Chapter 4

Optimal data-model inversion

Statistics are like a bikini. What they reveal is suggestive, but what they conceal is vital.

Aaron Levenstein

ABSTRACT
The data model describing the response of the LOFAR telescope to the intensity distribution of the sky is characterized by the non-linearity of the parameters and the large level of noise compared to the desired cosmological signal. The application of the primary calibration needs to take those issues into account. In this chapter, we discuss the implementation of a statistically optimal map-making process and its properties. The basic assumptions of this method are that the noise is Gaussian and independent between the stations and frequency channels and that the dynamic range of the data can been improved significantly during the off-line LOFAR processing. These assumptions match our expectations for the LOFAR EoR experiment.

4.1 Introduction

The main scientific goal of LOFAR is the detection of the redshifted 21-cm signal from the Epoch of Reionization, which is the epoch when the first radiating sources in the Universe formed [Madau et al., 1997b]. Despite our lack of direct observations to date of this epoch, except a number of drop-out sources at $z \gtrsim 6.5$ [Oesch et al., 2009] and a gamma-ray burst at $z = 8.2$ [Salvaterra et al., 2009], we have a number of indirect observational constraints. We know from Lyman-$\alpha$ forest data that the universe is highly ionized at redshifts $\lesssim 6$. At higher redshifts the Sloan Digital Sky Survey quasars clearly indicate an increase of the Intergalactic Medium (IGM) neutral fraction, demarcating the tail of the reionization process [Fan et al., 2003, 2006]; the amount of neutral hydrogen
at $z \gtrsim 6$ is quite uncertain (e.g. Mesinger & Dijkstra 2008). Moreover, the WMAP polarization data show that the optical depth for Thomson scattering $\tau$ of the CMB off the electrons released during reionization is about 0.09 (Dunkley et al., 2009b), which puts a strong integral constraint on the EoR process but says very little about the details of the reionization process. This evidence is not enough if one wishes to uncover the answers to key questions about the sources of reionization, its timeline and its effect on subsequent structure formation.

It is widely recognized that the redshifted 21-cm line of the hyperfine transition of neutral hydrogen is the most promising probe of the detailed evolution of the EoR. This line follows, in principle, the reionization process as it evolved in space and time. Unfortunately, the strength of the expected signal is very weak and it is normally masked by bright foregrounds.

The detection of the EoR signal fits marginally within the sensitivity limits of the upcoming generation of low-frequency radio arrays like LOFAR and the MWA. In these experiments, the signal extraction is also hampered by the level of thermal noise plus instrumental and ionospheric distortions. Therefore, the initial aim of the first generation of EoR experiments is to detect the cosmic signal statistically. At the frequency range covered by these instruments, dipoles and phased arrays are sufficient to receive radio waves in the upper VHF band (30-300 MHz) and more complex antenna designs like dishes with feed horns etc. are not necessary.

Phased dipole-array designs, however, pose several challenges that one does not encounter in classical dish-based radio telescopes (Jeffs et al., 2008; Maaskant et al., 2006). The beam of phased arrays varies much more than the beams of large dish antennas (Bhatnagar et al., 2008). The variations in the beam pattern are more pronounced and less predictable (i.e. some elements or sub-arrays might fail while observing). This requires much more sophisticated direction-dependent calibration. In addition, the ionospheric scintillation and refraction is much more significant at low-frequencies (Koopmans, 2010). The next generation radio arrays also have intrinsically large Fields of View (FOV) and approximating the sky as a plane, thus ignoring its curvature, leads to undesired distortions in the maps, a problem that was addressed in a limited manner during the analysis of interferometric data by current telescopes. Finally, the unique spatial and frequency sampling of the new generation of low-frequency instruments mean that the amount of data generated will exceed the amount of data generated by classical interferometers by two to three orders of magnitude. The time required to analyze the data can in many cases exceed the total integration time by an order of magnitude as well.

One might therefore ask the question: What are the prospects and limitations of EoR imaging or power-spectrum estimation with a phased array? Even if the array is perfectly calibrated the result will be noise-limited at diffraction-limited resolution. The fidelity of the map is difficult to assess even in this simple case, due to the improper sampling of the uv plane. This leads to a number of questions regarding the calibration:

- How can we correct, to the desired level of accuracy, for the systematic errors in the presence of noise?
- How can we achieve this in a computationally efficient way, given the large number of visibilities that the experiment produces.

1http://www.mwatelescope.org
• What is the theoretical limit on calibration accuracy (expressed via the information theoretic Cramér-Rao bound) and can we reach it in practice?

• Is the detection of the EoR signal feasible under the various instrumental distortions, incomplete uv-plane sampling and large noise power?

The large amount of data (∼ 1 – 2 petabytes) and the complex imaging problem lead to certain limitations in the selection of the data analysis methods. In the case of the MWA, where the amount of correlated elements is 512, the correlator output is so high that it is necessary to perform calibration and imaging in real time. In contrast this is not a problem for LOFAR as it utilizes fewer correlated elements and the data can be stored for further processing. This enables us to use iterative methods and revisit the data for a number of iterations.

During the history of synthesis imaging a handful of shortcuts like the w-projection (Bhatnagar et al., 2008), faceted imaging and the CLEAN approach to deconvolution (Hög bom, 1974; Clark, 1980; Voronkov & Wieringa, 2004) have been used to alleviate the computational burden of interferometric data processing. The validity and application of those methods is hard to assess mathematically and their validity is often shown only through practical experience and rules of thumb.

In any case, the properties of the noise have to be understood to a very high level of accuracy. The noise of a map is a combination of thermal noise from the receivers, the sky noise, residual calibration errors and confusion noise. This is the “effective” noise (see also Wijnholds 2010) and we shall refer to this whenever we mention the term noise, unless stated otherwise explicitly.

In array signal processing, the problem of estimating the parameters of multiple sources emitting signals that are received is addressed. There are several estimators proposed in the literature, but in this chapter we focus on the Maximum Likelihood (ML) estimator. The method was pioneered by R. A. Fisher in 1922. The maximum likelihood estimator (MLE) selects the parameter value which gives the observed data the largest possible probability density in the absence of prior information, although the latter can be easily incorporated to make the analysis fully Bayesian. For small numbers of samples, the bias of maximum likelihood estimators can be substantial, but for fairly weak regularity conditions it can be considered asymptotically optimal (Mackay, 2003). Thus, for large samples the MLE has been shown to achieve the Cramér-Rao lower bound and yield asymptotically efficient parameter estimates (Stoica & Nehorai, 1990). With large numbers of data points (of the order of 10⁹ contrasted to 10⁶ estimated parameters), as in the case of the LOFAR EoR KSP, the bias of the method tends to be very small (Central Limit Theorem).

In this chapter, we assess how accurate a regularized ML inversion of the calibrated uv data-set can reconstruct the sky intensity as function of frequency and how accurate the EoR signal can be retrieved from the residual (foreground-subtracted) data-cube.

In general, it is not feasible to estimate the size of the data needed in order to obtain a good enough level of approximation of the likelihood function to a multivariate Gaussian. We assume that errors associated with each are independent, but they have different variances and memory of previous values over time and/or frequency. Many of these problems can be overcome by Markov Chain Monte Carlo (MCMC) or nested sampling of the posterior (Skilling, 2004; Feroz et al., 2009). This process could even be
iterated, but we expect convergence after a couple of iterations, since we start with an already good approximation of the model parameters after reprocessing (Chapter 3). With the addition of extra regularization terms, a deconvolved image can be obtained, from which the foreground and point-sources can be subtracted or filtered, to leave only the EoR signal and the noise.

4.2 Map-making

All the planned EoR detection experiments are characterized by a large volume of data, produced by a new generation of radio arrays. The data requires elaborate analysis, which can be done by following these basic steps: (i) produce maps from the time-ordered visibilities after correcting for the instrumental errors (Chapter 3), (ii) combine the maps into 4/5D (three spatial dimensions plus frequency and polarization) data cubes, (iii) remove the foregrounds and finally (iv) study the residuals with the goal of extraction information about the reionization process. In this chapter, we focus mostly on the map-making step rather than on the primary calibration. After the calibration and provided that it performs in a satisfactory manner, the data model takes the form of a set of linear equations. The traditional approach to solve such a problem is through a brute force, direct matrix inversion. However this is not feasible in the context of modern arrays as the number of data-points is huge (i.e. of the order of $10^9$ visibilities per frequency channel and per station beam for LOFAR) and the complexity of the inversion operation is $O(N_{\text{vis}}^3)$. Even if we had the required computing power to perform a direct inversion, the numerical errors of the computation would magnify the noise of the produced maps by a large factor as we will demonstrate in a following section. We can circumvent this in a memory efficient way by an iterative algorithm. In this section we discuss the underlying mathematical principles of map-making and in the subsequent sections we specialize in the case of the LOFAR EoR experiment.

In this section we shall present three flavors of likelihood estimators. The Least Squares estimator (LSE) (i) has minimum variance amongst all linear unbiased estimators of a given parameter and is known as the best linear unbiased estimator (BLUE). If the random variables have a normal distribution, then the Least Squares estimator of that parameter is the Maximum Likelihood estimator, has a normal distribution and is the (ii) MVUE (Minimum Variance Unbiased Estimator). However, this might not always be the case. It is known that when a transformation is applied to one parameter (i.e. square root), then the estimator of that parameter need not be unbiased. This is important because e.g. estimating the real and imaginary part of a parameter is not equivalent to estimating the amplitude and phase. The mean-square error (MSE) of an MVUE estimator $\hat{x}$ is given by $\text{var}(\hat{x}) + \text{bias}(\hat{x})$. A biased estimator can have lower MSE because its variance can be significantly lower. Finally, we discuss the Asymptotic Likelihood estimator (ALE) (iii) which deals with statistical inference as the sample size approaches infinity. EoR experiments plan to increase the total integration time and consequently the number of samples. It is thus important to take the asymptotic properties of parameter estimation into account. The ALE is shown (Plackett, 1950) to be asymptotically Normal, unbiased and have the minimum asymptotic variance.

We shall start with the Best Linear Unbiased estimator. This estimator has by defini-
tion the minimum variance among all unbiased linear estimators estimator of a parameter. We also discuss the minimum variance unbiased estimator

The observed visibilities can be written in the form of the following narrowband model:

\[ V_{\text{obs}} = A(p) \mathbf{v} A^H(p) + \sigma_w^2 \mathbf{I}, \]  

(4.1)

where the \( 2 \times 2 \) matrix \( A(p) = G(t,f) \prod_m J_m(p) \) is the ordered product of the Jones matrices of the uv plane (\( G \)) and image plane (\( J_m \)) effects (Chapter 2), \( \mathbf{I} \) is the identity matrix, \( \mathbf{v} \) are the true underlying visibilities of the sky and \( \sigma_w^2 \) is the standard deviation of spatially-white noise. It is therefore logical to cast the relationship between the sky brightness distribution (or its Fourier transform) and the observed visibilities as a linear algebra system. This is achieved by using the column-wise Khatri-Rao product (\( \otimes \)) and we thus get (Boonstra, 2005):

\[ \hat{V}_{\text{obs}} = \left[ A^T(p) \otimes A(p) \right] \vec{\text{diag}}(\mathbf{v}) + \vec{\text{diag}}(\sigma_w^2 \mathbf{I}). \]  

(4.2)

The optimal map-making is essentially the estimation of the map \( \mathbf{v} \) given our measured data \( \hat{V}_{\text{obs}} \) and the properties of the noise. To be statistically optimal, the map choice has to maximize the posterior probability of a deduced set of map parameters which is

\[
P(\mathbf{v}|\hat{V}_{\text{obs}}) = \frac{P(\hat{V}_{\text{obs}}|\mathbf{v}) P(\mathbf{v})}{P(\hat{V}_{\text{obs}})}, \quad \text{or posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{norm. constant}}.
\]

Without assuming a prior explicitly, \( \mathbf{v} \) follows a uniform distribution, which means that the ratio \( \frac{P(\mathbf{v})}{P(\hat{V}_{\text{obs}})} \) is constant. There are four cases that we must distinguish: (i) the noise level is known and Gaussian, (ii) the noise level is unknown and Gaussian, (iii) the noise level is unknown and non-Gaussian and finally (iv) that the noise is known and non-Gaussian. We shall concentrate on the Gaussian noise case with either known or unknown noise level. This is because the amount of data is so large that the central-limit theorem justifies this choice and each visibility can be assumed to have a Gaussian error on its complex and imaginary parts (Thompson et al., 2001). In this case the parameter probability density function (PDF) approaches Gaussianity as well. Moreover, it is also safe to assume that visibilities are independent samples of the sky power spectrum and that there is scale mixing introduced due to the convolution of the visibilities with the image plane effects. For this we ignore the effects of RFI and mutual coupling, which affect short baselines more. We can thus write:

\[
P(\mathbf{v}|\hat{V}_{\text{obs}}) \propto P(\hat{V}_{\text{obs}}|\mathbf{v}) = \exp\left(-\frac{1}{2} (\hat{V}_{\text{obs}} - A(p) \mathbf{v})^T C_N^{-1} (\hat{V}_{\text{obs}} - A(p) \mathbf{v})\right),
\]

(4.3)

which is the likelihood function in the case of Gaussian noise and also the posterior in case of a flat prior on the model parameters. The normalizing constant is given by the inverse of \( P(\hat{V}_{\text{obs}}) = \sum_k P(\hat{V}_{\text{obs}}|\mathbf{v}_k) P(\mathbf{v}_k) \). This means that the best (minimum variance) unbiased (in the sense of the minimal difference between the expectation of the estimator
and then true parameter being estimated) estimator of $v$ is the one that maximizes the above probability function given the noise covariance matrix $C_N^{-1}$ and the data model equation. We should remark that this approach implies implicit belief in the data model. However, this might not always be reasonable. For example, the number of the Jones matrices is deduced from our knowledge of the physical processes along the signal path (Hamaker, 1999). Those processes are not always easily identifiable or independent i.e. the number of operational elements affect both the beam-shape and the Point Spread Function (PSF). If the matrix $A$ (we drop the explicit dependence of $A$ on the calibration parameters $p$ from hereon) has full-rank then the best linear unbiased estimator (BLUE) is equivalent to the least-square solution (Feller, 1968) and has the form:

$$\hat{V}_{BLUE} = \left[ A^T C_N^{-1} A \right]^{-1} A^T C_N^{-1}.$$ (4.4)

This is the Gauss-Markov theorem (Plackett, 1950). In the case where the noise between all telescopes is the same, we arrive at the solution of a weighted, deterministic least-squares problem. The inverse of the noise covariance matrix plays the role of a metric in the multi-dimensional parameter space. The second factor of the estimator gives the dirty “map” and the first factor which is enclosed in square brackets, is the deconvolution step (Boonstra, 2005). It can be viewed as an inversion of a beam-forming operation that weighs the dirty map parameters according to the instrumental corruptions. The CLEAN algorithm attempts to solve the deconvolution problem by iteratively removing models of point-sources from the measured data. Its computational expense scales as the number of map elements. However CLEANing yields less optimal solutions (e.g. Starck et al. 2002).

We could incorporate more priors to the inversion process. For example if we assume a Gaussian prior for the underlying data (Zaroubi et al., 1995) then:

$$\hat{V}_{BLUE} = \left[ C_v^{-1} + A^T C_N^{-1} A \right]^{-1} A^T C_N^{-1}.$$ (4.5)

We can get this solution by further Wiener filtering of the Maximum Likelihood solution without priors as well (Legmark & Efstathiou, 1996; Bouchet et al., 1998). Note that the Wiener filtering is related to the Tikhonov regularization for a certain selection of weights (Kitaura & Enßlin, 2008).

In light of the above discussion, the BLUE is unsuitable for the solution of noisy, ill-posed conditioned systems such as the one we are dealing with. To remedy this, we assume that the signal is a realization of a Gaussian random vector. What follows can be viewed as a linear least-squares analog of Bayesian estimation. Suppose that both $\hat{V}_{obs}$ and $v$ are jointly distributed Gaussian random vectors whose source spectral matrix $S \triangleq E [y_v y_v^H]$ is unknown and whose components have finite expected squares. $y_v$ is the output of the array element for a given frequency $v$. The log-likelihood function is (Trees, 2002):

$$\mathcal{L}(p, v) \sim -\ln \det V_{obs} - \frac{1}{N_{data}} \sum_{N_{data}} Q^T V_{obs}^{-1} Q.$$ (4.6)

$^2$Even in the presence of a large number of data, rows of the array response vector can be dependent and moreover the difference in magnitude of the data can be very large.
where $Q$ is the measured electric signals of each station. The estimation problem requires to maximize the above equation with respect to the variance of the noise, the source parameters and measured coherencies. The minimization problem is greatly simplified by employing the result of Bohme (1984). The maximizing arguments are presented explicitly in that paper and it is important to note that those expressions reduce the dimensionality of the estimation problem by $n^2 + 1$. The drawback is that this method does not consider Hermitian positive (semi)-definite matrices as it should.

The assumption of the spatially-white Gaussian noise still holds. In this case the Asymptotic Likelihood Estimator (ALE) is obtained by maximization of the above likelihood function [Trees 2002]:

$$\hat{V}_{ALE} = \left[ A^T (p) A (p) \right]^{-1} A^T \left[ \hat{V}_{obs} - \sigma^2_w I \right] A \left[ A^T (p) A (p) \right]^{-1},$$

(4.7)

where $(A^T A)^{-1} A^T$ is the Moore-Penrose pseudo-inverse. If the number of map elements is too large the Moore-Penrose pseudo-inverse is singular and thus a high resolution image of a crowded field cannot be constructed. From the above equation we also derive the AML estimator for the calibration parameters. They have been derived in [Trees 2002]:

$$\hat{\sigma}_w^2 = \frac{1}{m - n} Tr \left[ \left( I - AA^T A \right) \hat{V}_{obs} \right]_{\hat{p} = \hat{p}},$$

$$\hat{p} = \arg \min_p \left( A \hat{V}_{obs} A^T + \hat{\sigma}^2 I \right)_{\sigma = \hat{\sigma}},$$

where $\hat{\sigma}$ and $\hat{p}$ are the estimated variance of the noise and parameter vector respectively. $\text{Tr}$ denotes the trace of a matrix, $n$ is the number of sources (pixels) and $m$ is the number of data points. If we assume that the noise is zero-mean (not necessarily Gaussian) and that it is independent from the signal, then we obtain the expression of the Minimum Variance linear estimator (MVAR):

$$\hat{V}_{MVAR} = C_v A^T \left[ A^T (p) C_v A (p) + C_N^{-1} \right]^{-1}.$$

(4.8)

When the covariance of the signal is non-singular then this is the Maximum a Posteriori (MAP) estimator, which is the Bayesian equivalent of the Wiener filtering approach that we discussed earlier. We reiterate that in the presence of Gaussian errors all the above estimators are equivalent. In this case, we show that the local MSE approaches the Cramér-Rao bound in the asymptotic regime and thus justifies the approach of currently planned EoR detection experiments to increase the number of measured visibilities by long observation runs and large number of baselines. There are many other approaches that can be used for the estimation of ML solutions like the Expectation Maximization and Space Alternating Generalized Expectation-maximization algorithm [Yatawatta et al. 2008, Kazemi et al., in prep.], and the spatial ARMA process (see discussion in Chapter 3) and each of them has different accuracy and convergence properties. However, what we must focus our attention on is that any estimator has to perform satisfactorily and be computationally efficient. In this chapter we concentrate on the accuracy of the estimation. The computational issues are discussed in chapter 5. Geometrically, the signal subspace should be independent from the noise sub-space. In the presence of noise and errors on the calibration parameters this independence is not always achieved. We discuss those issues in the following sections.
Table 4.1: Simulation parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinates of the 3C196 field centre (J2000.0)</td>
<td>α = 8h13m36s.0, δ = 48°13′03″</td>
</tr>
<tr>
<td>Galactic coordinates</td>
<td>l ≃ 171°, b ≃ 33°</td>
</tr>
<tr>
<td>Number of spectral bands</td>
<td>128</td>
</tr>
<tr>
<td>Frequency coverage (MHz)</td>
<td>120 - 184</td>
</tr>
<tr>
<td>Width of each band (MHz)</td>
<td>1</td>
</tr>
<tr>
<td>Frequency resolution (MHz)</td>
<td>0.5</td>
</tr>
<tr>
<td>Time resolution (sec)</td>
<td>30</td>
</tr>
<tr>
<td>FoV</td>
<td>~ 10 deg.</td>
</tr>
<tr>
<td>Noise at 150 MHz (mK)</td>
<td>840 mK</td>
</tr>
<tr>
<td>Obs. duration (hrs)</td>
<td>4</td>
</tr>
</tbody>
</table>

4.3 Sky realizations

The sky realization used in this work is created by using the ChopChop simulation pipeline described in Chapter 3. The cosmological EoR signal is modeled using the 200 Mpc² squared simulation produced by Thomas & Zaroubi (2008). For the foregrounds, we use a polarized model described in Jelić et al. (2008). Except for the diffuse emission we include 350 resolved and approximately 20,000 unresolved and unpolarized point-sources which contribute to the confusion noise. The data model includes a set of corruptions, namely station gains which are modeled as Autoregressive Moving Average (ARMA) processes whose parameters are estimated from real LOFAR solutions, a 3D ionospheric model with wedges and turbulence, as well as an analytical model for the station beam taking into account the polarization distortion due to projection (see Chapter 3). However, we currently ignore cross-talk and mutual coupling between the elements. Mutual coupling tends to introduce correlations between the receiver noise of two elements and has to be corrected very well at the station level. However, here we are dealing with station correlations and the stations are much further apart than the individual elements. The simulation is described in detail in Chapter 3. A summary of the “mock” LOFAR-EoR observation specifications is given in table 4.1.

4.4 Pixel size, choice of domain and regularization

A linear problem is said to be well-posed when the following two conditions hold: (ia) for every observed datum there is a solution or (ib) the solution is unique and (ii) the solution is stable under perturbations. If any of these conditions do not hold the problem is ill-posed. The condition number is a measure of how well-posed a system is. It is defined as the ratio between the largest and smallest eigenvalues of the inversion matrix:

\[ \kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \]  

(4.9)
Physically, it is a measure of the "magnification" of perturbations due to numerical or systematic errors and thus it is an intrinsic property of the problem even in the case where the noise is insignificant. There are some profound implications that result from this. If the problem is ill-posed, additional constraints are needed in order to obtain a proper solution. There are numerous types of extra constraints that can be applied and depending on the type or the strength with which they are applied, they can give different results. Trying to estimate more parameters than the data allows, only makes the ill-posed problem worse. We shall shortly discuss the method of regularization and then return to the discussion of the issues raised above. Regularization is commonly used to remedy the uninvertibility of matrices in linear systems. That said, there are certain caveats that one must be careful of. If the dynamic range of the image is very large (i.e. there are very bright sources together with very faint ones), the weakest sources can be affected by the artifacts arising from the imperfectly removed dirty beams of the strongest sources (Fig 4.3). We shall demonstrate that shortly.

The regularized solution for the BLUE estimator (all the above mentioned ML estimators are statistically equivalent in the presence of Gaussian errors) that we employ in this Chapter and the next is given by

$$\hat{V}_{\text{BLUE}} = \left[ A^T C_N^{-1} A + \alpha R^T R \right] A^T C_N^{-1}$$

(4.10)

where $\alpha > 0$ is the regularization parameter and the positive-definite matrix $R^T R$ is the penalty operator. Since $A$ is non-linear one can replace it with the Fréchet derivative. For a non-quadratic penalty functional, $R^T R$ can be replaced by the Hessian or another linearization. The standard Tikhonov penalty functional (Tychonoff, 1943), for $R = I$, is then given by

$$\frac{1}{2} \| \mathbf{v} \|^2$$

(4.11)

When the regularization parameter is small, filtering of the noise is inadequate and the solution is highly oscillatory. On the other hand, when the regularization parameter is large, the noise components are partially filtered out. In that case though, components of the solution are suppressed too. In an image with very high dynamic range this will suppress very bright point-sources, and thus leads to results of questionable fidelity. Hence, if the dynamic range can be improved (e.g. through accurate point-source subtraction), better regularized solutions can be obtained.

There are several ways of selecting the regularization parameter. In the realm of simulations one can apply different methods and conduct a Monte Carlo study based on the simulated data. However, a more systematic approach is necessary when dealing with data. Our analysis is based on the analysis of the regularized solution total error. Under the assumption of a known data-model, we determine the attributes and the asymptotic rate of convergence of the regularized solution. The regularized solution error is given by:

$$e_{\alpha} = V_{\text{obs,} \alpha} - V_{\text{true,} \alpha} = e_{\alpha}^{\text{trunc}} + e_{\alpha}^{\text{noise}}$$

The above definition states that the error between the regularized solution and the true data consists of the truncation error $e_{\alpha}^{\text{trunc}}$, which describes the loss of information of
Figure 4.1: Foreground maps generated using different regularization methods and resolutions. The top row shows the reconstruction using diffusion operators and the bottom row using Tikhonov functionals. The maps at the left column are at a resolution of $80 \times 80$ pixels and on the right at $160 \times 160$ pixels:
4.4 Pixel size, choice of domain and regularization

Figure 4.2: The same map as Figure 4.1 but for an image size of 240 × 240 pixels and 320 × 320 respectively.

Figure 4.3: The Tikhonov method over-corrects around the position of bright point-sources as it tries to suppress the oscillatory points in the image. In this image the difference between the true map and the map obtain through the ML inversion is shown. There are clear artifacts shown at the positions of point-sources.
Figure 4.4: The choice of the regularization parameter as a function of frequency. We see that the parameter is relatively stable.

the solution due to regularization and the noise amplification error $e_{\text{noise}}^a$. Our aim is to choose a parameter such that both errors converge to zero as the noise level (norm of the noise) tends to zero. Using the Singular Value Decomposition (see also Chapter 5) of the regularized solution one can prove that the noise error is less or equal than $\sqrt{a^{-1}} \delta$, where $\delta = \|n\|$ is the error level and $n$ is the noise vector. Similarly the truncation error converges to zero as $a$ tends to zero. This means that we can obtain a map with minimal errors, if we choose $a = \delta^m$, with $0 < m < 2$.

This is an a priori regularization parameter selection rule in the sense that it depends on information which we already know about the solution. An a posteriori selection of a parameter depends only on the data. One such selection method arises from the Morozov discrepancy principle (Anzengruber & Ramlau, 2010), which selects the largest regularization parameter $a$, for which $\|Av^a - V_{\text{obs}}\| \leq \delta$. In this case $a = \delta / \|V_{\text{obs}}\|$. 

The other regularization method that we will consider in this paper is a method that penalizes non-smooth solutions and can be viewed as a generalization of Tikhonov regularization. For very large and ill-conditioned systems such as those arising during the processing of interferometric observations, it is often impractical to implement regularization by filtering as this requires the computation of the decomposition of a very large matrix. However, the Tikhonov solution can be written in a variational form with respect to $v$:

$$v^a = \arg \min \|Av^a - V_{\text{obs}}\|^2 + \alpha P,$$

where $P \equiv v^T R^T R v$ is a penalty functional. Penalty functionals can also incorporate information about the solution i.e. the non-negativity of intensity in the imaging problem. In this chapter we consider penalty operators $\hat{L}$ of the diffusion (Vogel & Oman, 1996) type:

$$\hat{L} = - \sum_i \frac{\partial}{\partial p_i} \left( \rho \frac{\partial}{\partial p_i} \right),$$

(4.13)
4.4 Pixel size, choice of domain and regularization

where $\rho$ is the local curvature, so that

$$P = \langle \tilde{L}v, v \rangle.$$

Diffusion operators can be viewed as a local averaging operators. They depend on the local, fine-scale geometry of the data and introduce a notion of smoothness on the data. In chapter 5 we will discuss an iterative regularization algorithm for applying diffusion-type regularization, that is a generalization of total variation regularization that avoids the need of line searches which arises in the application of other iterative regularization methods. The benefit of this algorithm is that it is easy to implement on hardware accelerators. However, iterative methods are generally impractical for large-scale problems.

For a more detailed discussion of regularization we refer the reader to the tutorial by Neumaier (1998).

Figures 4.1 and 4.2 show the comparison of the Tikhonov and diffusion regularization methods for final maps for different map resolutions. We see that Tikhonov regularization is particularly harsh on suppressing extreme values. The MLE is performed on the uv-plane and we present the direct Fourier transform of the visibilities (dirty map). To get statistically optimal results the dynamic range of the data has to be 100:1 or larger. In Figure 4.4 we plot the optimal regularization parameter $\alpha$ for each map as a function of frequency.

From the above discussion we see that working on the Fourier (uv) domain is more convenient as the condition number of the map-making matrix (in this case the map refers to the true underlying visibilities). The matrices are smaller and the number of independents measurements is roughly given by the maximum baseline divided by the station diameter, which for the LOFAR EoR experiment is of the order of $80 \times 80$. Those parameters are constrained by $10^9$ visibilities and thus we have an over-constrained
Optimal data-model inversion problem. Figure 4.5 shows the change of the condition number of the deconvolution matrix as a function of the number of the map elements for each linear dimension. Each visibility contributes to the value of a single pixel in the image domain as it is described through the deconvolution matrix. The foreground removal performs better and there are no deconvolution artifacts (i.e. side-lobe confusion noise) when it is done in the uv plane, since there is no spatial scale error introduced due to different uv sampling for every frequency (Harker et al., 2009a).

4.5 Results

In this section we discuss the results from the data inversion procedure. The sky realization that we use is a realization of the LOFAR 3C196 field (see Section 4.3). This field is centered on the bright (75 Jy at 138 MHz) quasar 3C 196 and is a very useful field to study for advancing our understanding of LOFAR. It is located at a moderate declination of 48 degrees and is populated by approximately 20 bright point sources of the order of a Jansky as well as hundreds of fainter ones. This field has also been studied extensively by the Westerbork Synthesis Radio telescope (Bernardi et al., 2010), which is one of the most stable interferometers available.

4.5.1 Properties of the noise

We mentioned that the errors affect the separation of the noise and signal subspace. This is particularly important for the calibration step because the successful estimation of the relevant maximum likelihood estimator depends on the ability to separate those two subspaces. For a more thorough discussion we refer the reader to Trees (2002) and Stoica & Nehorai (1990). We also used several assumptions for the noise in constructing the estimators of Section 4.2. It is therefore imperative to discuss the noise properties before continuing. Ideally, the noise is Gaussian and spatially-white and each telescope, as well as each snapshot, should have independent noise. When we refer to noise, we mean the effective noise (Wijnholds, 2010), which is a combination of calibration residuals, the confusion noise and the thermal noise. We shall discuss each of these contributions shortly. The noise of the dirty map is equal to the average thermal noise per baseline. The map-making process involves a deconvolution step which is effectively weighing and re-arranging of the data. This essentially means that the noise on the image is pixel-correlated. Even in the simplified case where all the stations behave in the same manner and thus have identical gains and beams, the noise on the image depends on the row sums of $\left(\tilde{A}^T \tilde{A} \odot A^T A\right)^{-1}$, where $\odot$ stands for the Hadamard product (Wijnholds, 2010). If this product is not constant (i.e. in the case of non-symmetric uv coverage), the noise is distributed in a more complex way.

The thermal noise is given by the radiometer equation and scales as the inverse of the number of data points, $\sigma_{\text{img}} = \frac{T_{\text{sys}}}{\sqrt{N_{\text{tel}}(N_{\text{tel}}-1)\Delta \tau \text{BW}}}$, where $\Delta \tau$ is the averaging time and BW is the bandwidth. Increasing any of the parameters in the denominator decreases the noise, at least theoretically. If the number of sources is large enough, at some point the source distribution will occupy the whole image and the map starts to resemble diffuse
Figure 4.6: A map of the thermal noise at 150 MHz for a field at declination 48 and an integration time of 300 hours.

Figure 4.7: The confusion noise resulting from a crowded field with 60,000 sources with fluxes between 1 and $10^{-6}$ Jansky. Bright sources have been removed to the noise level from the uv data. The map is shown at 150 MHz and corresponds to an observation with the LOFAR core.
emission. There is a maximum number of separable sources, given a certain map reso-

lution, and this sets the classical confusion limit (Condon [1974]). Another complication

is that if the number of pixels is large and each pixel is occupied by a point-source, then

the problem becomes ill-posed as for each source we need to estimated its flux and po-

sition inside the pixel. Since the LOFAR EoR experiment, will use around 40 to 48 High

Band Antenna stations during the EoR experiment the maximum number of sources

that the LOFAR core can separate is approximately 500. This is because an array with

$N_{tel}$ elements provides $N_{tel}^2$ correlations. Each source can be described with four param-

eters (intensity and position (in l,m,n,) coordinates) relative to the central source and

also the noise power needs $N_{tel}$ parameters (Wijnholds, 2010). Using longer baselines

within the Netherlands the number can be increased to 6200 to 8200 sources. The equiv-

alent theoretical limit for the maximum number of sources that the MWA can separate is

approximately 250,000! However, these computations need to be adjusted for the redun-

dancy of the array configuration. Redundant baselines translate into linearly dependent

measurements and therefore linearly dependent rows in the deconvolution matrix.

The covariance of the calibration residuals on the image is given by:

$$\text{cov}(v) = \left(\frac{\partial v}{\partial p}\right) C_p \left(\frac{\partial v}{\partial p}\right)^T,$$

(4.14)

where $C_p$ is the covariance of the calibration parameters. We need to estimate the partial
derivative with respect to the calibration parameters and this can be done analytically in
the case of Gaussian noise. The analytical equations are derived in Wijnholds, 2009:

$$\frac{\partial v}{\partial \theta^T} = -2D^{-1} \text{Im} \left\{ J^T \odot J G^2 J C_N J^T \right\} G^2,$$

$$\frac{\partial v}{\partial \sigma^T} = -D^{-1} \left( J^T J \odot J^T J \right)^{-1} I,$$

$$\frac{\partial v}{\partial g^T} = 2D^{-1} \text{Re} \left\{ J^T \odot J G^2 J C_N J^T \right\} (2I - G) G,$$

(4.15)

where $G$ is the uv-plane effect Jones matrix, $D \equiv [J^T G^2 J] \odot [J^T G^2 J]$ is the decon-

volution matrix, $\theta$ are the image-plane array response vector parameters, $g$ are the uv-plane
parameters and $\sigma$ is the noise variance. The bar above a matrix denotes complex conju-
gation and $\odot$ the Hadamard product (element-wise multiplication).

Calibration residual errors follow the beam-forming redistribution and thus the pat-
terns of the dirty map. Moreover, the image covariance decreases with an increase in the
number of data points in a way similar to the thermal noise, provided that the estimator
is unbiased. In the LOFAR EoR experiment we are going to observe the same windows
of the sky for about 100 nights. In that case we will be able to determine the PDF of the
parameter vector $p$. We know that the PDF of the gains is non-Gaussian and we expect a
similar behavior for the ionospheric phase and the beam error to some extent. In this case
we have to calculate numerically the global Cramér-Rao bound that does not depend on
the specific measurements. The analytical derivation of such a bound is not possible.

4.5.2 Correlation between the original and the recovered maps

As an initial test we study the correlation of the map obtained via the Fourier transform
of the corrected visibilities with the dirty map of the true model that we used for the
simulation. In Figure 4.8 we present the correlation coefficient of those maps as a func-
tion of frequency. We see that the plot is not trended, which means that the inversion
process does not introduce frequency dependent errors. We also see that the corre-
lation coefficient is relatively high (\( \sim 0.86 \)), which means that the inversion performs well.
It also has a low change from channel to channel which implies that the foregrounds
are probably smooth along the frequency direction. We then compare the correlation
between two adjacent channels in the true cube and in the reconstructed cube. A funda-
mentl assumption in the EoR signal extraction method introduced by Jelić et al. (2008);
Harker et al. (2009a) is that the foregrounds are strongly correlated in adjacent channels.
In Figure 4.9 we show that the residual calibration errors do not affect the validity of this
assumption.

In Fig. 4.9 we show the correlation of adjacent frequency channels for the original
and reconstructed maps. We also plot the Kullback-Leibler divergence (KLD), which is
defined as:

\[
D_{KL} = \sum_k I_{\nu}(k) \log \frac{I_{\nu}(k)}{I_{\nu+\Delta\nu}(k)},
\]

where \( I_{\nu}(k) \) is the intensity histogram of the map for a given frequency \( \nu \). The KLD is
a measure of the information distance between two models and is a measure of the in-
formation lost when assuming that due to strong correlation of the sky signal, adjacent
channels carry almost the same information. To calculate the KLD we estimate the boot-
strap PDF for each map at a given frequency and we generate 100 maps in Monte Carlo
sense that follow that PDF. Both the KLD and the correlation between adjacent channels
drop by a small amount in the reconstructed maps and this indicates that the inversion
step is leaving some chromatic error on the final maps. However, this error is rather
random and does not affect the signal extraction significantly.

4.5.3 EoR signal extraction

The ultimate benchmark of the EoR experiment is set by the ability to extract the cos-
smological HI signal. In this section we perform a simple extraction of the signal using a
polynomial fitting method in the uv-plane (Harker et al., 2009a). To do that we need to
select scales on the uv-plane that are sampled for a large fraction of the available observa-
tion bandwidth and with high SNR. Figure 4.10 shows the mask that we use for selecting
visibilities. We achieve the best fit with a polynomial of 7th or 8th order which is higher
than what Harker et al. (2009b) and Jelić et al. (2008) suggest. This is due to the higher
effective noise level that is present in the current simulations. In the aforementioned
work the authors did not include the effects of calibration residual or the confusion noise
and they assumed that the properties of the noise are known to a very high precision. In
our case that does not hold true. The effective noise has different components that are
affected in different ways by the data model and they also have different spatial, tem-
poral and frequency behaviors. One way to quantify the noise properties is to use the
difference of successive narrow-channel data. If the frequency resolution is high enough
we can assume that the astrophysical signal does not change significantly and that the
difference is purely determined by the noise. However, it is unfeasible to generate sim-
ulated data at such a high frequency resolution. To remedy this, we estimate the noise
for each of the 128 channels of the simulation using the true underlying maps. We note
that for the real observations, we will obtain very narrow frequency channels (10 kHz)
from which we can accurately assess the noise level as function of frequency, time and
baseline. We then estimate the PDF of the effective noise using bootstrapping. The result
is shown in Figure 4.11.

4.5.4 Cramér-Rao lower bound and the power-spectrum

In order to demonstrate the potential gain in estimation accuracy of the ML estimator
we need to evaluate numerically the asymptotic covariance. Equation 4.5.1 provides the
derivatives of the observed visibilities. In order to calculate the CRB, which states that the
variance of an unbiased estimator $\hat{\theta}$ is bounded by the inverse of the Fisher information
matrix (FIM) $F(\theta)$, the FIM has to be computed. The elements of the inverse CRB that is
the FIM are given by the Bangs formula (Stoica & Moses, 2005):

$$
F = \left( \frac{\partial \text{vec} \hat{V}_{obs}}{\partial \mathbf{m}} \right)^T \left( \hat{\mathbf{V}}_{obs} \otimes \hat{\mathbf{V}}_{obs} \right) \left( \frac{\partial \text{vec} \hat{V}_{obs}}{\partial \mathbf{m}} \right)
$$

$$
\mathbf{m} = \begin{bmatrix} g & \theta & \sigma \end{bmatrix},
$$

where $\mathbf{m}$ is the instrumental parameter vector and $g, \theta$ and $\sigma$ describe uv-plane, image-
plane effects and noise. The CRB is obtained by applying block-matrix inversion identi-
ities:
Figure 4.9: The correlation coefficient between the adjacent channels of the original (top) and reconstructed (bottom) dirty maps. The correlation coefficient is reduced after reconstruction and the same pattern can be seen for the Kullback-Leibler divergence.
Figure 4.10: The sampling of the uv-plane by the LOFAR core along frequency after 6 hours of synthesis. The top set of figures show the average number of visibilities per uv cell for 8, 16, 32 and 64 MHz of total bandwidth (the instantaneous bandwidth of LOFAR is 48 MHz). We assume that the data are delivered at 0.1 MHz resolution. The color-bar shows the number of visibilities per grid point. The bottom set of figures shows the area in which less than 5% of the data along frequency is lost due to the scaling of the uv coverage with frequency, compared to the total bandwidth. The black points represent regions where the visibilities and their Fourier conjugates occupy the same place, while the grey points represent true visibility measurements. This distinction is made because the Fourier conjugates do not contribute to the SNR.
4.5 Results

Figure 4.11: Signal extraction using polynomials on the uv-plane. In the figure above the \textit{rms} as a function of frequency is shown. The red dots correspond to the measured \textit{rms} of the residuals after foreground subtraction. The green line shows the \textit{rms} of the effective noise measured by differencing adjacent channels. The blue line is the evolution of the \textit{rms} of the cosmic signal and the magenta line the recovered signal.

\[
\text{CRB} = F_{\theta\theta} - F_{\theta\varnothing} F_{\varnothing\varnothing}^{-1} F_{\varnothing\theta} \\
- \frac{\left( F_{\theta\theta} - F_{\theta\varnothing} F_{\varnothing\varnothing}^{-1} F_{\varnothing\varnothing} \right) \left( F_{\theta\varnothing} - F_{\theta\varnothing} F_{\varnothing\varnothing}^{-1} F_{\varnothing\varnothing} \right)^T}{F_{\varnothing\varnothing} - F_{\varnothing\varnothing} F_{\varnothing\varnothing}^{-1} F_{\varnothing\varnothing}} 
\]

The computation of $F$ seems straight-forward but in practice it is more complicated. The matrix is usually rank-deficient and thus a reparametrization might be needed.

In Figure 4.12 we present the variance of the noise and the CRB as a function of SNR. In the case of interferometry the SNR scales with the number of data points as $N^{-\frac{1}{2}}$. Thus we can move to the asymptotic convergence regime by either integrating more or increasing the number of stations. We have assumed that we observe for 4 hours per night using 0.5 MHz of bandwidth and 30 seconds of averaging. That means that within each integration time interval we accumulate $\sim 10^4$ visibilities per frequency. The blue line shows the CRB computed from the parameters used for the simulation of Chapter 3. We then estimated the standard deviation using a varying number of visibilities to generate maps. We see that we approach the CRB for a number of visibilities that approaches $\sim 10^9$, which corresponds to 400 hours of integration. However, we must raise attention to the following issues: In this work we have ignored RFI and mutual coupling. These effects can add coherent contamination to the observed visibility and that would result in a Fisher information matrix with less than full-rank. In that case the MLE estimator would not be unbiased, although it might exhibit the same asymptotic behavior. The effect is more prominent for dipoles and tile correlation that are relatively close to each other, but for stations that are separated by a large enough distance this should not be a problem. This issue must be addressed through new, more sophisticated simulations. A Monte-Carlo type simulation is also needed in order to calculate the global FIM.
4.6 Conclusions

The extraction of the EoR signal presents two fundamental difficulties: on the one hand the SNR is extremely low, even for the new generation of radio telescopes and on the other hand the problem of imaging is not well posed. Furthermore, data-processing is limited by the computational power available at a given time. The new generations of radio interferometers and especially the SKA are highlighting these issues. In this Chapter we presented a Maximum-Likelihood estimation framework for post-processing the visibility data after calibration and bright source removal have been performed. To stabilize the solution we used Tikhonov and diffusion penalty functionals. A regularization parameter of the order of $10^{-4}$ is needed to stabilize the data, and though it is relatively small it is indeed necessary to producible sensible maps. The advantage of such a method is that it provides the MLE of the visibilities after a matrix inversion. The method can also be formulated in an iterative manner, that in principle requires a small number of iterations. Iterative methods that require very large number of iterations are impractical due to their computational overhead, given the huge data sizes involved. The computational issues related to the inversion are discussed in Chapter 5.

Larger maps require also the processing of much larger data volumes. In the case of the LOFAR EoR experiment, the use of the LOFAR core means that the resolution requirements are moderate. Making maps or corrected uv datasets with 80-120 elements in each linear direction is not an issue. However, higher resolution images becoming less accurate because the condition number of the deconvolution matrix explodes rapidly.
4.6 Conclusions

with increasing resolution.

In the case of LOFAR Core imaging, the ML method performs well and the problem is still tractable. Unfortunately, that cannot be said for large, wide-field surveys with LOFAR and the SKA. The method performs well when the dynamic range of the data has been reduced to $10^{2-3}$, after the brightest sources have been removed. Thus, accurate calibration is important and one needs to estimate the calibration parameters to an accuracy of $10^{-3}$.

We have also shown that the MLE becomes asymptotically optimal. By extrapolating the points of Fig. 4.12 we can see that integrating $O(10)$ more with LOFAR, one can get optimal estimates. This should in principle be within the grasp of the future SKA EoR experiment.

The next step would be to use and adaptive grids for parts of the image plane with low and high SNR. In any case, working on the uv plane tends to yield more accurate results than when working on the image plane.

Another issue of importance is the use of long-baselines. With LOFAR one can use longer baselines and this has a number of advantages. Long baselines enable us to sample the sky at smaller higher spatial frequencies and thus resolve and remove many more point sources from the image. On top of that, the RFI, the ionosphere and other near-field nuisances are uncorrelated for separated stations. It is still rather unclear to what degrees these effects are affecting the LOFAR data but in any case the presence of longer baselines is a reassuring factor.

4.6.1 Future work.

The next step and forerunner of the LOFAR EoR analysis is a fully fledged simulation generating a mock LOFAR EoR data set corresponding to 100 nights of observing. Longer baselines and far-sidelobe contamination of the maps as well as the effects of RFI and the ionosphere at such scales need to be introduced as well. After generating this data set, it has to be fed blindly into the standard LOFAR-EoR pipeline and processed in the same way as the actual LOFAR observations. This can enable us to test the calibration pipeline, fine-tune its computational performance and test the various signal extraction strategies in a realistic setting. Given our current experience and the results of the previous section we foresee no show-stopper, but this has to be verified by the actual data. From such a data set we can extract the observed statistics of the calibration parameters and compare them with the input ones and moreover estimate the global FIM.

In this paper we used two regularization methods: Tikhonov and diffusion operators. The Tikhonov method requires an accurate selection of the regularization parameter $\alpha$ in order to avoid over-amplification of the noise. It is easy to fold in the uv-plane sampling function of the interferometer in the regularization process. On top of that, we can even use only the spatial scales that are sampled well for most of the frequencies over a finite bandwidth in order to avoid the spatial-frequency scale mixing introduced by a varying uv coverage. If $P_H$ is the projector to the sub-space of spatially band-limited functions (equivalent to selecting a mask (spatial bandpass filter) on the uv-plane) then, this can be incorporated via the constraint

$$(I - P_Hv) = 0.$$
By using the above constraint all the available visibilities are taken into account even if they are redundant.

Based on the above discussion on band-limited regularization a similar remark can be made for the diffusion method. The diffusion method preserves the relationship between smoothness and sparsity in a “Fourier” basis. In this scenario the result depends on both the properties and geometry of the data space as well as the ME. We can use properties of the ME (via eigenvectors, multi-scale basis etc) to modify the data in such a way that the diffusion operator is linear. In the future, we plan to investigate such a regularization problem. The problem can be formulated as a discretized diffusion partial differential equation where the discretization is based on a finite differences method that provides certain filtering properties, well-posedness, mean conservation etc. A different method would be a dual approach based on representations and basis functions.