{r}} - \mathbf{M}\boldsymbol{\sigma}\|_2^2 \leq \epsilon$$

or the equivalent

$$\min_{\boldsymbol{\sigma}} \|\hat{\mathbf{r}} - \mathbf{M}\boldsymbol{\sigma}\|_2^2 + \lambda \|\boldsymbol{\sigma}\|_1$$

These are versions of the Basis Pursuit problem. Like the KLT, the results depend on the chosen noise threshold  $\epsilon$  (or regularization parameter  $\lambda$ ). The sparsity assumption poses that the sky is mostly empty. Although it has already long been suspected that CLEAN is related to  $\ell_1$ -optimization [27] (in fact, it is now recognized as a Matching Pursuit algorithm [25]), CS theory states the general conditions under which this assumption is likely to recover the true image [21, 43]. Extensions are needed in case of extended emissions [23].

2. *Requiring the resulting image to be non-negative*. This is physically plausible, and to some extent already covered by CLEAN [27]. It is an explicit condition in a Non-Negative Least Squares (NNLS) formulation [7], which searches for a Least Squares fit while requiring that the solution  $\boldsymbol{\sigma}$  has all entries  $\sigma_i \geq 0$ . This turns out to be a strong constraint, readily incorporated into other formulations (e.g., CLEAN, MEM, and  $\ell_1$ -optimization).

Some experimental results using these algorithms are shown in [23, 35].

## 6 Calibration

### 6.1 Non-ideal measurements

The previous section showed that there are many options to make an image from radio interferometer data. However, there are in fact several effects that make matters more complicated.

#### Instrumental effects

So far we ignored the beam shape of the individual elements (antennas or dishes) of the array. In fact, any antenna has its own directional response,  $b(\mathbf{p})$ . This function is called the primary beam (to distinguish it from the dirty beam that results from beamforming during the synthesis operation). It is generally assumed that the primary beam is equal for all elements in the array. With  $Q$  point sources, we will collect the resulting samples of the primary beam into a vector  $\mathbf{b} = [b(\mathbf{p}_1), \dots, b(\mathbf{p}_Q)]^T$ . These coefficients are seen as gains that (squared) will multiply the source powers  $\sigma_q^2$ . The general shape of the primary beam  $b(\mathbf{p})$  is known from electromagnetic modeling during the design of the telescope. If this is not sufficiently accurate, then it has to be calibrated.

Initially the direction independent electronic gains and phases of the receiver chain of each element in the array are unknown and have to be estimated. They are generally different from element to element. We thus have an unknown vector  $\mathbf{g}$  (size  $J \times 1$ ) with complex entries that each multiply the output signal of each telescope.

Also the noise powers of each element are unknown and generally unequal to each other. We will still assume that the noise is independent from element to element. We can thus model the noise covariance matrix by an (unknown) diagonal  $\Sigma_n$ .

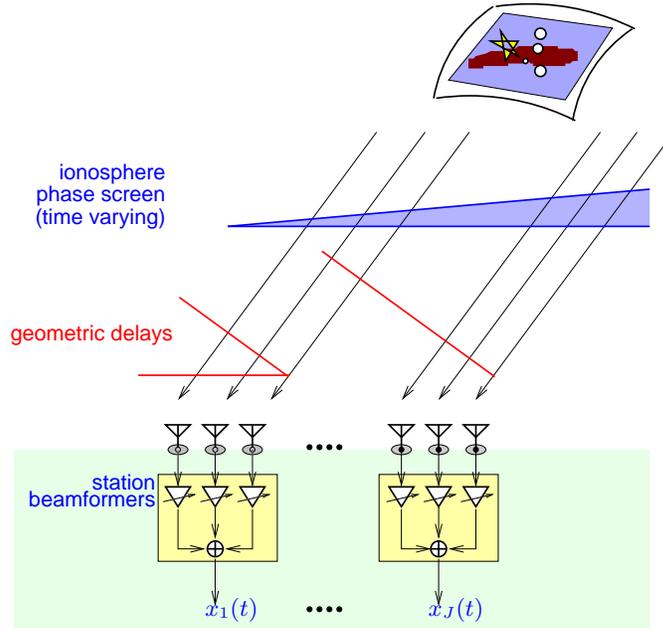
The modified data model that captures the above effects and replaces (12) is

$$\mathbf{R} = (\mathbf{\Gamma} \mathbf{A} \mathbf{B}) \Sigma_s (\mathbf{B}^H \mathbf{A}^H \mathbf{\Gamma}^H) + \Sigma_n \quad (32)$$

where  $\mathbf{\Gamma} = \text{diag}(\mathbf{g})$  is a diagonal with unknown receiver complex gains, and  $\mathbf{B} = \text{diag}(\mathbf{b})$  contains the samples of the primary beam. Usually,  $\mathbf{\Gamma}$  and  $\mathbf{B}$  are considered to vary only slowly with time  $m$  and frequency  $k$ , so that we can combine multiple covariance matrices  $\mathbf{R}_{m,k}$  with the same  $\mathbf{\Gamma}$  and  $\mathbf{B}$ .

#### Propagation effects

Ionospheric and tropospheric turbulence cause time-varying refraction and diffraction, which has a profound effect on the propagation of radio waves. In



**Fig. 9** A radio interferometer where stations consisting of phased array elements replace telescope dishes. The ionosphere adds phase delays to the signal paths. If the ionospheric electron density has the form of a *wedge*, it will simply shift the apparent positions of all sources.

the simplest case, the ionosphere is modeled as a thin layer at some height (say 100 km) above the Earth, causing delays that can be represented as phase shifts. At the low frequencies used for LOFAR, this effect is more pronounced. Generally it is first assumed that the ionosphere is “constant” over about 10 km and about 10 s. A better model is to model the ionospheric delay as a “wedge”, a linear function of the distance between piercing points (the intersection of the direction vectors  $\mathbf{p}_q$  with the ionospheric phase screen). As illustrated in figure 9, this modifies the geometric delays, leading to a shift in the apparent position of the sources. For larger distances, higher-order functions are needed to model the spatial behavior of the ionosphere, and if left uncorrected, the resulting image distortions are comparable to the distortions one sees when looking at lights at the bottom of a swimming pool.

Previously, we described that the array response matrix  $\mathbf{A}$  is really a function of the source direction vectors  $\mathbf{p}_q$ , and we wrote  $\mathbf{A}(\boldsymbol{\theta})$  where the vector  $\boldsymbol{\theta}$  is a suitable parametrization of the  $\mathbf{p}_q$  (typically two direction cosines per source). If a linear model for the ionospheric disturbance is sufficient, then it is sufficient to replace  $\mathbf{A}(\boldsymbol{\theta})$  by  $\mathbf{A}(\boldsymbol{\theta}')$ , where  $\boldsymbol{\theta}'$  differs from  $\boldsymbol{\theta}$  due to the shift in apparent direction of each source.

The modified data model that captures the above effects is thus

$$\mathbf{R} = (\mathbf{\Gamma} \mathbf{A}(\boldsymbol{\theta}') \mathbf{B}) \boldsymbol{\Sigma}_s (\mathbf{B}^H \mathbf{A}(\boldsymbol{\theta}')^H \mathbf{\Gamma}^H) + \boldsymbol{\Sigma}_n. \quad (33)$$

If we wish to be very general, we can write

$$\mathbf{R} = (\mathbf{G} \odot \mathbf{A}(\boldsymbol{\theta})) \boldsymbol{\Sigma}_s (\mathbf{G} \odot \mathbf{A}(\boldsymbol{\theta}))^H + \boldsymbol{\Sigma}_n \quad (34)$$

where  $\odot$  indicates an entrywise multiplication of two matrices (Schur-Hadamard product). Here,  $\mathbf{G}$  is a full matrix that captures all non-linear measurement effects. Equation (32) is recovered if we write  $\mathbf{G} = \mathbf{g} \mathbf{b}^H$  (i.e., a rank-1 matrix), and equation (33) if we write  $\mathbf{G} = \mathbf{g} \mathbf{b}^H \odot \mathbf{A}'$ , where  $\mathbf{A}'$  is a matrix consisting of phase corrections such that  $\mathbf{A}(\boldsymbol{\theta}') = \mathbf{A}(\boldsymbol{\theta}) \odot \mathbf{A}'$ .

Calibration is the process of identifying the unknown parameters in  $\mathbf{G}$ , and subsequently correcting for  $\mathbf{G}$  during the imaging step. The model (34) in its generality is not identifiable unless we make assumptions on the structure of  $\mathbf{G}$  (in the form of a suitable parametrization) and describe how it varies with time and frequency, e.g., in the form of (stochastic) models for these variations.

In practice, calibration is an integral part of the imaging step, and not a separate phase. In the next subsection, we will first describe how models of the form (32) or (33) can be identified. This step will serve as a stepping stone in the identification of a more general  $\mathbf{G}$ .

## 6.2 Calibration algorithms

### Estimating the element gains and directional responses

Let us assume a model of the form (32), where there are  $Q$  dominant calibration sources within the field of view. For these sources, we assume that their positions and source powers are known with sufficient accuracy from tables, i.e., we assume that  $\mathbf{A}$  and  $\boldsymbol{\Sigma}_s$  are known. We can then write (32) as

$$\mathbf{R} = \mathbf{\Gamma} \mathbf{A} \boldsymbol{\Sigma} \mathbf{A}^H \mathbf{\Gamma}^H + \boldsymbol{\Sigma}_n \quad (35)$$

where  $\boldsymbol{\Sigma} = \mathbf{B} \boldsymbol{\Sigma}_s \mathbf{B}$  is a diagonal with apparent source powers. With  $\mathbf{B}$  unknown,  $\boldsymbol{\Sigma}$  is unknown, but estimating  $\boldsymbol{\Sigma}$  is precisely the problem we studied before when we discussed imaging. Thus, once we have estimated  $\boldsymbol{\Sigma}$  and know  $\boldsymbol{\Sigma}_s$ , we can easily estimate the directional gains  $\mathbf{B}$ . The problem thus reduces to estimate the diagonal matrices  $\mathbf{\Gamma}$ ,  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Sigma}_n$  from a model of the form (35).

For some cases, e.g., arrays where the elements are traditional telescope dishes, the field of view is quite narrow (degrees) and we may assume that there is only a single calibrator source in the observation. Then  $\boldsymbol{\Sigma} = \sigma^2$  is a scalar and the problem reduces to

$$\mathbf{R} = \mathbf{g}\sigma^2\mathbf{g}^H + \boldsymbol{\Sigma}_n$$

and since  $\mathbf{g}$  is unknown, we could even absorb the unknown  $\sigma$  in  $\mathbf{g}$  (it is not separately identifiable). The structure of  $\mathbf{R}$  is a rank-1 matrix  $\mathbf{g}\sigma^2\mathbf{g}^H$  plus a diagonal  $\boldsymbol{\Sigma}_n$ . This is recognized as a “rank-1 factor analysis” model in multivariate analysis theory [26, 18]. Given  $\mathbf{R}$ , we can solve for  $\mathbf{g}$  and  $\boldsymbol{\Sigma}_n$  in several ways [4, 5, 48]. For example, any submatrix away from the diagonal is only dependent on  $\mathbf{g}$  and is rank 1. This allows direct estimation of  $\mathbf{g}$ . This property is related to the gain and phase closure relations often used in the radio astronomy literature for calibration (in particular, these relations express that the determinant of any  $2 \times 2$  submatrix away from the main diagonal will be zero, which is the same as saying that this submatrix is rank 1).

In general, there are more calibrator sources ( $Q$ ) in the field of view, and we have to solve (35). We resort to an Alternating Least Squares approach. If  $\boldsymbol{\Gamma}$  would be known, then we can correct  $\mathbf{R}$  for it, so that we have precisely the same problem as we considered before, (27), and we can solve for  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Sigma}_n$  using the techniques discussed in section 5.3. Alternatively, with  $\boldsymbol{\Sigma}$  known, we can say we know a reference model  $\mathbf{R}_0 = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^H$ , and the problem is to identify the element gains  $\boldsymbol{\Gamma} = \text{diag}(\mathbf{g})$  from a model of the form

$$\mathbf{R} = \boldsymbol{\Gamma}\mathbf{R}_0\boldsymbol{\Gamma}^H + \boldsymbol{\Sigma}_n$$

or, after applying the  $\text{vec}(\cdot)$ -operation,

$$\text{vec}(\mathbf{R}) = \text{diag}(\text{vec}(\mathbf{R}_0))(\bar{\mathbf{g}} \otimes \mathbf{g}) + \text{vec}(\boldsymbol{\Sigma}_n).$$

This leads to the Least Squares problem

$$\hat{\mathbf{g}} = \arg \min_{\mathbf{g}} \|\text{vec}(\hat{\mathbf{R}} - \boldsymbol{\Sigma}_n) - \text{diag}(\text{vec}(\mathbf{R}_0))(\bar{\mathbf{g}} \otimes \mathbf{g})\|^2.$$

This problem cannot be solved in closed form. Alternatively, we can first solve an unstructured problem: define  $\mathbf{x} = \bar{\mathbf{g}} \otimes \mathbf{g}$  and solve

$$\hat{\mathbf{x}} = \text{diag}(\text{vec}(\mathbf{R}_0))^{-1}\text{vec}(\hat{\mathbf{R}} - \boldsymbol{\Sigma}_n)$$

or equivalently, if we define  $\mathbf{X} = \mathbf{g}\mathbf{g}^H$ ,

$$\hat{\mathbf{X}} = (\hat{\mathbf{R}} - \boldsymbol{\Sigma}_n) \oslash \mathbf{R}_0.$$

where  $\oslash$  denotes an entrywise matrix division. After estimating the unstructured  $\mathbf{X}$ , we enforce the rank-1 structure  $\mathbf{X} = \mathbf{g}\mathbf{g}^H$ , via a rank-1 approximation, and find an estimate for  $\mathbf{g}$ . The pointwise division can lead to noise enhancement; this is remediated by only using the result as an initial estimate for a Gauss-Newton iteration [13] or by formulating a *weighted* least squares problem instead [45, 48].

With  $\mathbf{g}$  known, we can again estimate  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Sigma}_n$ , and make an iteration. Overall we then obtain an alternating least squares solution. A more optimal solution can be found by solving the overall problem (35) as a covariance matching problem with a suitable parametrization, and the more general algorithms in [31] lead to an asymptotically unbiased and statistically efficient solution.

The resulting algorithms are related to the classical self-calibration (Self-Cal) algorithm [10, 32] widely used in the radio astronomy literature, in particular for a single calibrator source. In that algorithm,  $\mathbf{R}_0$  is a reference model, obtained from the best known map at that point in the iteration. In the SelfCal iteration, the telescope gains are estimated, the corrections on  $\mathbf{R}$  are made, the next best image is constructed leading to a new reference model  $\mathbf{R}_0$ , etc.

### Estimating the ionospheric perturbation

The more general calibration problem (33) follows from (32) by writing  $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta}')$  where  $\boldsymbol{\theta}'$  are the apparent source locations. This problem can be easily solved in quite the same way: in the alternating least squares problem we solve for  $\mathbf{g}$ ,  $\boldsymbol{\theta}'$ ,  $\boldsymbol{\sigma}_s$  and  $\boldsymbol{\sigma}_n$  in turn, keeping the other parameters fixed at their previous estimates. After that, we can relate the apparent source locations to the (known) locations of the calibrator sources  $\boldsymbol{\theta}$ .

The resulting phase corrections  $\mathbf{A}'$  to relate  $\mathbf{A}(\boldsymbol{\theta}')$  to  $\mathbf{A}(\boldsymbol{\theta})$  via  $\mathbf{A}(\boldsymbol{\theta}') = \mathbf{A}(\boldsymbol{\theta}) \odot \mathbf{A}'$  gives us an estimate of the ionospheric phase screen in the direction of each source. These “samples” can then be interpolated to obtain a phase screen model for the entire field of view. This method is limited to the regime where the phase screen can be modeled as a linear gradient over the array. An implementation of this algorithm is called Field-Based Calibration [11].

Other techniques are based on “peeling” [28]. In this method of successive estimation and subtraction calibration, parameters are obtained for the brightest source in the field. The source is then removed from the data, and the process is repeated for the next brightest source. This leads to a collection of samples of the ionosphere, to which a model phase screen can be fitted.

### Estimating the general model

In the more general case (34), viz.

$$\mathbf{R} = (\mathbf{G} \odot \mathbf{A}) \boldsymbol{\Sigma}_s (\mathbf{G} \odot \mathbf{A})^H + \boldsymbol{\Sigma}_n,$$

we have an unknown full matrix  $\mathbf{G}$ . We assume  $\mathbf{A}$  and  $\boldsymbol{\Sigma}_s$  known. Since  $\mathbf{A}$  pointwise multiplies  $\mathbf{G}$  and  $\mathbf{G}$  is unknown, we might as well omit  $\mathbf{A}$  from the equations without loss of generality. For the same reason also  $\boldsymbol{\Sigma}_s$  can be

omitted. This leads to a problem of the form

$$\mathbf{R} = \mathbf{G}\mathbf{G}^H + \boldsymbol{\Sigma}_n,$$

where  $\mathbf{G} : J \times Q$  and  $\boldsymbol{\Sigma}_n$  (diagonal) are unknown. This problem is known as a rank- $Q$  factor analysis problem. For reasonably small  $Q$ , as compared to the size  $J$  of  $\mathbf{R}$ , the factor  $\mathbf{G}$  can be solved for, again using algorithms for covariance matching such as in [31]. We discuss this problem in more detail in section 7.

It is important to note that  $\mathbf{G}$  can be identified only up to a unitary factor  $\mathbf{V}$  at the right:  $\mathbf{G}' = \mathbf{G}\mathbf{V}$  would also be a solution. This factor makes the gains unidentifiable unless we introduce more structure to the problem.

To make matters worse, note that this problem is used to fine-tune earlier coarser models (33). At this level of accuracy, the number of dominant sources  $Q$  is often not small anymore, making  $\mathbf{G}$  not identifiable.

As discussed in [30] and studied in more detail in [39], more structure needs to be introduced to be able to solve the problem. Typically, what helps is to consider the problem for a complete observation (rather than for a single snapshot  $\mathbf{R}$ ) where we have many different frequencies  $f_k$  and time intervals  $m$ . The directional response matrix  $\mathbf{A}_{m,k}$  varies with  $m$  and  $k$  in a known way, and the instrumental gains  $\mathbf{g}$  and  $\mathbf{b}$  are relatively constant. The remaining part of  $\mathbf{G} = \mathbf{g}\mathbf{b}^H \odot \mathbf{A}'$  is due to the ionospheric perturbations, and models can be introduced to describe its fluctuation over time, frequency, and space using some low order polynomials. We can also introduce stochastic knowledge that describe a correlation of parameters over time and space.

New instruments such as LOFAR and SKA will only reach their full potential if this general calibration problem is solved. For LOFAR, a complete calibration method that incorporates many of the above techniques was recently proposed in [16]. In general, calibration and imaging need to be considered in unison, leading to many potential directions, approaches, and solutions. This promises to be a rich research area in years to come.

## 7 Factor analysis

### 7.1 Introduction

Many array signal processing algorithms are at some point based on the eigenvalue decomposition, which is used e.g., to make a distinction between the “signal subspace” and the “noise subspace”. By using orthogonal projections, part of the noise is projected out and only the signal subspace remains. This can then be used for applications such as high-resolution direction-of-arrival estimation, blind source separation, etc. In these applications, it is commonly

assumed that the noise is spatially white. However, this is valid only after suitable calibration.

Factor analysis considers covariance data models where the noise is uncorrelated but has unknown powers at each sensor, i.e., the noise covariance matrix is an arbitrary diagonal with positive real entries. In these cases the familiar eigenvalue decomposition (EVD) has to be replaced by a more general “Factor Analysis” decomposition (FAD), which then reveals all relevant information. It is a very relevant model for the early stages of data processing in radio astronomy, because at that point the instrument is not yet calibrated and the noise powers on the various antennas may be quite different. We saw two examples in section 6.

As it turns out, this problem has been studied in the psychometrics, biometrics and statistics literature since the 1930s (but usually for real-valued matrices) [18, 26]. The problem has received much less attention in the signal processing literature. In this section, we briefly describe the FAD and some algorithms for computing it.

## 7.2 Problem formulation

Assume as before that we have a set of  $Q$  narrow-band Gaussian signals impinging on an array of  $J$  sensors. The received signal can be described in complex envelope form by

$$\mathbf{x}(n) = \sum_{q=1}^Q \mathbf{a}_q s_q(n) + \mathbf{n}(n) = \mathbf{A}\mathbf{s}(n) + \mathbf{n}(n) \quad (36)$$

where  $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_Q]$  contains the array response vectors. In this model,  $\mathbf{A}$  is unknown, and the array response vectors are unstructured, i.e., we do not consider a directional model for them. The source vector  $\mathbf{s}(n)$  and noise vector  $\mathbf{n}(n)$  are considered i.i.d. Gaussian, i.e., the corresponding covariance matrices are diagonal. Without loss of generality, we can scale the source signals such that the source covariance matrix  $\mathbf{\Sigma}_s$  is identity.

The data covariance matrix thus has the form

$$\mathbf{R} = \mathbf{A}\mathbf{A}^H + \mathbf{D} \quad (37)$$

where we assume  $Q < J$  so that  $\mathbf{A}\mathbf{A}^H$  is rank deficient. Many signal processing algorithms are based on computing an eigenvalue decomposition of  $\mathbf{R}$  as  $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ , where  $\mathbf{U}$  is unitary and  $\mathbf{\Lambda}$  is a diagonal matrix containing the eigenvalues in descending order.

If  $\mathbf{D} = \mathbf{0}$  (no noise), then  $\mathbf{R}$  has rank  $Q$  and the eigenvalue decomposition specializes to

$$\mathbf{R} = \mathbf{U}\mathbf{A}_0\mathbf{U}^H = [\mathbf{U}_s \quad \mathbf{U}_n] \begin{bmatrix} \mathbf{A}_s & \\ & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_s^H \\ \mathbf{U}_n^H \end{bmatrix}$$

where  $\mathbf{A}_s$  contains the  $Q$  nonzero eigenvalues and  $\mathbf{U}_s$  the corresponding eigenvectors. The range of  $\mathbf{U}_s$  is called the signal subspace, its orthogonal complement  $\mathbf{U}_n$  the noise subspace.

For spatially white noise,  $\mathbf{D} = \sigma^2\mathbf{I}$ , we can write  $\mathbf{D} = \sigma^2\mathbf{U}\mathbf{U}^H$ , and the eigenvalue decomposition becomes

$$\mathbf{R} = \mathbf{U}\mathbf{A}\mathbf{U}^H = \mathbf{U}(\mathbf{A}_0 + \sigma^2\mathbf{I})\mathbf{U}^H = [\mathbf{U}_s \quad \mathbf{U}_n] \begin{bmatrix} \mathbf{A}_s + \sigma^2\mathbf{I} & \\ & \sigma^2\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{U}_s^H \\ \mathbf{U}_n^H \end{bmatrix}.$$

Hence, all eigenvalues are raised by  $\sigma^2$ , but the eigenvectors are unchanged. Algorithms based on  $\mathbf{U}_s$  can thus proceed as if there was no noise, thus leading to the use of the EVD and related subspace estimation algorithms in many array signal processing applications.

If the noise is not uniform, then  $\mathbf{D}$  is an unknown diagonal matrix, and the EVD does not reveal the signal subspace  $\mathbf{U}_s$ . The objective of factor analysis is, for given  $\mathbf{R}$ , to identify  $\mathbf{A}$  and  $\mathbf{D}$ , as well as the factor dimension  $Q$ . This can be seen as an extension of the eigenvalue decomposition, to be used if the noise covariance is a diagonal.

It is clear that for an arbitrary Hermitian matrix  $\mathbf{R}$ , this factorization can exist in its exact form only for  $Q \geq J$ , in which case we can set  $\mathbf{D} = \mathbf{0}$ , or any other value, which makes the factorization useless. Hence, for a noise-perturbed matrix, we wish to detect the smallest  $Q$  which gives a “reasonable fit”, and we will assume that  $Q < J$  is sufficiently small so that unique decompositions exist.

Furthermore, we cannot estimate  $\mathbf{A}$  uniquely, since  $\mathbf{A}$  can be replaced by  $\mathbf{A}\mathbf{V}$  for an arbitrary unitary matrix  $\mathbf{V}$ . If we denote  $\mathbf{A}\mathbf{A}^H = \mathbf{U}_s\mathbf{A}_s\mathbf{U}_s^H$ , it is clear that we can only estimate the column span of  $\mathbf{A}$ , i.e.,  $\text{ran}(\mathbf{A}) = \text{ran}(\mathbf{U}_s)$ , as well as the “signal eigenvalues”  $\mathbf{A}_s$ .

Suppose we have estimated  $\mathbf{D}$ , then we can whiten  $\mathbf{R}$ :

$$\tilde{\mathbf{R}} := \mathbf{D}^{-1/2}\mathbf{R}\mathbf{D}^{-1/2} = (\mathbf{D}^{-1/2}\mathbf{A})(\mathbf{A}^H\mathbf{D}^{-1/2}) + \mathbf{I}.$$

At this point, we can introduce the usual eigenvalue decomposition of  $\tilde{\mathbf{R}}$ :

$$\tilde{\mathbf{R}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{U}}^H$$

and identify  $\mathbf{D}^{-1/2}\mathbf{A}\mathbf{V} = \tilde{\mathbf{U}}$ , or  $\mathbf{A} = \mathbf{D}^{1/2}\tilde{\mathbf{U}}\mathbf{V}^H$ , where  $\mathbf{V}$  is an arbitrary unitary factor. If we choose  $\mathbf{V} = \mathbf{I}$ , we obtain  $\mathbf{A}^H\mathbf{D}^{-1}\mathbf{A} = \tilde{\mathbf{\Lambda}}$  is diagonal, which is a constraint that is sometimes used to obtain a more unique parametrization of  $\mathbf{A}$ . Note that  $\mathbf{A}$  is not yet quite unique, because in the complex case each column of  $\mathbf{A}$  can be scaled by an arbitrary complex phase, and the columns may be reordered as well. However, the point is that once  $\mathbf{D}$  is known, we are back on familiar grounds.

Regarding identifiability, we generally require to have more “equations” than “unknowns”. Here, the number of available equations is equal to the number of (real) parameters in  $\hat{\mathbf{R}}$ , which is  $J$  (real) entries on the main diagonal and  $J(J-1)$  parameters for the off-diagonal (complex) entries, taking into account Hermitian symmetry. The number of unknowns is  $2JQ$  (real) parameters for  $\mathbf{A}$ , and  $J$  parameters for  $\mathbf{D}$ , minus the number of constraints to make  $\mathbf{A}$  unique. Taking the constraint that  $\mathbf{A}^H \mathbf{D}^{-1} \mathbf{A}$  is diagonal gives  $Q^2 - Q$  constraints on the parameters of  $\mathbf{A}$ , and further setting the first row of  $\mathbf{A}$  to be real gives another  $Q$  constraints. In total we have for the number of equations minus the number of unknowns

$$s = J + J(J-1) - (2JQ + J - (Q^2 - Q + Q)) = (J - Q)^2 - J.$$

Requiring  $s > 0$  leads to the condition  $Q < J - \sqrt{J}$ . This is an upper bound on the factor rank.

In Factor Analysis, there are two problems:

1. Detection: given  $\hat{\mathbf{R}}$ , estimate  $Q$ . The hypothesis that the factor rank is  $q$  is denoted by  $\mathcal{H}_q$ .
2. Identification: given  $\hat{\mathbf{R}}$  and  $Q$ , estimate  $\mathbf{D}$  and  $\mathbf{A}$ , or  $\mathbf{A}_s$  and  $\mathbf{U}_s$ .

We consider the latter problem first.

### 7.3 Computing the Factor Analysis decomposition

Assume we know  $Q$ . Let  $\boldsymbol{\theta}$  be a minimal parametrization of  $(\mathbf{A}, \mathbf{D})$ , dependent on  $Q$ , such that  $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{A}^H + \mathbf{D}$ . If we start from a likelihood perspective, we obtain after standard derivations that the maximum likelihood estimate of  $\mathbf{R}$  is obtained by finding the model parameters  $\boldsymbol{\theta}$  such that

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} N \left( \ln |\mathbf{R}(\boldsymbol{\theta})| + \text{tr}(\mathbf{R}(\boldsymbol{\theta})^{-1} \hat{\mathbf{R}}) \right).$$

where  $\hat{\mathbf{R}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n)\mathbf{x}(n)^H$  is the sample covariance matrix. This is exactly the same problem as we saw before in (26), and we can follow the same solution strategy.

In particular, we can use the result from [31] that the ML problem is asymptotically (large  $N$ ) equivalent to the Weighted Least Squares problem

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\mathbf{C}_w^{-1/2}(\hat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta}))\|^2 = \arg \min_{\boldsymbol{\theta}} (\hat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta}))^H \mathbf{C}_w^{-1} (\hat{\mathbf{r}} - \mathbf{r}(\boldsymbol{\theta})) \quad (38)$$

where as before  $\mathbf{r} = \text{vec}(\mathbf{R})$ ,  $\hat{\mathbf{r}} = \text{vec}(\hat{\mathbf{R}})$ , and the weighting matrix  $\mathbf{C}_w$  is the covariance of  $\hat{\mathbf{r}}$ , i.e.,  $\mathbf{C}_w = (1/N)(\bar{\mathbf{R}} \otimes \mathbf{R})$ . This is precisely in context of [31],

and we can use the algorithms proposed there: Gauss-Newton iterations, the scoring algorithm or sequential estimation algorithms.

It is also possible to propose an alternating least squares approach. Given an estimate for  $\mathbf{D}$ , then, as mentioned above, we can whiten  $\hat{\mathbf{R}}$  by  $\mathbf{D}$ , do an eigenvalue decomposition on  $\hat{\mathbf{R}} = \mathbf{D}^{-1/2} \hat{\mathbf{R}} \mathbf{D}^{-1/2}$ , and estimate  $\mathbf{A}$  of size  $J \times Q$ , taking into account some suitable constraints to make  $\mathbf{A}$  unique. For  $\mathbf{A}$  known, the optimal  $\mathbf{D}$  in turn is given by  $\text{diag}(\mathbf{R} - \mathbf{A}\mathbf{A}^H)$ . Given a reasonable initial point (e.g.,  $\mathbf{D}^{(0)} = \text{diag}(\hat{\mathbf{R}})$ ), we can easily alternate between these two solutions. Convergence is to a local optimum and may be very slow.

An alternative approach was recently proposed in [36]. The ML cost function is shown to be equivalent to the Kullback-Leibler norm as often used in information theory, and a suitable algorithm is the Expectation Maximization (EM) algorithm. This is an iterative estimation algorithm which is shown in [36] to reduce, for current estimates  $(\mathbf{A}_k, \mathbf{D}_k)$ , to

$$\begin{aligned} \mathbf{R}_k &:= \mathbf{A}_k \mathbf{A}_k^H + \mathbf{D}_k \\ \boldsymbol{\Phi}_k &:= \mathbf{I} - \mathbf{A}_k^H \mathbf{R}_k^{-1} \mathbf{A}_k + \mathbf{A}_k^H \mathbf{R}_k^{-1} \hat{\mathbf{R}} \mathbf{R}_k^{-1} \mathbf{A}_k \\ \mathbf{A}_{k+1} &:= \hat{\mathbf{R}} \mathbf{R}_k^{-1} \mathbf{A}_k \boldsymbol{\Phi}_k^{-1} \\ \mathbf{D}_{k+1} &:= \text{diag}(\hat{\mathbf{R}} - \mathbf{A}_{k+1} \mathbf{A}_{k+1}^H \mathbf{R}_k^{-1} \hat{\mathbf{R}}). \end{aligned}$$

As any EM algorithm, it will converge to a local optimum, where convergence is guaranteed and typically reasonably fast.

#### 7.4 Rank detection

The detection problem is to estimate the factor rank  $Q$ . The largest permissible value of  $Q$  is that for which the number of unknown (real) parameters  $s = (J - Q)^2 - J \geq 0$ , or  $Q \leq J - \sqrt{J}$ . For larger  $Q$ , there is no identifiability of  $\mathbf{A}$  and  $\mathbf{D}$ : any sample covariance matrix  $\hat{\mathbf{R}}$  can be fitted.

To find  $Q$ , we can define a collection of hypotheses

$$\mathcal{H}_q : \quad \mathbf{x}(k) \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_q) \quad q = 1, 2, \dots \quad (39)$$

which are tested in turn against the null hypothesis

$$\mathcal{H}' : \quad \mathbf{x}(k) \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}'),$$

where  $\mathcal{CN}(\mathbf{0}, \mathbf{R})$  denotes the zero-mean complex normal distribution with covariance  $\mathbf{R}$ , and  $\mathbf{R}_q$  is the covariance matrix of the model with  $q$  interferers,

$$\mathbf{R}_q = \mathbf{A}\mathbf{A}^H + \mathbf{D}, \quad \text{where } \mathbf{A} : J \times q, \quad \mathbf{D} \text{ diagonal.}$$

$\mathcal{H}'$  corresponds to a default hypothesis of an arbitrary (unstructured) positive definite matrix  $\mathbf{R}'$ .

The generalized likelihood ratio test (GLRT) [17] is applicable. In this test, we have to insert maximum likelihood estimates for each of the unknown parameters, under each of the hypotheses. For  $\mathcal{H}_q$ , we can use the estimation techniques from the previous subsection. For  $\mathcal{H}'$ , the ML estimate  $\mathbf{R}'$  is equal to the sample covariance,  $\mathbf{R}' = \hat{\mathbf{R}}$ .

Under  $\mathcal{H}_q$ , respectively  $\mathcal{H}'$ , the log-likelihood is (dropping constants)

$$\begin{aligned}\log(L_q) &= -N \log |\mathbf{R}_q^{-1}| - N \text{tr}(\mathbf{R}_q^{-1} \hat{\mathbf{R}}) \\ \log(L') &= -N \log |\hat{\mathbf{R}}^{-1}| - Np.\end{aligned}$$

The log-likelihood ratio is then

$$\log(\lambda) := \log\left(\frac{L'}{L_q}\right) = N \text{tr}(\mathbf{R}_q^{-1} \hat{\mathbf{R}}) + N \log |\mathbf{R}_q^{-1} \hat{\mathbf{R}}| - Np. \quad (40)$$

Here,  $\lambda = L'/L_q$  is the test statistic (likelihood ratio), and we will reject  $\mathcal{H}_q$  and accept  $\mathcal{H}'$  if  $\lambda > \gamma$ , where  $\gamma$  is a predetermined threshold. Typically,  $\gamma$  is determined such that we obtain an acceptable “false-alarm” rate (i.e., the probability that we accept  $\mathcal{H}'$  instead of  $\mathcal{H}_q$ , while  $\mathcal{H}_q$  is actually true). To establish  $\gamma$ , we need to know the statistics of  $\lambda$  under  $\mathcal{H}_q$ .

Generalizing the results from the real-valued case [18, 26], we obtain that for moderately large  $N$  (say  $N > 50$ ), the test statistic  $2 \log(\lambda)$  has approximately a  $\chi_s^2$  distribution, where  $s$  is equal to “the number of free parameters” under  $\mathcal{H}_q$  (the number of equations minus the number of unknowns). For the complex case, this number is  $s = (J - q)^2 - J$  degrees of freedom.

In view of results of Box and Bartlett, a better fit of the distribution of  $2 \log(\lambda)$  to a  $\chi_s^2$  distribution is obtained by replacing  $N$  in (40) by [18, 26]

$$N' = N - \frac{1}{6}(2J + 11) - \frac{2}{3}Q.$$

To detect  $Q$ , we start with  $q = 0$ , and apply the test for increasing values of  $q$  until it is accepted, or until  $s = (J - q)^2 - J$  is negative. In that case, the hypothesis  $\mathcal{H}'$  is accepted, i.e., the given  $\hat{\mathbf{R}}$  is an unstructured covariance matrix. A disadvantage of this process is that the model parameters for each  $q$  have to be estimated, which can become quite cumbersome if  $J$  is large. Also, as for any sequential hypothesis test, the actual false alarm rate that is achieved is unknown, because the tests are not independent.

## 7.5 Application to interference cancellation

In the context of radio astronomy, factor analysis shows up in a number of applications. We already mentioned calibration in section 6. Another application is interference cancellation. In general, this is a large topic with many aspects. Here, we consider a simple case where we take short STIs and an uncalibrated array. Since astronomical sources are weak and much below the background noise level, if we integrate only over short intervals, the noise is dominant. Therefore, in the absence of interference, the STI data covariance matrix  $\mathbf{R} = \mathbf{R}_m$  could be modeled as a diagonal  $\mathbf{D}$ . Assuming  $Q$  independent interfering signals gives us a contribution  $\mathbf{A}\mathbf{A}^H$ . The approach for interference cancellation using spatial filtering is to estimate  $\text{ran}(\mathbf{A})$ , and to apply to  $\mathbf{R}$  a projector  $\mathbf{P}_\mathbf{A}^\perp$  onto the orthogonal complement of the span, i.e.,  $\mathbf{R}' = \mathbf{P}_\mathbf{A}^\perp \mathbf{R} \mathbf{P}_\mathbf{A}^\perp$ . That should remove the interference. The filtered covariance matrices are further averaged, and corrections need to be applied since also the astronomical data has been filtered. Details on this approach can be found in [34, 40].

Here, we describe only a limited-scope simulation on synthetic data, where we estimate a rank-1 subspace (*i*) using factor analysis, and for comparison (*ii*) using eigendecomposition assuming that  $\mathbf{D} = \sigma^2 \mathbf{I}$ , or (*iii*) using the eigendecomposition after whitening by  $\mathbf{D}^{-1/2}$ , assuming the true  $\mathbf{D}$  is known from calibration. The correct rank is  $Q = 1$ , and we show the residual interference power after projection, i.e.,  $\|\mathbf{P}_\mathbf{a}^\perp \mathbf{a}\|$  as a function of number of samples  $N$ , mean noise power, and deviation in noise power. The noise powers are randomly generated at the beginning of the simulation, uniformly in an interval. Legends in the graphs indicate the nominal noise power and the maximal deviation. All simulations use  $J = 8$  sensors, and a nominal interference to noise ratio per channel of 0 dB.

The results are shown in figure 10. The first graph shows the residual interference power for varying maximal deviations, the second graph shows

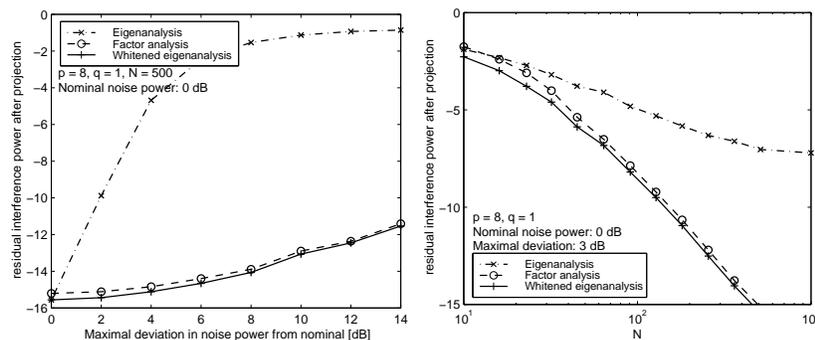


Fig. 10 Residual interference power after projections.

the residual for varying number of samples  $N$ , and a maximal deviation of 3 dB of the noise powers. The figures indicate that already for small deviations of the noise powers it is essential to take this into account, by using the FAD instead of the EVD. Furthermore, the estimates from the factor analysis are nearly as good as can be obtained via whitening with known noise powers.

## 8 Concluding remarks and further reading

In this chapter, we presented a signal processing viewpoint on radio astronomy. We showed how, with the right translations, the “measurement equations” are connected to covariance matrix data models used in the phased array signal processing literature. In this presentation, the resulting data models are very compact and clean, in the sense that the most straightforward covariance data models, widely studied in the signal processing literature as theoretical models, already seem valid. This is because far field assumptions clearly hold, and the propagation channels are very simple (no multipath), in contrast to other array processing applications such as seismology, synthetic aperture radar, or biomedical tomography.

However, this does not mean that radio astronomy is a “simple” application: data volumes are massive, and the requirements on resolution and accuracy are mind-boggling. Current telescopes, developed in the 1970s, start with signals sampled at 1–2 bits accuracy (because anyway the signals are mostly noise), and after data reduction and map making routinely end up with images with a dynamic range of  $10^5$ .

So far, radio astronomy has done very well without explicit connection to the array signal processing literature. However, we expect that, by making this connection, a wealth of new insights and access to “new” algorithms can be obtained. This will be beneficial, and possibly essential, for the development of new instruments like LOFAR and SKA.

For further reading we suggest, first of all, the classical radio astronomy textbooks, e.g., Thompson [37] and Perley [33]. The August 2009 issue of the *Proceedings of the IEEE* was devoted to the presentation of new instruments. The January 2010 issue of *IEEE Signal Processing Magazine* gave a signal processing perspective. For general insights into imaging and deconvolution, we suggest Blahut [2].

Challenges for signal processing lie in (1) imaging, (2) calibration, (3) interference suppression. These problems are really intertwined. It is interesting to note that, especially for calibration and interference suppression, factor analysis is an essential tool. Our contributions in these areas have appeared in [20, 19, 41, 1, 22, 4, 39, 47, 48, 46] and are summarized in the PhD theses [3, 44, 38], which should provide ample details for further reading.

## References

1. Ben-David, C., Leshem, A.: Parametric high resolution techniques for radio astronomical imaging. *IEEE J. Sel. Topics in Signal Processing* **2**(5), 670–684 (2008)
2. Blahut, R.E.: *Theory of remote image formation*. Cambridge University Press (2004). ISBN 0521553733
3. Boonstra, A.J.: Radio frequency interference mitigation in radio astronomy. Ph.D. thesis, TU Delft, Dept. EEMCS (2005). ISBN 90-805434-3-8
4. Boonstra, A.J., van der Veen, A.J.: Gain calibration methods for radio telescope arrays. *IEEE Trans. Signal Processing* **51**(1), 25–38 (2003)
5. Boonstra, A.J., Wijnholds, S.J., van der Tol, S., Jeffs, B.: Calibration, sensitivity and RFI mitigation requirements for LOFAR. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Philadelphia (Penn.), USA (2005)
6. Borgiotti, G.B., Kaplan, L.J.: Superresolution of uncorrelated interference sources by using adaptive array techniques. *IEEE Trans. Antennas Propagat.* **27**, 842845 (1979)
7. Briggs, D.S.: High fidelity deconvolution of moderately resolved sources. Ph.D. thesis, New Mexico Inst. of Mining and Technology, Socorro (NM) (1995). URL <http://www.aoc.nrao.edu/dissertations/dbriggs/>
8. Cornwell, T., Braun, R., Briggs, D.S.: Deconvolution. In: G.B. Taylor, C.L. Carilli, R.A. Perley (eds.) *Synthesis Imaging in Radio Astronomy II, Astronomical Society of the Pacific Conference Series*, vol. 180, pp. 151–170 (1999)
9. Cornwell, T.J.: Multiscale CLEAN deconvolution of radio synthesis images. *IEEE J. Sel. Topics in Signal Processing* **2**(5), 793–801 (2008)
10. Cornwell, T.J., Wilkinson, P.N.: A new method for making maps with unstable radio interferometers. *Mon. Not. R. Astron. Soc.* **196**, 1067–1086 (1981)
11. Cotton, W.D., Condon, J.J., Perley, R.A., Kassim, N., Lazio, J., Cohen, A.S., Lane, W., Erickson, W.C.: Beyond the isoplanatic patch in the VLA Low-frequency Sky Survey. In: *Proc. SPIE*, vol. 5489, pp. 180–189. Glasgow (2004)
12. Dewdney, P., Hall, P., Schilizzi, R., Lazio, T.: The square kilometre array. *Proceedings of the IEEE* **97**(8), 1482–1496 (2009)
13. Fuhrmann, D.R.: Estimation of sensor gain and phase. *IEEE Trans. Signal Processing* **42**(1), 77–87 (1994)
14. Hayes, M.H.: *Statistical Digital Signal Processing and Modeling*. John Wiley and Sons (1996)
15. Hogbom, J.A.: Aperture synthesis with non-regular distribution of interferometer baselines. *Astronomy and Astrophysics Suppl.* **15**, 417–426 (1974)
16. Intema, H.T., van der Tol, S., Cotton, W.D., Cohen, A.S., van Bemmell, I.M., Rottgering, H.J.A.: Ionospheric calibration of low frequency radio interferometric observations using the peeling scheme. I. Method description and first results. *Astronomy & Astrophysics* **501**(3), 1185–1205 (2009)
17. Kay, S.M.: *Fundamentals of Statistical Signal Processing. Volume II: Detection Theory*. Prentice Hall PTR, Upper Saddle River, NJ (1998)
18. Lawley, D.N., Maxwell, A.E.: *Factor Analysis as a Statistical Method*. Butterworth & Co, London (1971)
19. Leshem, A., van der Veen, A.J.: Radio-astronomical imaging in the presence of strong radio interference. *IEEE Transactions on Information Theory* **46**(5), 1730–1747 (2000)
20. Leshem, A., van der Veen, A.J., Boonstra, A.J.: Multichannel interference mitigation techniques in radio astronomy. *Astrophysical Journal Supplements* **131**(1), 355–374 (2000)
21. Levanda, R., Leshem, A.: Radio astronomical image formation using sparse reconstruction techniques. In: *IEEE 25th convention of Elec. Electron. Eng. in Israel (IEEEI 2008)*, pp. 716–720 (2008)

22. Levanda, R., Leshem, A.: Image formation in radio astronomy. *IEEE Signal Processing Magazine* **27**(1) (2010)
23. Li, F., Cornwell, T.J., de Hoog, F.: The application of compressive sampling to radio astronomy; I deconvolution. *Astronomy and Astrophysics* (528), A31 1–10 (2011)
24. Lonsdale, C., et al.: The Murchison Widefield Array: Design overview. *Proceedings of the IEEE* **97**(8), 1497–1506 (2009)
25. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. *IEEE Tr. Signal Processing* **41**(12), 3397–3415 (1993)
26. Mardia, K.V., Kent, J.T., Bibby, J.M.: *Multivariate Analysis*. Academic Press (1979)
27. Marsh, K.A., Richardson, J.M.: The objective function implicit in the CLEAN algorithm. *Astronomy and Astrophysics* **182**(1), 174–178 (1987)
28. Mitchell, D.A., Greenhill, L.J., Wayth, R.B., Sault, R.J., Lonsdale, C.J., Cappallo, R.J., Morales, M.F., Ord, S.M.: Real-time calibration of the Murchison Widefield Array. *IEEE Journal of Selected Topics in Signal Processing* **2**(5), 707–717 (2008)
29. Moon, T.K., Stirling, W.C.: *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall (2000). ISBN 0201361868
30. Noordam, J.E.: Generalized self-calibration for LOFAR. In: XXVIIth General Assembly of the International Union of Radio Science (URSI). Maastricht (The Netherlands) (2002)
31. Ottersten, B., Stoica, P., Roy, R.: Covariance matching estimation techniques for array signal processing applications. *Digital Signal Processing, A Review Journal* **8**, 185–210 (1998)
32. Pearson, T.J., Readhead, A.C.S.: Image formation by self-calibration in radio astronomy. *Annual Rev. Astronomy and Astrophysics* **22**, 97–130 (1984)
33. Perley, R.A., Schwab, F.R., Bridle, A.H.: *Synthesis Imaging in Radio Astronomy, Astronomical Society of the Pacific Conference Series*, vol. 6. BookCrafters Inc. (1994)
34. Raza, J., Boonstra, A.J., van der Veen, A.J.: Spatial filtering of RF interference in radio astronomy. *IEEE Signal Processing Letters* **9**(2), 64–67 (2002)
35. Schwardt, L.C.: Compressed sensing imaging with the KAT-7 array. In: *Int. Conf. on Electromagnetics in Adv. Appl. (ICEAA)*, pp. 690–693. IEEE (2012)
36. Seghouane, A.K.: An iterative projections algorithm for ML factor analysis. In: *IEEE Workshop on Machine Learning for Signal Processing (MLSP08)*, pp. 333–338 (2008)
37. Thompson, A.R., Moran, J.M., Swenson, G.W.: *Interferometry and Synthesis in Radio Astronomy*, 2nd edn. Wiley, New York (2001)
38. van der Tol, S.: Bayesian estimation for ionospheric calibration in radio astronomy. Ph.D. thesis, TU Delft, Dept. EEMCS (2009)
39. van der Tol, S., Jeffs, B.D., van der Veen, A.J.: Self-calibration for the LOFAR radio astronomical array. *IEEE Transactions on Signal Processing* **55**(9), 4497–4510 (2007)
40. van der Tol, S., van der Veen, A.J.: Performance analysis of spatial filtering of RF interference in radio astronomy. *IEEE Transactions on Signal Processing* **53**(3), 896–910 (2005)
41. van der Veen, A.J., Leshem, A., Boonstra, A.J.: Array signal processing for radio astronomy. *Experimental Astronomy* **17**(1-3), 231–249 (2004)
42. de Vos, M., Gunst, A., Nijboer, R.: The LOFAR telescope: System architecture and signal processing. *Proceedings of the IEEE* **97**(8), 1431–1437 (2009)
43. Wiaux, Y., Jacques, L., Puy, G., Scaife, A., Vanderghynst, P.: Compressed sensing imaging techniques for radio interferometry. *Monthly Notices of the Royal Astronomical Society* **395**, 17331742 (2009)

44. Wijnholds, S.J.: Fish-eye observing with phased array radio telescopes. Ph.D. thesis, TU Delft, Dept. EEMCS (2010). ISBN 978-90-9025180-6
45. Wijnholds, S.J., Boonstra, A.J.: A multisource calibration method for phased array telescopes. In: Fourth IEEE Workshop on Sensor Array and Multi-channel Processing (SAM). Waltham (Mass.), USA (2006)
46. Wijnholds, S.J., van der Tol, S., Nijboer, R., van der Veen, A.J.: Calibration challenges for the next generation of radio telescopes. *IEEE Signal Processing Magazine* **27**(1), 32–42 (2010)
47. Wijnholds, S.J., van der Veen, A.J.: Fundamental imaging limits of radio telescope arrays. *IEEE J. Sel. Topics in Signal Processing* **2**(5), 613–623 (2008)
48. Wijnholds, S.J., van der Veen, A.J.: Multisource self-calibration for sensor arrays. *IEEE Tr. Signal Processing* **57**(9), 3512–3522 (2009)