Circuits and Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ens.ewi.tudelft.nl/

CAS-2019-4742001

M.Sc. Thesis

Bayesian Learning Applied to Radio Astronomy Image Formation

Yajie Tang

Abstract

Radio astronomy image formation can be treated as a linear inverse problem. However, due to physical limitations, this inverse problem is ill-posed. To overcome the ill-posedness, side information should be involved. Based on the sparsity assumption of the sky image, we involve ℓ_1-regularization. We formulate the image formation problem into a ℓ_1-regularized weighted least square (WLS) problem and associate each variable with one regularization parameter. We use Bayesian learning to learn the regularization parameters from data by maximizing the posterior density. With the iterative update of the regularization parameters, the solution is updated until convergence of the regularization parameters. We involve a stopping rule based on the noise level to improve the computational efficiency and control the sparsity of the solution. We compare the performance of this Bayesian learning method with other existing imaging methods by simulations. Finally, we propose some future research directions in improving the performance of this Bayesian learning method.

Bayesian Learning Applied to Radio Astronomy Image Formation

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Signals and Systems

by

Yajie Tang
born in Panzhihua, China

This work was performed in:

Circuits and Systems Group
Department of Microelectronics & Computer Engineering
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Delft University of Technology

Delft University of Technology
Department of
Microelectronics & Computer Engineering

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Bayesian Learning Applied to Radio Astronomy Image Formation"** by **Yajie Tang** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 15 October 2019

Chairman: _____
prof.dr.ir. A.J. van der Veen

Advisors: _____
dr.ir. S.

_____
dr.ir. A. Sardarabadi

Committee Members: _____
prof.dr.ir. G.J.T. Leus

_____
prof.dr.ir. M.H.G. Verhaegen

Abstract

Radio astronomy image formation can be treated as a linear inverse problem. However, due to physical limitations, this inverse problem is ill-posed. To overcome the ill-posedness, side information should be involved. Based on the sparsity assumption of the sky image, we involve ℓ_1-regularization. We formulate the image formation problem into a ℓ_1-regularized weighted least square (WLS) problem and associate each variable with one regularization parameter. We use Bayesian learning to learn the regularization parameters from data by maximizing the posterior density. With the iterative update of the regularization parameters, the solution is updated until convergence of the regularization parameters. We involve a stopping rule based on the noise level to improve the computational efficiency and control the sparsity of the solution. We compare the performance of this Bayesian learning method with other existing imaging methods by simulations. Finally, we propose some future research directions in improving the performance of this Bayesian learning method.

Acknowledgments

I would like to thank my supervisor prof.dr.ir. A.J. van der Veen for his guidance and encouragement throughout this thesis project. I would like to thank dr.ir. S. Naghibzadeh for the inspiration she gave me at the beginning of this thesis project and her suggestions when I was lost. I would like to thank dr.ir. A. Sardarabadi for his patient and selfless help when I was stuck in troubles. Without them, neither could I finish this thesis project nor could I know how scientific research should be. Last but not least, I would like to thank my parents for their continuous support.

Yajie Tang
Delft, The Netherlands
15 October 2019

Contents

List of Figures

List of Tables

1 Introduction

Everything in the world begins with the Big Bang around 13.8 billion years ago. Figure 1.1 shows the expansion process of the universe since the Big Bang. Astronomers are always expecting to unravel the history and depict the future of the universe. Engineers have been making effort to offer astronomers observational information to reveal the mysteries of the universe. The radio emissions in the universe carry much information of the early stage of the universe. To extract the information hidden in the radio emissions, a more accurate picture of the universe in radio frequencies is required, which promotes the development of more and more powerful radio telescopes. Along with the development of telescopes, radio interferometric imaging methods are continuously proposed and improved.

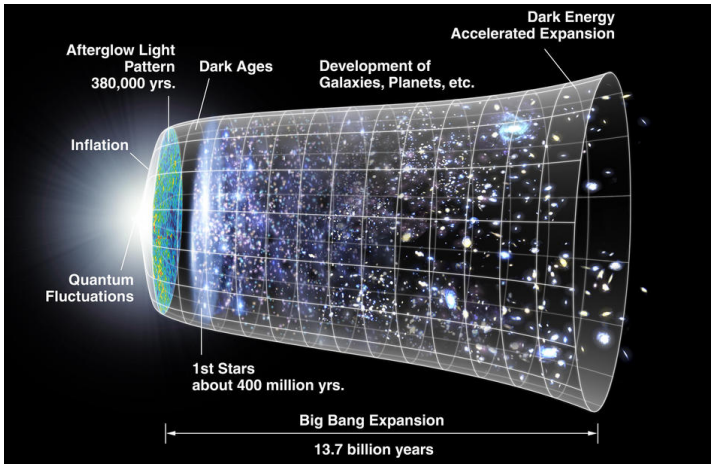


Figure 1.1: Big Bang and expansion of the universe (Image courtesy of NASA Jet Propulsion Laboratory)

In this introduction chapter, we introduce what is radio interferometric imaging. We also introduce some state-of-the-art methods of radio interferometric imaging. Based on the background of radio astronomy imaging and existing methods, we propose our research goals.

1.1 Radio Interferometric Imaging

The formation of images of the sky in radio frequencies are important to astronomers, since the images can provide much quantified information for studying the celestial objects and phenomenas. The interferometric imaging aims at estimating the intensity and Direction of Arrival (DoA) of the celestial signal. With the measured correlations,

this imaging problem can be transferred into an inverse problem: inverting the measured correlation data into the pixel intensities of the sky image. The relation between the intensities and correlations is stated in a linear equation as:

$$\tilde{\boldsymbol{r}} = \boldsymbol{M}\boldsymbol{\sigma} + \boldsymbol{e}, \tag{1.1}$$

where $\tilde{\boldsymbol{r}}$ denotes the measured correlation data, $\boldsymbol{\sigma}$ denotes the pixel intensities over the field of view (FoV), $\boldsymbol{M}$ denotes the system matrix and $\boldsymbol{e}$ denotes the noise. More explanations of this linear equation will be presented in the next chapter.

1.2 State-of-the-Art Imaging Algorithms

Existing radio astronomy imaging methods can be divided into three categories: (i) greedy sparse reconstruction methods such as CLEAN (ii) methods based on convex optimization such as MEM and SARA; (iii) projection-based algorithms such as PRIFIRA [1].

CLEAN is a classical method of deconvolution in radio astronomy [2]. It imposes a sparse prior by assuming that the radio sky is an almost empty field with a small number of point sources. These point sources are found iteratively and are subtracted from the residual map. The final result is the sum of the point "CLEAN" components reconvolved with a "CLEAN" beam. Variants with different strategy of searching the point sources have been developed [3]. Furthermore, variants have been developed to improve deconvolution results of extended structures [4].

Another common method is the maximum entropy method (MEM) [5]. The MEM regularizes the ill-posed problem via an entropic prior $\boldsymbol{\sigma}\log(\boldsymbol{\sigma})$. While CLEAN often struggles to reconstruct extended structures, MEM struggles to reconstruct point sources.

We look for a sparse solution of the radio astronomy imaging. The ℓ_0 constraint leads to a sparse solution due to the nonzero support it presents. This constraint can be weakened to an ℓ_1 constraint which still leads to a sparse solution but is much easier to solve due to the convexity of the ℓ_1-norm. A more general regularization form $\|\boldsymbol{\Psi}^T\boldsymbol{\sigma}\|_1$ can be adopted, where $\boldsymbol{\Psi}$ is an overcomplete dictionary. The sparsity averaging reweighted analysis (SARA) is a method defined in this framework [6]. The algorithm relies on the conjecture that cosmic signals are synchronously sparse in multiple bases, particularly the Dirac basis, wavelet bases, or in their gradient, so that promoting average signal sparsity over multiple wavelet bases presents an extremely powerful prior.

PRIor-conditioned Fast Iterative Radio Astronomy (PRIFIRA) is a new method that embodies the Bayesian-based regularization into the linear model via right preconditioning and solves the resulting system via projection onto Krylov subspaces [7]. The preconditioner can be beamforming-based estimation such as MVDR dirty image [8]. An outer loop can be defined to apply generalized reweighted prior-conditioning. This algorithm can well resolve diffuse and compact sources.

1.3 Research Goals and Tasks

Various imaging methods are being proposed which rely on accurate modeling of the radio sky to guarantee high reconstruction quality. Modeling of point sources is fairly easy while no straightforward model of the extended emissions can be presupposed. Bayesian learning opens the doors to adaptive modeling of the extended emissions where the model has to be learned from the data. Bayesian learning treats model parameters as random variables. In Bayesian learning, parameter estimation amounts to computing posterior distributions for these random variables based on the observed data. We would like to derive an automated accurate model and take into account the trade-off between the estimation accuracy and the computational complexity.

The tasks of this project can be divided into the following steps: (i) deriving our model based on Bayesian framework; (ii) designing the regularized imaging problem according to the derived model; (iii) designing an efficient method to solve the resulting problem; (iv) using simulations to compare our algorithm with state-of-the-art methods.

1.4 Outline

My thesis report is organized as follows. In Chapter 2, we introduce the interferometric measurement model that is widely used in astronomy, and we also employ the signal processing data model. In Chapter 3, based on the measurement model developed in the previous chapter, we analyze the ill-posedness of this problem and introduce some common formulations of this problem. In Chapter 4, we reformulate the problem and propose our Bayesian learning framework. We further compare our framework with PRIFIRA. In Chapter 5, we introduce the details of our Bayesian learning method. In Chapter 6, we compare our method with other existing methods by simulations. In Chapter 7, we draw conclusions of this project and present some future work.

2 Data Model

In this chapter, we introduce the data model of radio interferometric imaging. We start by introducing some important symbols, operators and matrix product relations, which are widely used in this chapter and the rest of this report. Then, we introduce the measurement model of radio astronomy. We first employ the conventional measurement model which astronomers are familiar with. Furthermore, we explain the radio interferometric imaging problem in terms of array processing and attain our array processing model which is the basis of the rest of this report.

2.1 Notations

We use similar representations to what Sardarabadi used in [9] and what Naghibzadeh used in [1]. The representations and corresponding interpretations are listed as follows.

2.1.1 Symbols

a, A: plain lowercase and uppercase letters denote scalars
$\mathcal{A}$: calligraphic letters denote continuous functions
$\boldsymbol{\mathcal{A}}$: calligraphic boldface letters denote operators
$\boldsymbol{a}$: boldface lowercase letters denote column vectors
$\boldsymbol{A}$: boldface uppercase letters denote matrices
$\boldsymbol{a}_i$: for a matrix $\boldsymbol{A}$ denotes the ith column of $\boldsymbol{A}$
$a_{i,j}/a_{ij}/A_{ij}$: for a matrix $\boldsymbol{A}$ denotes the i, jth entry
$\mathbf{1}$: vector with all the elements equal to 1
$\boldsymbol{I}$: identity matrix
$\boldsymbol{I}_P$: $P \times P$ identity matrix
$\boldsymbol{e}_i$: the ith column of the identity matrix
$\mathbf{0}$: vector with all the elements equal to 0
j: square root of -1
$\mathbb{R}$: blackboard bold letter R denotes the set of real numbers
$\mathbb{R}^{m\times n}$: denotes the set of real-values m by n arrays
$\mathbb{C}$: denotes the complex numbers
$\mathbb{C}^{m\times n}$: denotes the set of complex-values m by n arrays
$\mathscr{N}(\cdot)$: denotes the null space of a matrix
$\mathscr{R}(\cdot)$: denotes the range of a matrix

2.1.2 Opertators

$\mathrm{E}\{\cdot\}$: expectation operator
$\mathrm{Cov}\{\cdot\}$: covariance

$(\cdot)^T$: transpose operator
$(\cdot)^*$: complex conjugate operator
$(\cdot)^H$: Hermitian Transpose operator
$(\cdot)^{-1}$: inverse operator
$(\cdot)^\dagger$: Moore-Penrose psuedo-inverse operator
$\|\boldsymbol{a}\|_p$: p-norm of vector $\boldsymbol{a}$ defined as $\|\boldsymbol{a}\|_p^p = \sum |a_i|^p$
$\|\boldsymbol{A}\|_F$: Frobenius norm of matrix $\boldsymbol{A}$ defined as $\|\boldsymbol{A}\|_F = \sqrt{\sum \sum |a_{i,j}|^2}$
trace($\boldsymbol{A}$): computes matrix the sum of the diagonal elements of $\boldsymbol{A}$
det($\boldsymbol{A}$): the determinant of matrix $\boldsymbol{A}$
vect($\boldsymbol{A}$): stacks the columns of matrix $\boldsymbol{A}$ to form a vector
vectdiag($\boldsymbol{A}$): stacks the diagonal elements of matrix $\boldsymbol{A}$ to form a vector
diag($\boldsymbol{a}$): takes the vector $\boldsymbol{a}$ as the diagonal elements to form a diagonal matrix
diag($\boldsymbol{A}$): = diag(vectdiag($\boldsymbol{A}$)
$\otimes$: Kronecker product
$\circ$: Khatri-Rao product
$\odot$: Hadamard product
$*$: convolution
$\in$: belongs to

2.1.3 Matrix Product Relations

The Kronecker product is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}. \tag{2.1}$$

The Khatri-Rao product is defined as

$$\boldsymbol{A} \circ \boldsymbol{B} = \begin{bmatrix} \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 & \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 & \cdots \end{bmatrix}, \tag{2.2}$$

where $\boldsymbol{a_i}$ and $\boldsymbol{b}_j$ denote the ith and jth column of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. The Hadamard product is defined as

$$\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots \\ a_{21}b_{21} & a_{22}b_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{2.3}$$

which is the element-wise product of $\boldsymbol{A}$ and $\boldsymbol{B}$.

The following properties are useful in the rest of this report:

$$
\begin{aligned}
(\boldsymbol{B}^T \otimes \boldsymbol{A})\,\mathrm{vect}(\boldsymbol{X}) &= \mathrm{vect}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) \\
(\boldsymbol{B} \otimes \boldsymbol{A})^H &= (\boldsymbol{B}^H \otimes \boldsymbol{A}^H) \\
(\boldsymbol{B} \otimes \boldsymbol{A})^{-1} &= (\boldsymbol{B}^{-1} \otimes \boldsymbol{A}^{-1}) \\
(\boldsymbol{B}^T \circ \boldsymbol{A})\boldsymbol{x} &= \mathrm{vect}(\boldsymbol{A}\,\mathrm{diag}(\boldsymbol{x})\boldsymbol{B}) \\
(\boldsymbol{B}\boldsymbol{C} \otimes \boldsymbol{A}\boldsymbol{D}) &= (\boldsymbol{B} \otimes \boldsymbol{A})(\boldsymbol{C} \otimes \boldsymbol{B}) \\
(\boldsymbol{B}\boldsymbol{C} \circ \boldsymbol{A}\boldsymbol{D}) &= (\boldsymbol{B} \otimes \boldsymbol{A})(\boldsymbol{C} \circ \boldsymbol{B}) \\
(\boldsymbol{B}^H\boldsymbol{C} \odot \boldsymbol{A}^H\boldsymbol{D}) &= (\boldsymbol{B} \circ \boldsymbol{A})^H(\boldsymbol{C} \circ \boldsymbol{B}) \\
\mathrm{vectdiag}(\boldsymbol{A}^H\boldsymbol{X}\boldsymbol{A}) &= (\boldsymbol{A}^* \circ \boldsymbol{A})^H\,\mathrm{vect}(\boldsymbol{X})
\end{aligned}
$$

.

2.2 Measurement Model of Radio Interferometric Imaging

There are two models of interferometric imaging: (i) developed in the conventional radio astronomy; (ii) constructed in the array processing framework. The latter is the basis of our work.

2.2.1 Conventional Measurement Model

Let us consider an interferometer composed of a pair of antennas as shown in Figure 2.1 and consider this interferometry working with a narrow observing frequency band centered at f. The unit vector $\boldsymbol{s}$ denotes the direction vector towards the celestial source. The intensity of the source is represented by a function of $\boldsymbol{s}$ as $I(\boldsymbol{s})$. The location difference between the two antennas is represented by $\boldsymbol{b}$. The time delay of the celestial signal arriving on two antennas is $\tau = \frac{\boldsymbol{s}^T\boldsymbol{b}}{c}$, where c denotes the speed of light. Then the output of a complex correlator is the power received per unit bandwidth from an element of the celestial source as

$$
\begin{aligned}
V_{12} &= A(\boldsymbol{s})I(\boldsymbol{s})e^{-j2\pi f\tau}d\boldsymbol{s} \\
&= A(\boldsymbol{s})I(\boldsymbol{s})e^{-j2\pi\boldsymbol{s}^T\boldsymbol{b}/\lambda}d\boldsymbol{s},
\end{aligned}
\tag{2.4}
$$

where $A(\boldsymbol{s})$ denotes the reception pattern, and λ denotes the observing wavelength. The baseline vector can be defined here as $\boldsymbol{b}_\lambda = \frac{\boldsymbol{b}}{\lambda}$.

Since the sources are so far away from the earth that we can assume that those sources are placed in an imaginary sphere called "celestial sphere" [1]. Then the total response is obtained by integrating over the solid angle subtended by the source as

$$
\mathcal{V}(\boldsymbol{b}_\lambda) = \int_{4\pi} A(\boldsymbol{s})I(\boldsymbol{s})e^{-j2\pi\boldsymbol{s}^T\boldsymbol{b}_\lambda}d\Omega. \tag{2.5}
$$

Here, the correlation output $\mathcal{V}(\boldsymbol{b}_\lambda)$ is known as "visibility" in radio astronomy. $\mathcal{V}(\boldsymbol{b}_\lambda)$ is a function of baseline and baseline depends on the observing frequency. The solid angle 4π means that the whole sphere is observed. A Fourier Transform relationship

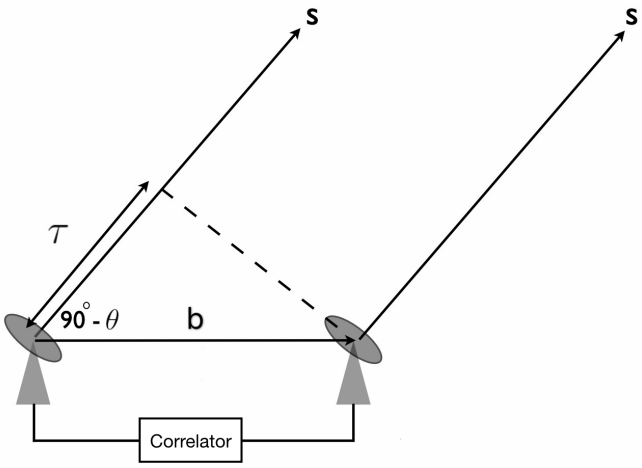


Figure 2.1: A two-element interferometer

between the intensity distribution of celestial sources and the correlation measurements has been built through this equation.

To further determine the measurement equation, a FoV is always defined and a Cartesian coordinate system is employed as shown in Figure 2.2. Two parallel planes, image plane and uv plane are defined. The location of a point in FoV is defined by l and m components, which are the direction cosines. The location of a baseline is defined by u and v components. The size of a FoV is represented by the values of l and m. An auxiliary axis n can be introduced here as $n = \sqrt{1 - l^2 - m^2}$. Now, (2.5) can be rewritten as

$$\mathcal{V}(u, v, w) = \int_l \int_m A(l, m) I(l, m) e^{-j2\pi[ul+vm+w(n-1)]} \frac{dldm}{n}, \tag{2.6}$$

where $\mathcal{V}(u, v, w)$ is the visibility, $I(l, m)$ is the source intensity distribution and $A(l, m)$ is the reception pattern.

In practice, discrete samples of the continuous function $\mathcal{V}$ are collected by the interferometer. Radio interferometric imaging is to reconstruct the image over the FoV based on these discrete samples. To obtain more samples, the earth's rotation can be used to collect time-dependent measurement data, which is called "earth rotation synthesis" [10]. Furthermore, different frequency band can be chosen to obtain different data sets. Here, we consider one frequency band and one snapshot. After discretizing (2.6), stacking the two-dimensional image grid into one dimension and omitting the constant terms, we can obtain the discrete linear measurement model as

$$\boldsymbol{r} = \boldsymbol{M}\boldsymbol{\sigma}, \tag{2.7}$$

where,

$$\boldsymbol{r} = \begin{bmatrix} \mathcal{V}(u_1, v_1, w_1) \\ \mathcal{V}(u_2, v_2, w_2) \\ \vdots \\ \mathcal{V}(u_K, v_K, w_K) \end{bmatrix} \tag{2.8}$$

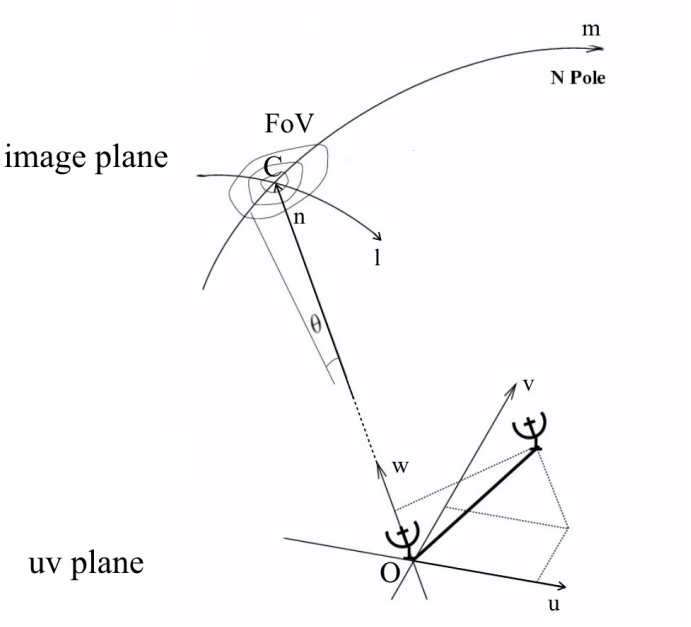


Figure 2.2: Coordinate system of baselines and sources

,

$$\boldsymbol{\sigma} = \begin{bmatrix} \frac{I(l_1,m_1)}{\sqrt{1-l_1^2-m_1^2}} \\ \frac{I(l_2,m_2)}{\sqrt{1-l_2^2-m_2^2}} \\ \vdots \\ \frac{I(l_Q,m_Q)}{\sqrt{1-l_Q^2-m_Q^2}} \end{bmatrix}, \tag{2.9}$$

and

$$m_{k,q} = e^{-j2\pi[u_k l_q + v_k m_q + w_k(\sqrt{1-l_q^2-m_q^2}-1)]}, \tag{2.10}$$

the (k,q)th entry of the kernel matrix M.

Actually, the measured visibility contains noise and quantization error. The noise is caused by the interference in propagation, inherent noise of device, etc. The additional visibility caused by this noise is represented by $\boldsymbol{r}_n$ here and can be removed by calibration process. The quantization error is caused by discretizing the integral (2.6). The error is represented by $\boldsymbol{e}$ here. The measurement equation comes down to

$$\hat{\boldsymbol{r}} = \boldsymbol{M}\boldsymbol{\sigma} + \boldsymbol{e} + \boldsymbol{r}_n. \tag{2.11}$$

If we remove the contribution of noise by calibration, we can obtain the same measurement equation as we mentioned in Chapter 1:

$$\tilde{\boldsymbol{r}} = \boldsymbol{M}\boldsymbol{\sigma} + \boldsymbol{e}. \tag{2.12}$$

We will introduce the distribution of $\boldsymbol{e}$ in the next section.

2.2.2 Array Processing Model

In the previous section, a measurement model is attained through analyzing interferometry's measurement principle and finding the relationship between image plane and uv plane and discretizing the resulting equation. In this section, we discuss how to attain the measurement model from the point of view of array signal processing. The model we introduce here is the same with the one in [7].

Celestial signals arrive on each antenna with different time delays due to the placement of each antenna. The analog signals are digitized and partitioned into narrow frequency sub-bands. Once the narrow-frequency-band condition holds, the time delays can be substituted by complex phase shifts [11]. Even though the celestial sources are stationary, the relative positions between sources and antennas change with the rotation of earth. To correct this effect, the received data are divided into short snapshots and there are N samples in each snapshot. If we assume there are P antennas, sampled output data of one time snapshot and one frequency band can be stacked into a $P \times 1$ vector $\boldsymbol{x}_k[n]$, where $n = 1, 2, \cdots, N$ represents the sample index and $k = 1, 2, \cdots, K$ represents the time snapshot and frequency band index. If we assume there are Q independent sources $s_{k,q}[n]$, with $q = 1, 2, \cdots, Q$, arriving on the antennas, they can be stacked into a $Q \times 1$ vector $\boldsymbol{s}_k[n]$. We assume independent receiver noise $n_{k,p}[n]$ where $p = 1, 2, \cdots, P$, and stack the noise into a $P \times 1$ vector $\boldsymbol{n}_k[n]$. We assume the celestial source $s_{k,q}[n]$ and receiver noise $n_{k,p}[n]$ are both zero-mean wide sense stationary (WSS) white Gaussian random processes sampled at Nyquist rate [12].

The output of the antenna array is

$$\boldsymbol{x}_k[n] = \boldsymbol{A}_k \boldsymbol{s}_k[n] + \boldsymbol{n}_k[n], \tag{2.13}$$

where $\boldsymbol{A}_k$ is a $P \times Q$ matrix called array response matrix. The (p, q)th entry of $\boldsymbol{A}_k$ represents the delay of the qth source impinging on the pth antenna:

$$a_{p,q} = \frac{1}{\sqrt{P}} e^{-j\frac{2\pi}{\lambda}\boldsymbol{y}_p^T \boldsymbol{z}_q}, \tag{2.14}$$

where the scaling $\frac{1}{\sqrt{P}}$ guarantees $\|\boldsymbol{a}_q\| = 1$ with $\boldsymbol{a}_q$ as the qth column of $\boldsymbol{A}_k$. In (2.14), λ denotes the corresponding wavelength of observing frequency f, $\boldsymbol{y}_p$ denotes the coordinates of the pth antenna in the uv plane and $\boldsymbol{z}_q$ denotes the coordinates of the qth pixel in the image plane.

We can consider only one time snapshot and one frequency band here without loss of generality [7]. Assuming the sources and noise are uncorrelated, the sources are mutually uncorrelated and the noise of different receivers are mutually uncorrelated, then the covariance model can be defined and calculated as

$$\begin{aligned} \boldsymbol{R} :&= \mathrm{E}\{\boldsymbol{x}[n]\boldsymbol{x}^H[n]\} \\ &= \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \boldsymbol{R}_n, \end{aligned} \tag{2.15}$$

where the covariance matrix of signal $\boldsymbol{R}_s = \mathrm{diag}\,(\boldsymbol{\sigma})$ with $\boldsymbol{\sigma} = [\sigma_{s,1}^2, \sigma_{s,2}^2, \cdots, \sigma_{s,Q}^2]^T$, and the covariance matrix of noise $\boldsymbol{R}_n = \mathrm{diag}\,(\boldsymbol{\sigma}_n)$ with $\boldsymbol{\sigma}_n = [\sigma_{n,1}^2, \sigma_{n,2}^2, \cdots, \sigma_{n,P}^2]^T$. In

practice, it is not possible to obtain the true correlation since the number of samples is finite. We can use the received available data to estimate the covariance matrix as

$$\hat{\boldsymbol{R}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}[n]\boldsymbol{x}^H[n]. \tag{2.16}$$

To attain an linear measurement model for the radio interferometric imaging problem, we vectorize the covariance model (2.15) as

$$\begin{aligned} \boldsymbol{r} &= (\boldsymbol{A}^* \circ \boldsymbol{A})\boldsymbol{\sigma} + \boldsymbol{r}_n \\ &= \boldsymbol{M}\boldsymbol{\sigma} + \boldsymbol{r}_n, \end{aligned} \tag{2.17}$$

where $\boldsymbol{r} = \text{vect}\,(\boldsymbol{R})$ and $\boldsymbol{r}_n = \text{vect}\,(\boldsymbol{R}_n) = (\boldsymbol{I} \circ \boldsymbol{I})\boldsymbol{\sigma}_n$. $\boldsymbol{\sigma}$ denotes the pixel intensities in the FoV. $\boldsymbol{M}$ is the system matrix of this linear model and each entry of $\boldsymbol{M}$ is related to one baseline and one source as

$$\begin{aligned} m_{ij,q} &= a_{iq}^* a_{jq} \\ &= \frac{1}{P} e^{j\frac{2\pi}{\lambda}(\boldsymbol{y}_i - \boldsymbol{y}_j)\boldsymbol{z}_q}, \end{aligned} \tag{2.18}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are the position of the ith and jth antenna, respectively.

We can also vectorize the estimate covariance matrix to obtain

$$\hat{\boldsymbol{r}} = \text{vect}\,(\hat{\boldsymbol{R}}). \tag{2.19}$$

The noise covariance matrix $\boldsymbol{R}_n$ can be known from the calibration procedure, and we can use it to do a correction of the unwanted contribution of the noise power as

$$\tilde{\boldsymbol{R}} = \hat{\boldsymbol{R}} - \boldsymbol{R}_n. \tag{2.20}$$

Vectorizing the above equation and introducing an error item $\boldsymbol{e}$ caused by the finite sampling, we can obtain

$$\begin{aligned} \tilde{\boldsymbol{r}} &= \hat{\boldsymbol{r}} - \boldsymbol{r}_n \\ &= \boldsymbol{r} + \boldsymbol{e} - \boldsymbol{r}_n \\ &= \boldsymbol{M}\boldsymbol{\sigma} + \boldsymbol{r}_n + \boldsymbol{e} - \boldsymbol{r}_n. \end{aligned} \tag{2.21}$$

As a result, the measurement equation is attained as

$$\tilde{\boldsymbol{r}} = \boldsymbol{M}\boldsymbol{\sigma} + \boldsymbol{e}, \tag{2.22}$$

where $\boldsymbol{e}$ is zero-mean and the covariance of $\boldsymbol{e}$ is

$$\begin{aligned} \boldsymbol{C}_e &= \text{E}\{(\hat{\boldsymbol{r}} - \boldsymbol{r})(\hat{\boldsymbol{r}} - \boldsymbol{r})^H\} \\ &= \frac{1}{N}(\boldsymbol{R}^T \otimes \boldsymbol{R}). \end{aligned} \tag{2.23}$$

$\boldsymbol{e}$ is usually assumed to be zero-mean complex Gaussian distributed as $\boldsymbol{e} \sim \mathcal{CN}(0, \boldsymbol{C}_e)$ [13], where $\boldsymbol{C}_e$ can be estimated using $\hat{\boldsymbol{R}}$.

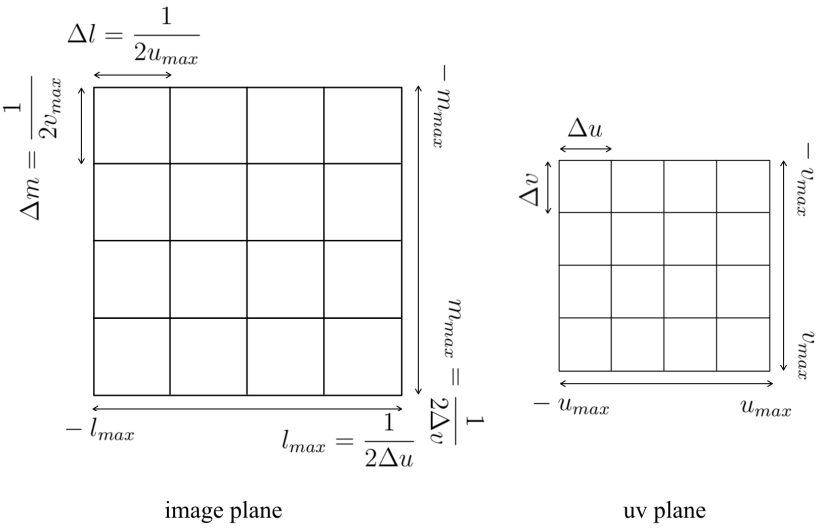


Figure 2.3: Grid on image plane and uv plane

The assumption of zero-mean Gaussian distribution of the sample error $\boldsymbol{e}$ is true only under two conditions: (i) the sample number N is large enough; (ii) the receiver noise is much stronger than the celestial signal $\boldsymbol{s}$. Since under these two conditions, the covariance of $\hat{\boldsymbol{r}}$ is independent on the mean of $\hat{\boldsymbol{r}}$, then we can assume a zero-mean Gaussian distribution of $\hat{\boldsymbol{r}}$. Furthermore, under the above conditions, we can approximate the covariance to $\boldsymbol{C}_e \approx \frac{1}{N}(\boldsymbol{R}_n^T \otimes \boldsymbol{R}_n)$ [12]. If the receiver noise is spatially white, i.e., $\boldsymbol{R}_n = \sigma_n^2 \boldsymbol{I}$, then the covariance can be further approximated to $\boldsymbol{C}_e \approx \frac{\sigma_n^4 \boldsymbol{I}}{N}$ [12]. The above assumptions usually hold in radio astronomy imaging problem.

2.3 Interpretations of the Measurement Model

In this section, we make some further interpretations of the measurement model to see how physical limitations affect imaging, and what role the system matrix $\boldsymbol{M}$ plays. These are the preparations for the next chapter.

2.3.1 Fourier Transform Relationship

For radio interferometric imaging problem, the Fourier Transform takes place between the image plane and the uv plane. If we consider image pixels and measurements are placed uniformly on rectangular grid, the Discrete Fourier Transform (DFT) relationship can be shown as Figure 2.3. In Figure 2.3, Δl and Δm represent the size of each image pixel, l_{max} and m_{max} determine the size of the FoV, Δu and Δv can be regarded as the minimum baseline length, and u_{max} and v_{max} denote the maximum baseline length.

According to the Nyquist-Shannon sampling rule, there should be

$$\begin{aligned} \Delta l &\leq \frac{1}{2u_{max}}, \\ \Delta m &\leq \frac{1}{2v_{max}}. \end{aligned} \tag{2.24}$$

The lower bound of the image pixel size is determined by the maximum baseline length. Therefore, if we aim at reconstructing a point source perfectly, an infinitely large aperture is required which is impossible. To make up the physical limitation, side information should be involved.

2.3.2 Non-Uniform and Sparse Sampling

Here we restate the explanation Naghibzadeh made in [1] to see what a role $\boldsymbol{M}$ plays besides Fourier Transform operator.

If vector $\boldsymbol{y}$ contains all the samples and $\boldsymbol{\Phi}$ denotes a sparse sampling matrix, the sub-sample vector is

$$\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{y}. \tag{2.25}$$

According to the covariance model (2.15), the correlation can be presented as

$$\begin{aligned} \boldsymbol{R}_y &= \mathrm{E}\{\boldsymbol{y}\boldsymbol{y}^H\} \\ &= \boldsymbol{F}\boldsymbol{R}_s\boldsymbol{F}^H, \end{aligned} \tag{2.26}$$

where $\boldsymbol{R}_s = \mathrm{diag}(\boldsymbol{\sigma})$ and $\boldsymbol{F}$ is the DFT matrix. Then we go back to the correlation of $\boldsymbol{x}$ to obtain

$$\begin{aligned} \boldsymbol{R}_x &= \mathrm{E}\{\boldsymbol{\Phi}\boldsymbol{y}\boldsymbol{y}^H\boldsymbol{\Phi}^h\} \\ &= \boldsymbol{\Phi}\boldsymbol{F}\boldsymbol{R}_s\boldsymbol{F}^H\boldsymbol{\Phi}^H. \end{aligned} \tag{2.27}$$

Vectorizing the above equation, we obtain the noiseless measurement model of subsampling as

$$\begin{aligned} \boldsymbol{r} &= (\boldsymbol{\Phi}^* \otimes \boldsymbol{\Phi})(\boldsymbol{F}^* \circ \boldsymbol{F})\boldsymbol{\sigma} \\ &= (\boldsymbol{\Phi}^*\boldsymbol{F}^* \circ \boldsymbol{\Phi}\boldsymbol{F})\boldsymbol{\sigma}, \end{aligned} \tag{2.28}$$

Comparing the above equation with (2.22), we find that system matrix $\boldsymbol{M}$ plays the role both of baseline sampling operator and Fourier Transform operator. In practice, the sampling is sparse and non-uniform. Therefore if we consider reconstructing the sky image by direct back-projection as $\hat{\boldsymbol{\sigma}} = \boldsymbol{M}^H\tilde{\boldsymbol{r}}$, we will obtain a bad result.

2.4 Conclusions

We started from illustrating the two-element interferometer, to conclude that the measured visibility is related to the source intensity distribution via Fourier Transform. To further define the imaging problem, we introduced the image plane and uv plane established in a Cartesian coordinate system. Then we moved to establish the measurement model in the array processing framework by defining the covariance model. Finally, we discussed the potential difficulties in image reconstruction due to physical limitations.

3 Problem Formulation

The radio interferometric imaging problem is a linear inverse problem, estimating pixel intensities $\boldsymbol{\sigma}$ from the measured and calibrated correlations $\tilde{\boldsymbol{r}}$ with the known system matrix $\boldsymbol{M}$ as stated in (2.22). Due to physical limitations such as the size of aperture, some information is lost and contaminated unavoidably. Therefore, this inverse problem is ill-posed. To modify the ill-posedness, side information is required, and we consider regularization.

3.1 Ill-posedness Analysis

A well-posed problem should have the properties: (i) a solution exist; (ii) the solution is unique; (iii) the solution's behavior changes continuously with the initial conditions. In this section, we analyze how these properties of well-posedness are violated in radio interferometric image reconstruction from the point of view of linear system and Fourier Transform. We use some analysis methods that Naghibzadeh used in [1].

3.1.1 Linear System

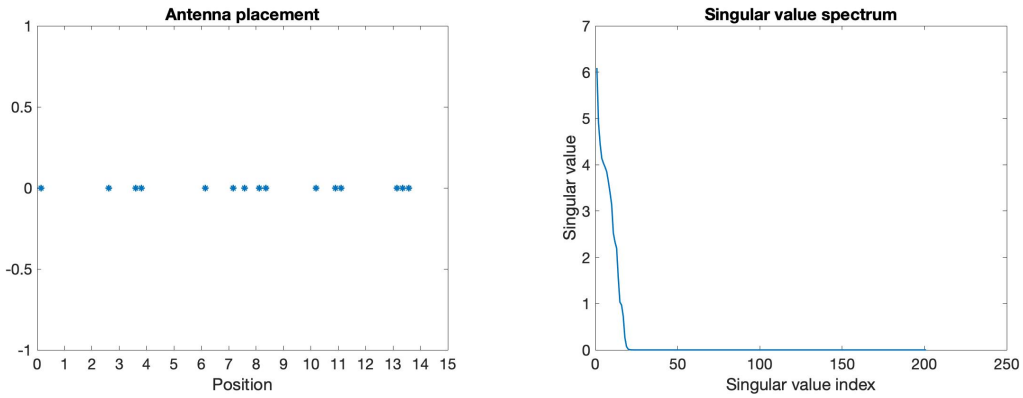
The image reconstruction problem is a linear inverse problem with $\boldsymbol{M}$ as the system matrix. If $\boldsymbol{M} \in \mathbb{C}^{P^2 \times Q}$ is a wide matrix, i.e., $P^2 < Q$, the system is underdetermined. In general there would be infinitely many solutions. This contradicts with the second property of well-posedness. To obtain a unique solution, side information on $\boldsymbol{\sigma}$ should be involved. If $\boldsymbol{M}$ is a tall matrix, i.e., $P^2 > Q$, we still cannot say that the inverse problem is well-posed, since we are not sure that a stable solution can be attained with perturbations.

A method to evaluate how good or bad is the behavior of the system matrix $\boldsymbol{M}$ against perturbations is to study the condition number of the matrix [14]. The condition number is defined as

$$cond(\boldsymbol{M}) = \frac{\varsigma_1}{\varsigma_\rho}, \tag{3.1}$$

where ς_1 and ς_ρ are the largest and smallest singular value of $\boldsymbol{M}$, respectively. A large condition number means that the matrix is ill-conditioned. The Singular Value Decomposition (SVD) of $\boldsymbol{M} \in \mathbb{C}^{P^2 \times Q}$ is

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H \\ &= \sum_{i=1}^{\rho} \boldsymbol{u}_i \varsigma_i \boldsymbol{v}_i^H, \end{aligned} \tag{3.2}$$



(a) Antenna placement (b) Singular value spectrum

Figure 3.1: SVD result of a one-dimensional case

where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{P^2}]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_Q]$ contains the left and right singular vectors, respectively, the diagonal matrix $\boldsymbol{\Sigma}$ contains the singular values on the diagonal in descending order and ρ denotes the rank of $\boldsymbol{M}$.

Here we use a simple case to show the singular value spectrum of $\boldsymbol{M}$. We do a one-dimensional experiment with 15 irregularly placed antennas ($P = 15$) and 201 pixels ($Q = 201$) ($P^2 > Q$). The corresponding antenna placement and SVD result are shown in Figure 3.1(a) and Figure 3.1(b), respectively. According to the singular value spectrum, many small singular values are close to zero, which means that the condition number computed as (3.1) is very large. Therefore, $\boldsymbol{M}$ is ill-conditioned and small perturbations will affect the solution.

Therefore, this linear inverse problem is ill-posed. We need side information to obtain a unique solution or suppress the effect of perturbations.

3.1.2 Fourier Transform

One straightforward reconstruction method is back-projection since the measurements and the pixel intensities are Fourier Transform pairs. If we assume a noiseless case, i.e., $\boldsymbol{e} = \boldsymbol{0}$, we reconstruct the sky image as

$$\begin{aligned} \hat{\boldsymbol{\sigma}} &= \boldsymbol{M}^H \tilde{\boldsymbol{r}} \\ &= \boldsymbol{M}^H \boldsymbol{M} \boldsymbol{\sigma}. \end{aligned} \tag{3.3}$$

However, as we mention in Section 2.3.2, $\boldsymbol{M}$ is not only a Fourier Transform operator but also a sampling operator. The image reconstructed in this way is called "dirty image" actually, and $\boldsymbol{M}^H \boldsymbol{M}$ can be regarded as a convolution operator. The dirty image is obtained by taking inverse Fourier Transform of the product of the continuous visibility function with the sampling function as

$$I_D(l, m) \approx \int_u \int_v \mathcal{V}(u, v) S(u, v) e^{j2\pi(ul+vm)} du dv, \tag{3.4}$$

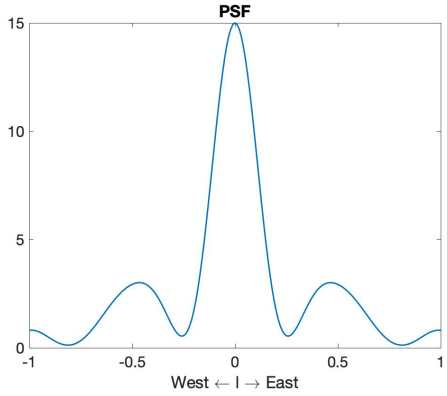


Figure 3.2: PSF

where $S(u, v)$ is the sampling function. The dirty image can be regarded as the convolution result of the true image and the inverse Fourier Transform of $S(u, v)$:

$$I_D(l, m) = I(l, m) * B(l, m), \tag{3.5}$$

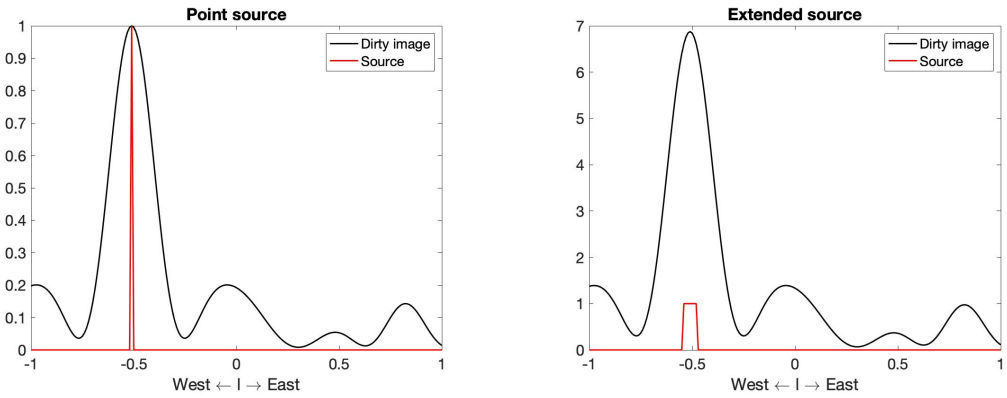
where $I(l, m)$ is the true image and $B(l, m)$ is the inverse Fourier Transform of $S(u, v)$, called dirty beam. Therefore, if we consider reconstructing the sky image through Fourier Transform, we should pay attention to the affect of sampling. The limited lengths of baselines act as a truncation on the spatial frequency signals (visibility), which leads to leakage effect in the reconstructed image. Leakage effect means that a single space component produces a set of components. Sparse sampling can be regarded as a low sampling frequency, which leads to aliasing effect. Aliasing effect means that a set of space components fold back onto a single component [15].

According to (3.4), when we set the visibility function to the constant 1, we obtain the dirty beam. Therefore, the dirty beam can be computed as

$$\boldsymbol{b} = \boldsymbol{M}^H \mathbf{1}. \tag{3.6}$$

The dirty beam can also be regarded as the impulse response of an imaging system, which is called Point Spread Function (PSF) [16]. As Naghibzadeh mentioned in [1], any two imaging systems with the same PSF are equivalent, thus we can evaluate the quality of an imaging system via evaluating the PSF. Furthermore, we evaluate the back-projection results of a point source and an extended source in noiseless case as Naghibzadeh did.

The PSF of the image system described in Section 3.1.1 is shown in Figure 3.2. We can observe conspicuous side lobes of the PSF. The reconstructed images of point source and extended source obtained by back-projecting are shown in Figure 3.3. The dirty image of the point source can be regarded as the PSF shifting to the position where the source is. The existence of side lobes makes the background dirty and the resolution is much lower than we expect. The extended source is much overestimated, even the values of side lobes are higher than the source. If there is noise, it will be amplified as well.



(a) Point source (b) Extended source

Figure 3.3: Dirty image of point source and extended source in one dimension

Utilizing Fourier Transform is not a proper way to reconstruct the sky image since the information lost in sampling is not taken into account. The above simple one-dimensional experiment proves that direct back-projection will lead to a bad result. To improve the reconstruction quality, side information should be considered.

3.2 Estimation Problem Formulations and Regularizations

As we discussed in the previous section, this radio interferometric imaging problem is an ill-posed inverse problem. To attain a stable and unique solution, we need involve side information, or additional constraints. Here we apply regularization to the objective function. In this section, we introduce different formulations of the intensity estimation problem, and consider involving regularization with some of these formulations. We restate the estimation methods discussed in [7].

3.2.1 Beamforming-Based Estimation

We first introduce the estimation methods based on beamforming. Given the available estimate of the covariance matrix $\hat{\boldsymbol{R}}$ and the noise covariance matrix obtained from the calibration procedure, the estimate is presented as

$$\hat{\sigma}_i = \boldsymbol{w}_i^H(\hat{\boldsymbol{R}} - \boldsymbol{R}_n)\boldsymbol{w}_i, \ i = 1, \cdots, Q, \tag{3.7}$$

where σ_i represents the intensity of the ith pixel and $\boldsymbol{w}_i$ is a beamformer. The estimate obtained via Matched Filter (MF) beamforming is

$$\hat{\sigma}_{MF,i} = \boldsymbol{a}_i^H(\hat{\boldsymbol{R}} - \boldsymbol{R}_n)\boldsymbol{a}_i, \tag{3.8}$$

where the beamformer is set as $\boldsymbol{w}_i = \boldsymbol{a}_i$. This estimate can be rewritten as

$$\hat{\boldsymbol{\sigma}}_{MF} = \boldsymbol{M}^H\tilde{\boldsymbol{r}}, \tag{3.9}$$

which is the estimate of dirty image (we call it dirty image in the rest of this report). The estimate obtained via Minimum Variance Distortionless Response (MVDR) beamforming is

$$\begin{aligned}\hat{\sigma}_{MVDR,i} &= \frac{\boldsymbol{a}_i\hat{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{R}}-\boldsymbol{R}_n)\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}_i}{(\boldsymbol{a}_i^H\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}_i)^2} \\ &= \frac{1}{\boldsymbol{a}_i^H\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}_i} - \frac{\boldsymbol{a}_i\hat{\boldsymbol{R}}^{-1}\boldsymbol{R}_n\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}_i}{(\boldsymbol{a}_i^H\hat{\boldsymbol{R}}^{-1}\boldsymbol{a}_i)^2}.\end{aligned} \tag{3.10}$$

where the beamformer is set to $\boldsymbol{w}_i = \frac{\boldsymbol{R}^{-1}\boldsymbol{a}_i}{\boldsymbol{a}_i^H\boldsymbol{R}^{-1}\boldsymbol{a}_i}$, and $\boldsymbol{R}$ is substituted by the estimate $\hat{\boldsymbol{R}}$. As stated in [17], without considering the correction by subtracting $\boldsymbol{R}_n$, there should be

$$\boldsymbol{0} \leq \boldsymbol{\sigma}_{true} \leq \boldsymbol{\sigma}_{MVDR} \leq \boldsymbol{\sigma}_{MF}. \tag{3.11}$$

3.2.2 Least Square Estimation

Formulating the estimation as a Least Square (LS) problem is a straightforward way. We make no additional assumptions on the sources in this simple method. The LS problem is stated as

$$\hat{\boldsymbol{\sigma}} = \arg\min_{\boldsymbol{\sigma}} \|\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma}\|_2^2. \tag{3.12}$$

The corresponding normal equation is

$$\boldsymbol{M}^H\boldsymbol{M}\hat{\boldsymbol{\sigma}} = \boldsymbol{M}^H\tilde{\boldsymbol{r}}, \tag{3.13}$$

where the LHS represents the convolution of the pixel intensities with the beam pattern, while the RHS is exactly the dirty image.

$\boldsymbol{M}^H\boldsymbol{M}$ is usually ill-conditioned in radio interferometric imaging, due to the nearly parallelity of the columns of $\boldsymbol{M}$ [7]. Furthermore, $\boldsymbol{M}^H\boldsymbol{M}$ is often non-invertible. Therefore, additional assumptions are required to obtain a unique and stable solution. A regularized LS problem can be stated as

$$\hat{\boldsymbol{\sigma}} = \arg\min_{\boldsymbol{\sigma}} \|\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma}\|_2^2 + \tau\boldsymbol{\mathcal{R}}(\boldsymbol{\sigma}), \tag{3.14}$$

where τ is the regularization parameter and $\boldsymbol{\mathcal{R}}(\cdot)$ is the regularization operator. The choice of τ is important since it adjusts the importance of the penalty term. If τ is too small, the corresponding small weight put on the penalty term would lead to overfitting of the model to the data. Conversely, if τ is too large, underfitting would be caused. There are many possibilities for $\boldsymbol{\mathcal{R}}(\cdot)$, e.g., $\|\boldsymbol{\sigma}\|_2^2$ or $\|\boldsymbol{\sigma}\|_1$. ℓ_1 regularization favors sparse solutions while ℓ_2 regularization favors evenly distributed solutions.

3.2.3 Maximum Likelihood Estimation

As described in Section 2.2.2, the noise $\boldsymbol{e}$ in (2.22) follows a zero-mean complex Gaussian distribution with the covariance matrix as $\boldsymbol{C}_e = \frac{1}{N}(\hat{\boldsymbol{R}}^T \otimes \hat{\boldsymbol{R}})$. We consider using the

Maximum Likelihood Estimation (MLE) to formulate this problem. The likelihood function is

$$p(\tilde{\boldsymbol{r}}|\boldsymbol{\sigma}) = \frac{1}{\pi^{P^2}\det(\boldsymbol{C}_e)} \exp\left\{-(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})^H \boldsymbol{C}_e^{-1}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\right\}, \tag{3.15}$$

Maximizing this likelihood function can be transformed into minimizing the cost function

$$J(\boldsymbol{\sigma}) = (\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})^H \boldsymbol{C}_e^{-1}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma}). \tag{3.16}$$

The resulting minimization problem is

$$\hat{\boldsymbol{\sigma}} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2, \tag{3.17}$$

where $\boldsymbol{C}_e^{-1} = \boldsymbol{\Gamma}^H\boldsymbol{\Gamma}$. This problem can be regarded as a Weighted Least Square (WLS) problem [12]. The difference between WLS and LS is that WLS takes the possibility that measurements may not be equally reliable due to different noise levels into account. The normal equation is

$$\boldsymbol{M}^H \boldsymbol{C}_e^{-1} \boldsymbol{M}\hat{\boldsymbol{\sigma}} = \boldsymbol{M}^H \boldsymbol{C}_e^{-1}\tilde{\boldsymbol{r}}, \tag{3.18}$$

where $\boldsymbol{M}^H\boldsymbol{C}_e^{-1}\boldsymbol{M}$ is often non-invertible. To solve this problem properly, a regularization term is involved. The regularized problem is

$$\hat{\boldsymbol{\sigma}} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \tau\boldsymbol{\mathcal{R}}(\boldsymbol{\sigma}), \tag{3.19}$$

3.2.4 Bayesian estimation

Another estimation method based on probabilistic assumptions is the Maximum A Posteriori (MAP). For MAP estimator, an assumptions on $\boldsymbol{\sigma}$ should be involved, in other words, the prior distribution $p(\boldsymbol{\sigma})$ should be offered. Thus, pixel intensities $\boldsymbol{\sigma}$ are treated as random variables in MAP while they are treated as deterministic values in MLE. Regularization is involved through including the prior distribution. The MAP estimation problem is established by maximizing the posterior as

$$\begin{aligned}
\hat{\boldsymbol{\sigma}} &= \arg\max_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}) \\
&= \arg\max_{\boldsymbol{\sigma}} \frac{p(\tilde{\boldsymbol{r}}|\boldsymbol{\sigma})p(\boldsymbol{\sigma})}{\int p(\tilde{\boldsymbol{r}}|\boldsymbol{\sigma})p(\boldsymbol{\sigma})} d\boldsymbol{\sigma} \\
&= \arg\max_{\boldsymbol{\sigma}} p(\tilde{\boldsymbol{r}}|\boldsymbol{\sigma})p(\boldsymbol{\sigma}).
\end{aligned} \tag{3.20}$$

The established maximization problem (3.20) is just a starting point of involving extra information of the unknown $\boldsymbol{\sigma}$. The key is how to design informative priors. There are different types of priors such as smoothness priors which usually lead to Gaussian prior model, sparsity-promoting priors which include ℓ_p-priors ($0 \leq p \leq 1$), etc. [18]. In the following, we will give the examples of Gaussian prior model and Laplace prior model.

If we assume that $\boldsymbol{\sigma}$ follows a Gaussian distribution with mean $\boldsymbol{\mu}_\sigma$ and covariance $\boldsymbol{C}_\sigma$, $\boldsymbol{\sigma} \sim \mathcal{N}(\boldsymbol{\mu}_\sigma, \boldsymbol{C}_\sigma)$. As a result,

$$p(\boldsymbol{\sigma}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)^T \boldsymbol{C}_\sigma^{-1}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)\right\}. \tag{3.21}$$

Furthermore,

$$\ln p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}) \propto -(\tilde{\boldsymbol{r}}-\boldsymbol{M}\boldsymbol{\sigma})^H \boldsymbol{C}_e^{-1}(\tilde{\boldsymbol{r}}-\boldsymbol{M}\boldsymbol{\sigma}) - \frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)^T \boldsymbol{C}_\sigma^{-1}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma). \tag{3.22}$$

With factorizing the covariance as $\boldsymbol{C}_\sigma^{-1} = \boldsymbol{L}^T\boldsymbol{L}$, (3.22) is expressed as

$$\ln p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}) \propto -\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}}-\boldsymbol{M}\boldsymbol{\sigma})\|_2^2 - \frac{1}{2}\|\boldsymbol{L}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)\|_2^2. \tag{3.23}$$

With a more general regularization parameter τ, the MAP estimation problem is formulated as a minimization problem as

$$\boldsymbol{\sigma} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}}-\boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \tau\|\boldsymbol{L}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)\|_2^2. \tag{3.24}$$

(3.24) is the basis of the imaging method PRIFIRA.

If we assume an independent Laplace prior distribution as

$$p(\sigma_i) = \frac{\beta_i}{2}\exp\left\{-\beta_i|\sigma_i-\mu_{\sigma,i}|\right\},\ \beta_i > 0,\ i = 1, 2, \cdots, Q, \tag{3.25}$$

where $\mu_{\sigma,i}$ is the location parameter and β_i is the scale parameter. Then the prior distribution of $\boldsymbol{\sigma}$ is

$$\begin{aligned} p(\boldsymbol{\sigma}) &= \frac{1}{\prod_{i=1}^{Q}\beta_i}\exp\left\{-\sum_{i=1}^{Q}\beta_i|\sigma_i-\mu_{\sigma,i}|\right\} \\ &= \frac{1}{\det(\boldsymbol{B})}\exp\left\{\|\boldsymbol{B}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)\|_1\right\}, \end{aligned} \tag{3.26}$$

where $\boldsymbol{B} = \mathrm{diag}(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_Q]^T$, and $\boldsymbol{\mu}_\sigma = [\mu_{\sigma,1}, \mu_{\sigma,2}, \cdots, \mu_{\sigma,Q}]^T$. The logarithm of the posteriori distribution is

$$\ln p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}) \propto -\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}}-\boldsymbol{M}\boldsymbol{\sigma})\|_2^2 - \|\boldsymbol{B}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)\|_1. \tag{3.27}$$

With a regularization parameter, the resulting minimization problem is

$$\boldsymbol{\sigma} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}}-\boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \tau\|\boldsymbol{B}(\boldsymbol{\sigma}-\boldsymbol{\mu}_\sigma)\|_1. \tag{3.28}$$

Compared to the Gaussian distribution, the Laplace distribution has fatter tails which leads to a sparser solution.

Bayesian inference provides a flexible way of incorporating extra information to supplement the noisy data by modeling the pixel intensities as random variables [18]. After determining the type of prior, we should move further to determine the variables

which the prior depends on. A complete hierarchical Bayesian model is involved. Now the posterior density can be stated as

$$p(\boldsymbol{\sigma}, \boldsymbol{\theta}|\tilde{\boldsymbol{r}}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}, \boldsymbol{\theta}), \tag{3.29}$$

where $\boldsymbol{\theta}$ denotes the variables such as $\boldsymbol{\beta}$ and $\boldsymbol{\mu}_\sigma$ which the prior depends on. $\boldsymbol{\theta}$ is called hyperparameter and $p(\boldsymbol{\theta})$ is called hyperprior [19]. There are three ways to proceed the estimation [18]. The first one is to marginalize the posterior density w.r.t $\boldsymbol{\sigma}$ as

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{r}}) = \int p(\boldsymbol{\sigma}, \boldsymbol{\theta}|\tilde{\boldsymbol{r}}) d\boldsymbol{\sigma}, \tag{3.30}$$

maximize this marginal density w.r.t $\boldsymbol{\theta}$ and maximize the posterior density $p(\boldsymbol{\sigma}, \boldsymbol{\theta}|\tilde{\boldsymbol{r}})$ with $\boldsymbol{\theta}$ fixed. The second way is to marginalize the posterior density w.r.t $\boldsymbol{\theta}$ as

$$p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}) = \int p(\boldsymbol{\sigma}, \boldsymbol{\theta}|\tilde{\boldsymbol{r}}) d\boldsymbol{\theta}, \tag{3.31}$$

and then estimate $\boldsymbol{\sigma}$ using this posterior density. Another way is to regard $(\boldsymbol{\sigma}, \boldsymbol{\theta})$ as a pair of unknowns, and maximize the posterior density $p(\boldsymbol{\sigma}, \boldsymbol{\theta}|\tilde{\boldsymbol{r}})$ w.r.t both of them. We will choose the first way and the details will be stated in the following two chapters.

3.3 Conclusions

In this chapter, we first analyzed why this image reconstruction problem would be an ill-posed inverse problem. No matter this linear measurement system is overdetermined or underdetermined, we cannot assure a stable unique solution. If we consider the inverse problem from the point of view of Fourier Transform, it is still ill-posed due to sampling deficiencies. Then we introduced some common formulations for this imaging problem and involve regularization in some of these formulations. Finally, we introduced the hierarchical Bayesian model.

4 Proposed Solution Method

In the previous chapter, we introduced some formulations to involve the prior distribution of the pixel intensities as regularization and introduced the hierarchical Bayesian model. In this chapter, we formulate the radio interferometric imaging problem into a sparsity-promoting WLS problem and relate it to Laplace prior model. We introduce the Bayesian inference framework to solve the imaging problem. We also discuss the similarities and differences between this method and PRIFIRA.

4.1 Problem Reformulation

The universe is dark with stars sparsely distributed in it. Therefore, we can involve the sparsity of $\boldsymbol{\sigma}$ as the prior knowledge. With the prior knowledge of the noise level, the problem becomes finding a sparse solution for the WLS problem (3.17). Adding a sparsity regularization term to the objective function is a widely used approach for enforcing sparse solutions. The natural choice of the sparsity regularization term is the ℓ_0 norm of $\boldsymbol{\sigma}$ which counts the nonzero elements. However, with ℓ_0 norm, the optimization problem is hard to solve. A relaxed sparsity regularization should be considered. ℓ_1 norm is an appropriate option since it is convex. With ℓ_1-regularization, the objective function is convex, thus we have many possible optimization methods to solve it. We write the regularized problem as

$$\boldsymbol{\sigma} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \beta\|\boldsymbol{\sigma}\|_1, \tag{4.1}$$

where β is the regularization parameter. β is not determined and there are two methods for determining it. One is heuristic approach [20] and another one is cross-validation [21]. Both of these methods have their limitations and do not involve the definition of the optimal sparsity. Therefore, we consider defining the optimal sparsity and inferring the regularization parameter directly from the data. Furthermore, we can associate each variable with one regularization parameter to promote the optimal sparsity as

$$\boldsymbol{\sigma} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \sum_{i=1}^{Q} \beta_i|\sigma_i|, \tag{4.2}$$

Comparing the above formulation with (3.28), we find that the above estimation problem can be attained by MAP estimation with independent zero-mean Laplace prior distributions. Therefore, we can infer the regularization parameters β_is and estimate the pixel intensities $\boldsymbol{\sigma}$ in a Bayesian inference framework.

4.2 Bayesian Inference Framework

Given the hyperparameters $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_Q]^T$, the joint posterior density is

$$p(\boldsymbol{\sigma}, \boldsymbol{\beta}|\tilde{\boldsymbol{r}}) \propto p(\boldsymbol{\beta})p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}, \boldsymbol{\beta}). \tag{4.3}$$

We adopt the first way we mentioned in Section 3.2.4 to solve the estimation problem as follows:

1. Marginalize the posterior density $p(\boldsymbol{\sigma}, \boldsymbol{\theta}|\tilde{\boldsymbol{r}})$ w.r.t $\boldsymbol{\sigma}$ as $p(\boldsymbol{\beta}|\tilde{\boldsymbol{r}}) = \int p(\boldsymbol{\sigma}, \boldsymbol{\beta}|\tilde{\boldsymbol{r}})d\boldsymbol{\sigma}$.
2. Maximize the above marginal likelihood $p(\boldsymbol{\beta}|\tilde{\boldsymbol{r}})$ w.r.t $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} p(\boldsymbol{\beta}|\tilde{\boldsymbol{r}})$.
3. Estimate $\boldsymbol{\sigma}$ by maximizing the posterior density $p(\boldsymbol{\sigma}, \boldsymbol{\theta}|\tilde{\boldsymbol{r}})$ with fixed $\boldsymbol{\beta}$ obtained from step 2.

We can easily find that the above step 3 is equivalent to solving (4.2).

Since $p(\boldsymbol{\beta}|\tilde{\boldsymbol{r}}) \propto p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})p(\boldsymbol{\beta})$, if we choose a non-informative hyperprior, maximizing $p(\boldsymbol{\beta}|\tilde{\boldsymbol{r}})$ is equivalent to maximizing $p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$. $p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$ is usually more computable. Since $\boldsymbol{\beta}$ denotes scale parameters, a proper hyperprior is Gamma distribution [18]:

$$\begin{aligned} p(\boldsymbol{\beta}) &= \prod_{i=1}^{Q} \text{Gamma}(\beta_i|a, b) \\ &= \prod_{i=1}^{Q} \Gamma(a)^{-1} b^a \beta_i^{a-1} e^{-b\beta_i}, \end{aligned} \tag{4.4}$$

where $\Gamma(\cdot)$ is the "gamma function". To make these hyperprior non-informative, i.e. flat distribution, we set a and b to small values, e.g. $x = y = 10^{-4}$ [19].

Therefore, the problem comes down to estimating hyperparametes from data by maximizing the marginal likelihood $p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$. We call this procedure as "Bayesian learning" and call this pixel intensity estimation method as "Bayesian learning method". We will introduce the details of Bayesian learning method in the next chapter.

4.3 Discussions

Usually, values of $\boldsymbol{\beta}$ which maximize $p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$ cannot be obtained in closed form. Instead, we iteratively reestimate them. Once we obtain the new estimate of $\boldsymbol{\beta}$, we can obtain the new estimate of $\boldsymbol{\sigma}$ by solving (4.2). We can summarize this iterative procedure as

$$\hat{\boldsymbol{\sigma}}^{(k)} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \|\boldsymbol{B}^{(k)}\boldsymbol{\sigma}\|_1. \tag{4.5}$$

where k denotes the iteration number and $\boldsymbol{B}^{(k)} = \text{diag}(\boldsymbol{\beta}^{(k-1)})$.

Here we briefly introduce the similar iterative procedure of IR1-PRIFIRA in [7]:

$$\hat{\boldsymbol{\sigma}}^{(k)} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \tau\|\boldsymbol{W}^{(k)}\boldsymbol{\sigma}\|_2^2, \tag{4.6}$$

where k is the iteration number and the iteratively updated weight matrix $\boldsymbol{W}^{(k)} = \operatorname{diag}\left(\left[\sqrt{\frac{1}{\hat{\sigma}_1^{(k-1)}}}, \sqrt{\frac{1}{\hat{\sigma}_2^{(k-1)}}}, \cdots, \sqrt{\frac{1}{\hat{\sigma}_Q^{(k-1)}}}\right]^T\right)$.

If we compare Bayesian learning method with IR1-PRIFIRA, we find that both of them solve the estimation problem with parameters obtained from the previous iteration. While Bayesian learning method solves a actual ℓ_1-regularized problem, IR1-PRIFIRA solves an approximate ℓ_1-regularized problem. While Bayesian learning method associates each $|\sigma_i|$ with one regularization parameter β_i, IR1-PRIFIRA associates every $|\sigma_i|$ with the same parameter τ. Therefore, we look for a better performance of Bayesian learning method due to its accuracy and generalization.

4.4 Conclusions

We reformulated the problem as a regularized WLS problem with ℓ_1-norm regularization. We involved Bayesian framework to obtain the regularization parameters from data and estimate the image. We also compared Bayesian learning method with IR1-PRIFIRA due to their similar iterative procedure.

5 Bayesian Learning Method

The proposed solution method from Chapter 4 is further developed into our Bayesian learning method in this chapter. We introduce the iterative learning procedure of the regularization parameters in detail. We restate our problem here as

$$\hat{\boldsymbol{\sigma}} = \arg\min_{\boldsymbol{\sigma}} \|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \sum_{i=1}^{Q} \beta_i |\sigma_i|, \tag{5.1}$$

where the regularization parameters β_is are attained via Bayesian learning iteratively.

5.1 Bayesian Learning

We estimate the regularization parameters $\boldsymbol{\beta}$ from data by maximizing the marginal likelihood $p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$. Since $\boldsymbol{\beta}$ cannot be obtained in closed form, we iteratively reestimate them. This iterative learning procedure is accomplished via EM algorithm.

5.1.1 EM Algorithm

As we described in Section 4.2, if the hyperprior distribution $p(\boldsymbol{\beta})$ is flat, we can estimate $\boldsymbol{\beta}$ by maximizing the marginal likelihood $p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$. We take logarithm of the marginal likelihood and present the estimation as

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta}). \tag{5.2}$$

Here we involve the hidden variables $\boldsymbol{\sigma}$ to present the marginal likelihood as

$$\begin{aligned} p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta}) &= \int p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta}) d\boldsymbol{\sigma} \\ &= \int p(\tilde{\boldsymbol{r}}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma}|\boldsymbol{\beta}) d\boldsymbol{\sigma} \\ &= \int \frac{\prod_{i=1}^{Q} \beta_i}{\pi^{P^2} \det(\boldsymbol{C}_e) 2^Q} \exp\left\{ -\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 - \sum_{i=1}^{Q} \beta_i |\sigma_i| \right\} d\boldsymbol{\sigma}. \end{aligned} \tag{5.3}$$

There is no close form for the above integral, thus we consider involving the EM algorithm to maximize the marginal likelihood iteratively [19]. The intuition behind EM algorithm is to create a lower bound of $\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$ (E-step) and then push the lower bound to increase $\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$ (M-step) [21]. The detailed derivation processes of the E-step and M-step are presented in Appendix A. The two resulting steps are as follows:

$$\text{E-step: } Z^{(k)}(\boldsymbol{\sigma}) = p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}, \boldsymbol{\beta}^{(k-1)}), \tag{5.4}$$

$$\text{M-step: } \beta_i^{(k)} = \frac{1}{\int Z^{(k)}(\boldsymbol{\sigma})|\sigma_i| d\boldsymbol{\sigma}}, \tag{5.5}$$

where k denotes the iteration number and the distribution $Z(\boldsymbol{\sigma})$ is formulated as

$$\begin{aligned} Z(\boldsymbol{\sigma}) &= \frac{p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta})}{\int p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta}) d\boldsymbol{\sigma}} \\ &= \frac{\exp\left\{-(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})^H \boldsymbol{C}_e^{-1}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma}) - \sum_{i=1}^{Q} \beta_i |\sigma_i|\right\}}{\int \exp\left\{-(\tilde{\boldsymbol{r}} - \boldsymbol{M}\tilde{\boldsymbol{\sigma}})^H \boldsymbol{C}_e^{-1}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\tilde{\boldsymbol{\sigma}}) - \sum_{i=1}^{Q} \beta_i |\tilde{\sigma}_i|\right\} d\tilde{\boldsymbol{\sigma}}}. \end{aligned} \tag{5.6}$$

The problem now is how to compute $Z(\boldsymbol{\sigma})$. Since there is no close form of $Z(\boldsymbol{\sigma})$, we are forced to adopt some approximation for $Z(\boldsymbol{\sigma})$. To obtain an approximate posterior $Z(\boldsymbol{\sigma})$, we consider expressing the logarithm of the joint density $p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta})$ in terms of a second-order Taylor approximation around its mode, i.e., its most likely value, $\boldsymbol{\sigma}^{MP}$ [22]. The mode is defined as

$$\begin{aligned} \boldsymbol{\sigma}^{MP} &= \arg\max_{\boldsymbol{\sigma}} \; Z(\boldsymbol{\sigma}) \\ &= \arg\min_{\boldsymbol{\sigma}} (\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})^H \boldsymbol{C}_e^{-1}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma}) + \sum_{i=1}^{Q} \beta_i |\sigma_i| \\ &= \arg\min_{\boldsymbol{\sigma}} \tilde{\boldsymbol{r}}^H \boldsymbol{C}_e^{-1} \tilde{\boldsymbol{r}} - 2\boldsymbol{M}^H \boldsymbol{C}_e^{-1} \tilde{\boldsymbol{r}} + \boldsymbol{M}^H \boldsymbol{C}_e^{-1} \boldsymbol{M} + \sum_{i=1}^{Q} \beta_i |\sigma_i|. \end{aligned} \tag{5.7}$$

To proceed, we need compute the mode of $Z(\boldsymbol{\sigma})$.

5.1.2 Determination of the Mode

Given that pixel intensity values are nonnegative and real, and term $\tilde{\boldsymbol{r}}^H \boldsymbol{C}_e^{-1} \tilde{\boldsymbol{r}}$ can be ignored, the minimization problem in (5.7) can be transformed to

$$\begin{aligned} \boldsymbol{\sigma}^{MP} &= \arg\min_{\boldsymbol{\sigma} \succeq 0} f(\boldsymbol{\sigma}) \\ &= \arg\min_{\boldsymbol{\sigma} \succeq 0} \frac{1}{2} \boldsymbol{\sigma}^T \boldsymbol{Q} \boldsymbol{\sigma} + \boldsymbol{b}^T \boldsymbol{\sigma} \end{aligned} \tag{5.8}$$

where $\boldsymbol{Q} = 2\boldsymbol{M}^H \boldsymbol{C}_e^{-1} \boldsymbol{M} \in \mathbb{R}^{Q \times Q}$ and $\boldsymbol{b} = -2\boldsymbol{M}^H \boldsymbol{C}_e^{-1} \tilde{\boldsymbol{r}} + \boldsymbol{\beta} \in \mathbb{R}^{Q \times 1}$. Since $\boldsymbol{Q}$ is a positive semidefinite matrix, this problem is a nonnegative quadratic convex problem. We will not solve this problem by taking the derivative w.r.t $\boldsymbol{\sigma}$ and setting it to zero, because it is improper to compute the inverse of a large-scale matrix $\boldsymbol{Q}$. To save memory, we consider just involving the matrix-vector product of $\boldsymbol{Q}$ in computational process rather than $\boldsymbol{Q}$ itself. Projected gradient descent method may be an option, but it involves choosing an appropriate step size. In the rest of this section, we will introduce an optimization method based on constructing auxiliary function, which is inspired by the multiplicative update method in [23]. Only matrix-vector product is involved and no heuristic or free parameter is involved.

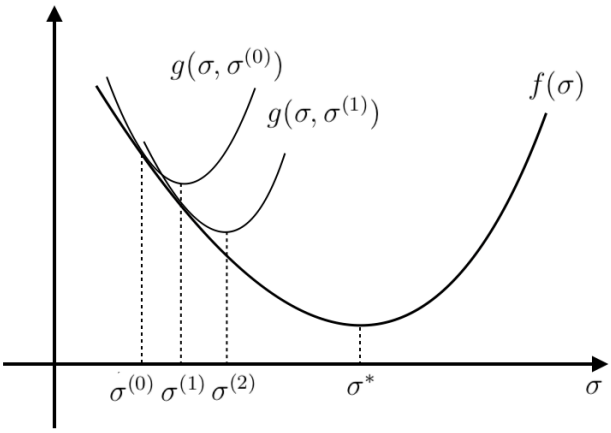


Figure 5.1: The iteration process to minimize $f(\boldsymbol{\sigma})$ via auxiliary function

The coefficient of the quadratic term of the above problem can be rewritten as

$$\boldsymbol{Q} = 2N(\boldsymbol{A}^H \hat{\boldsymbol{R}}^{-1} \boldsymbol{A})^* \odot (\boldsymbol{A}^H \hat{\boldsymbol{R}}^{-1} \boldsymbol{A}). \tag{5.9}$$

As a result, each element of $\boldsymbol{Q}$ is nonnegative. Accordingly, we construct our auxiliary function on vectors $\boldsymbol{\sigma}$ and $\tilde{\boldsymbol{\sigma}}$ of which every element is nonnegative:

$$g(\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}) = \frac{1}{2} \sum_i \frac{(\boldsymbol{Q}\tilde{\boldsymbol{\sigma}})_i}{\tilde{\sigma}_i} \sigma_i^2 + \boldsymbol{b}^T \boldsymbol{\sigma}. \tag{5.10}$$

There are two properties of this auxiliary function: (i) $f(\boldsymbol{\sigma}) \leq g(\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}})$; (ii) $f(\boldsymbol{\sigma}) = g(\boldsymbol{\sigma}, \boldsymbol{\sigma})$. With the iteration process $\tilde{\boldsymbol{\sigma}} \leftarrow \arg\min_{\boldsymbol{\sigma}} g(\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}})$, the solution will converge to the global minimum of $f(\boldsymbol{\sigma})$, which exists due to the convexity of $f(\boldsymbol{\sigma})$. The iteration procedure is shown in Figure 5.1. The convergence of this method is demonstrated in Appendix B.

It can be easily found that this auxiliary function is coordinate-wise separable, thus we can take the first derivative w.r.t each element and set it to zero to attain the minimizer of the current auxiliary function:

$$\sigma_i = -b_i \frac{\tilde{\sigma}_i}{(\boldsymbol{Q}\tilde{\boldsymbol{\sigma}})_i}, \ i = 1, 2, \cdots, Q. \tag{5.11}$$

Considering that σ_i and $\tilde{\sigma}_i$ are both nonnegative, if the result obtained in (5.11) is negative ($b_i > 0$), it should be projected onto the feasible set. In this case, to guarantee that $g(\sigma_i, \tilde{\sigma}_i)$ with nonnegative coefficients ($b_i > 0$, $\frac{(\boldsymbol{Q}\tilde{\boldsymbol{\sigma}})_i}{\tilde{\sigma}_i} > 0$) is minimal in the feasible set, the negative result should be projected to zero. The update rule is

$$\sigma_i^{(k+1)} = \sigma_i^{(k)} \frac{-b_i + |b_i|}{2(\boldsymbol{Q}\boldsymbol{\sigma}^{(k)})_i}. \tag{5.12}$$

This proposed optimization method can be efficiently and easily implemented, since there is no need to choose extra parameters such as step size and no need to store a potentially large-scale matrix $\boldsymbol{Q}$. In addition, this method can guarantee the non-negativity of the solutions.

5.1.3 Variational Approximation

After computing $\boldsymbol{\sigma}^{MP}$, $Z(\boldsymbol{\sigma})$ should be approximated around this mode. As we mentioned in Section 5.1.1, we approximate $Z(\boldsymbol{\sigma})$ by expressing the corresponding joint density $p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta})$ in a second-order Taylor expansion, which is called Laplace approximation [22]. However, we should note that when $\sigma_i^{MP} = 0$, the nonnegative constraint should be active.

The solution $\boldsymbol{\sigma}_{MP}$ partitions the elements of the vector $\boldsymbol{\sigma}$ into two subsets $\boldsymbol{\sigma}_I$ and $\boldsymbol{\sigma}_J$ with components $i \in I$ such that $(\boldsymbol{\sigma}^{MP})_i = 0$, and components $j \in J$ such that $(\boldsymbol{\sigma}^{MP})_j > 0$. Here, we use a factorial approximation method inspired by [24]. The distribution $Z(\boldsymbol{\sigma})$ is approximated as a factored form:

$$Z(\boldsymbol{\sigma}) \approx Z_I(\boldsymbol{\sigma}_I) Z_J(\boldsymbol{\sigma}_J). \tag{5.13}$$

Reconsidering the formulation of $Z(\boldsymbol{\sigma})$ in (5.6) and let

$$\begin{aligned} F(\boldsymbol{\sigma}) &= (\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})^H \boldsymbol{C}_e^{-1} (\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma}) + \sum_{i=1}^{Q} \beta_i |\sigma_i| \\ &= \tilde{\boldsymbol{r}}^H \boldsymbol{C}_e^{-1} \tilde{\boldsymbol{r}} + \frac{1}{2} \boldsymbol{\sigma}^T \boldsymbol{Q} \boldsymbol{\sigma} + \boldsymbol{b}^T \boldsymbol{\sigma}. \end{aligned} \tag{5.14}$$

$-F(\boldsymbol{\sigma})$ can be approximated by a second-order Taylor polynomial near $\boldsymbol{\sigma}^{MP}$ as

$$\begin{aligned} -F(\boldsymbol{\sigma}) = &- F\left(\boldsymbol{\sigma}^{MP}\right) - \left(\nabla F(\boldsymbol{\sigma})|_{\boldsymbol{\sigma}^{MP}}\right)\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{MP}\right) \\ &- \frac{1}{2}\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{MP}\right)^T \left(\nabla^2 F(\boldsymbol{\sigma})|_{\boldsymbol{\sigma}^{MP}}\right)\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{MP}\right). \end{aligned} \tag{5.15}$$

Since the first derivative $\left(\nabla F(\boldsymbol{\sigma})|_{\boldsymbol{\sigma}^{MP}}\right)_J = 0$, we can represent $Z_J(\boldsymbol{\sigma}_J)$ as

$$Z_J(\boldsymbol{\sigma}_J) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\sigma}_J - \boldsymbol{\sigma}_J^{MP}\right)^T \boldsymbol{Q}_{JJ}\left(\boldsymbol{\sigma}_J - \boldsymbol{\sigma}_J^{MP}\right)\right\}, \tag{5.16}$$

where $\boldsymbol{Q}_{JJ}$ is the sub-matrix of $\boldsymbol{Q}$ as $(\boldsymbol{Q}_{JJ})_{mn} = \boldsymbol{Q}_{J(m),J(n)}$ Since $\boldsymbol{\sigma}_J \succ 0$, there is no need to consider the nonnegative constraints. Now, according to the above equation, $Z_J(\boldsymbol{\sigma})$ can be approximated properly by Gaussian distribution as $Z_J(\boldsymbol{\sigma}_J) \sim \mathcal{N}(\boldsymbol{\sigma}_J|\boldsymbol{\sigma}_J^{MP}, \boldsymbol{Q}_{JJ}^{-1})$.

Since the first derivative $\left(\nabla F(\boldsymbol{\sigma})|_{\boldsymbol{\sigma}^{MP}}\right)_I \neq 0$ and $\boldsymbol{\sigma}_I^{MP} = 0$, $Z_I(\boldsymbol{\sigma})$ can be approximated as follows with nonnegative constraints:

$$Z_I(\boldsymbol{\sigma}_I) \propto \exp\left\{-\left(\boldsymbol{Q}\boldsymbol{\sigma}^{MP} + \boldsymbol{b}\right)_I^T \boldsymbol{\sigma}_I - \frac{1}{2}\boldsymbol{\sigma}_I^T \boldsymbol{Q}_{II} \boldsymbol{\sigma}_I\right\}, \ \boldsymbol{\sigma}_I \succeq 0, \tag{5.17}$$

where $\boldsymbol{Q}_{II}$ is the corresponding sub-matrix of $\boldsymbol{Q}$.

According to the naive mean-field approximation [25], $Z_I(\boldsymbol{\sigma}_I)$ can be approximated further by factorial exponential distribution [26]:

$$\hat{Z}_I(\boldsymbol{\sigma}_I) = \prod_{i\in I} \frac{1}{\mu_i} e^{-\sigma_i/\mu_i},\ \boldsymbol{\sigma}_I \succeq 0. \tag{5.18}$$

The mean field parameters $\boldsymbol{\mu}$ are obtained by minimizing the KL-divergence [21]:

$$\min_{\boldsymbol{\mu}\succeq 0} \int_{\boldsymbol{\sigma}_I\succeq 0} \ln \frac{\hat{Z}_I(\boldsymbol{\sigma}_I)}{Z_I(\boldsymbol{\sigma}_I)} \hat{Z}_I(\boldsymbol{\sigma}_I) d\boldsymbol{\sigma}_I. \tag{5.19}$$

This integral can yield a minimization problem as

$$\min_{\boldsymbol{\mu}\succeq 0} \ -\sum_{i\in I} \ln \mu_i + \hat{\boldsymbol{b}}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^T \hat{\boldsymbol{Q}} \boldsymbol{\mu}, \tag{5.20}$$

where $\hat{\boldsymbol{b}} = \left(\boldsymbol{Q}\boldsymbol{\sigma}^{MP} + \boldsymbol{b}\right)_I$ and $\hat{\boldsymbol{Q}} = \boldsymbol{Q}_{II} + \mathrm{diag}(\boldsymbol{Q}_{II})$. The details about how this minimization problem is obtained is displayed in Appendix C. To solve this minimization problem, we can construct an similar auxiliary function as:

$$g(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}) = \frac{1}{2}\sum_{i\in I} \frac{(\boldsymbol{Q}\tilde{\boldsymbol{\mu}})_i}{\tilde{\mu}_i} \mu_i^2 + \hat{\boldsymbol{b}}^T \boldsymbol{\mu} - \sum_{i\in I} \ln \mu_i. \tag{5.21}$$

To optimize $g(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}})$, we take the first derivative with respect to μ_i and set it to be zero. The iterative update rule is

$$\mu_i^{(k+1)} = \mu_i^{(k)} \frac{-\hat{b}_i + \sqrt{\hat{b}_i^2 + 4(\hat{\boldsymbol{Q}}\boldsymbol{\mu}^{(k)})_i/\mu_i^{(k)}}}{2(\boldsymbol{Q}\boldsymbol{\mu})_i^{(k)}}. \tag{5.22}$$

Now, the distribution $Z(\boldsymbol{\sigma})$ is approximated as $Z_J(\boldsymbol{\sigma}_J)\hat{Z}_I(\boldsymbol{\sigma}_I)$, then the mean of $\boldsymbol{\sigma}$ under this distribution, i.e., $\bar{\sigma}_i = \int Z(\boldsymbol{\sigma})\sigma_i d\boldsymbol{\sigma}$, is given by

$$\bar{\sigma}_i = \begin{cases} \sigma_i^{MP},\ i \in J \\ \mu_i,\ i \in I. \end{cases} \tag{5.23}$$

The update rule of $\boldsymbol{\beta}$ in (5.5) is given by:

$$\beta_i = \frac{1}{\bar{\sigma}_i}. \tag{5.24}$$

5.2 Stopping Rule and Algorithm Summary

We have introduced the details of how to obtain the regularization parameters $\boldsymbol{\beta}$ via Bayesian learning. If we compare the problem determining the mode $\boldsymbol{\sigma}^{MP}$ (5.7) with the problem estimating the image (5.1), we find they are equivalent. Therefore, we can obtain the estimate $\hat{\boldsymbol{\sigma}}$ when determining the mode $\boldsymbol{\sigma}^{MP}$ of $Z(\boldsymbol{\sigma})$.

We allow iterative procedure of Bayesian learning to continue until it converges. When it converges, we will obtain the optimally sparse solution. However, the solution will become sparser and sparser as the importance of the regularization term grows. This may cause underfitting, i.e., the model depends on the sparsity assumption too much while ignoring the data.

To avoid the underfitting, we can set a stopping criteria. We use a similar method setting the stopping rule to the method in [7], based on comparing the residual with the sample error. According to the Morozov discrepancy principle, when we have the knowledge of the noise level, the regularized solution should be as close to the data as the noise level permits [27]. This principle in our case can be stated as

$$\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\hat{\boldsymbol{\sigma}})\|_2^2 = \|\boldsymbol{\Gamma}\boldsymbol{e}\|_2^2. \tag{5.25}$$

We take expectations of both sides of the above equation to obtain:

$$\begin{aligned} \mathrm{E}\{\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\hat{\boldsymbol{\sigma}})\|_2^2\} &= \mathrm{E}\{\|\boldsymbol{\Gamma}\boldsymbol{e}\|_2^2\} \\ &= \mathrm{E}\{(\boldsymbol{\Gamma}\boldsymbol{e})^H\boldsymbol{\Gamma}\boldsymbol{e}\} \\ &= \mathrm{trace}(\mathrm{Cov}\{\boldsymbol{\Gamma}\boldsymbol{e}\}) \\ &= \mathrm{trace}(\boldsymbol{I}_{P^2\times P^2}). \end{aligned} \tag{5.26}$$

Therefore, we can allow the iteration to continue until $\frac{\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}}-\boldsymbol{M}\hat{\boldsymbol{\sigma}})\|_2^2 - P^2}{P^2} \leq \epsilon$, where ϵ is set according to the noise level to prevent the final regularized solution from being too far away from the data.

We summarize the our estimation method based on Bayesian learning as

1. Initialize the regularization parameters $\boldsymbol{\beta}$.
2. Determine the mode $\boldsymbol{\sigma}^{MP}$ by solving the minimization problem (5.7) (estimate $\hat{\boldsymbol{\sigma}}$ is also attained in this step);
3. Approximate the distribution $Z(\boldsymbol{\sigma}) \approx \hat{Z}_I(\boldsymbol{\sigma}_I)Z_J(\boldsymbol{\sigma}_J)$;
4. Calculate the mean $\bar{\boldsymbol{\sigma}}$ under the distribution $Z(\boldsymbol{\sigma})$ using (7.1);
5. Update $\boldsymbol{\beta}$ using (5.24);
6. Go back to step 2 if the stopping rule is not satisfied.

5.3 Implementation and Computational Complexity

For ease of computation, we first use Cholesky decomposition as $\hat{\boldsymbol{R}}^{-1} = \boldsymbol{L}^H\boldsymbol{L}$. Then we can express the involved matrix-vector multiplication as [7]

$$\begin{aligned} \boldsymbol{Q}\boldsymbol{x} &= 2\boldsymbol{M}^H\boldsymbol{C}_e^{-1}\boldsymbol{M}\boldsymbol{x} \\ &= 2\bar{\boldsymbol{M}}^H\bar{\boldsymbol{M}}\boldsymbol{x}, \end{aligned} \tag{5.27}$$

where $\bar{\boldsymbol{M}} = \bar{\boldsymbol{A}}^* \circ \bar{\boldsymbol{A}}$ with $\bar{\boldsymbol{A}} = N^{\frac{1}{4}}\boldsymbol{L}\boldsymbol{A}$. (5.27) can be further expressed as

$$\begin{aligned} \bar{\boldsymbol{M}}\boldsymbol{x} &= \mathrm{vect}(\bar{\boldsymbol{A}}\,\mathrm{diag}(\boldsymbol{x})\bar{\boldsymbol{A}}^H) = \boldsymbol{y}, \\ \bar{\boldsymbol{M}}^H\boldsymbol{y} &= \mathrm{vectdiag}(\bar{\boldsymbol{A}}^H\boldsymbol{Y}\bar{\boldsymbol{A}}), \end{aligned} \tag{5.28}$$

where $\boldsymbol{y} = \mathrm{vect}(\boldsymbol{Y})$.

The computational complexity of $\bar{\boldsymbol{M}}\boldsymbol{x}$ and $\bar{\boldsymbol{M}}^H\boldsymbol{y}$ are both $\mathcal{O}(P^2Q)$, where P is the number of antennas and Q is the number of sky image pixels. Therefore, the complexity of this Bayesian learning method is $\mathcal{O}(2T_1T_2P^2Q)$, where T_1 is the outer iteration number of learning $\boldsymbol{\beta}$ and T_2 is the inner iteration number of solving the nonnegative quadratic problem. In practice, T_1 and T_2 are small.

5.4 Conclusions

We introduced our proposed method based on Bayesian learning in detail. We involved the framework of EM algorithm to iteratively estimate the regularization parameters. Since there was no close form for the distribution in the E-step (5.4), we chose to approximate it around its mode via variational approximation. After updating $\boldsymbol{\beta}$ according to the M-step (5.5), the new regularization parameters were attained. We could iteratively attain the regularized solution via this Bayesian learning method.

Simulations and Experiment Results

6

In this section, we test our proposed method on one-dimensional noisy simulated data, two-dimensional noisy simulated data. We also compare the reconstruction results with other important methods of radio astronomy imaging.

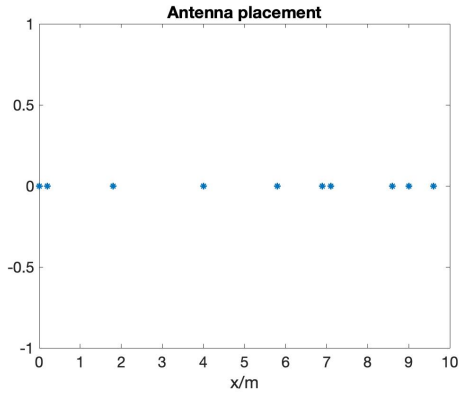
6.1 One-Dimensional Simulation

We test the Bayesian learning method and compare it with MVDR-PRIFIRA and IR1-PRIFIRA on one-dimensional case with $P = 10$ antennas and $Q = 201$ pixels. The antennas are placed non-uniformly, and the placement is shown in Figure 6.1(a).The observing frequency is set to 80 MHZ, and the frequency-dependent baseline locations are shown in Figure 6.1(b). The source shown in Figure 6.1(c) contains an extended source and point source, since one source is wider than the main beam while another one is narrower than the main beam [28]. The noise covariance is set to $\boldsymbol{R}_n = 100\boldsymbol{I}$. Figure 6.1(d) shows the covariance matrix $\hat{\boldsymbol{R}}$. From the figure of covariance matrix, we find that the noise $\boldsymbol{n}$ is dominant, which means that we can assume that $\boldsymbol{e}$ follows a complex Gaussian distribution. The SVD spectrum of $\boldsymbol{M}$ is shown in Figure 6.1(e), accordingly, $\boldsymbol{M}$ is ill-conditioned. The PSF of this imaging system is shown in Figure 6.1(f), and we can observe side lobes clearly.

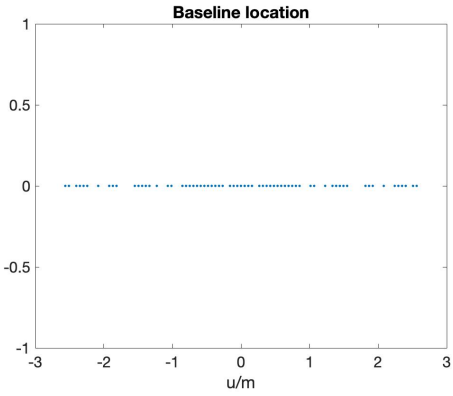
The reconstruction results are shown in Figure 6.2. The dirty image and MVDR dirty image are shown in Figure 6.2(a). Figure 6.2(b) shows the reconstruction result of our Bayesian learning method, here the iteration number is 3. Figure 6.2(c) and Figure 6.2(d) show the reconstruction results of MVDR-PRIFIRA and IR1-PRIFIRA, respectively. We set the iteration number of MVDR-PRIFIRA to 5, and we set the outer iteration number of IR1-PRIFIRA to 20. From the reconstruction results we can see that each pixel value of MF dirty image or MVDR dirty image is higher than the source, and the extended source is particularly overestimated. The point source is recovered well by the Bayesian learning while the edge of extended source becomes sharper and the reconstruction result is sparser. The MVDR-PRIFIRA can reconstruct the extended source well, but cannot reconstruct the point source and some reconstructed pixel values are negative. The IR1-PRIFIRA recovers the point source to some extent but the point source is widened and is underestimated. The IR1-PRIFIRA causes sharper edge of the extended source than the MVDR-PRIFIRA.

6.2 Two-Dimensional Simulation

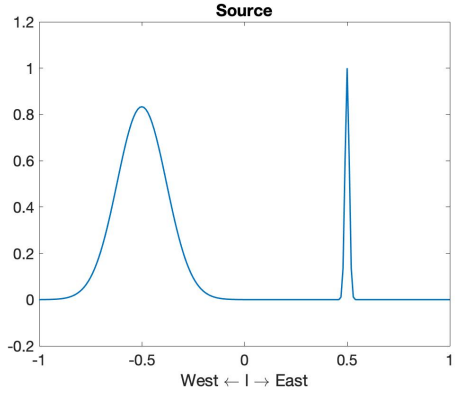
We test the Bayesian learning method on two-dimensional noisy data using the configuration of the core stations of the LOFAR as the antenna placement and an image of the W28 supernova remnant as the test image [7]. There are $P = 273$ antennas and



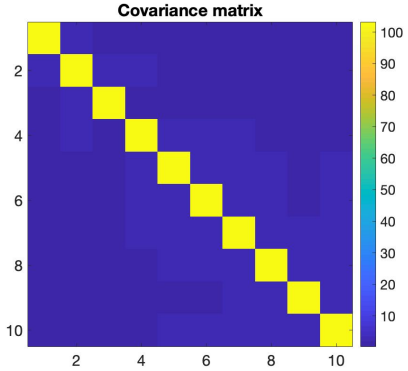
(a) Antenna placement



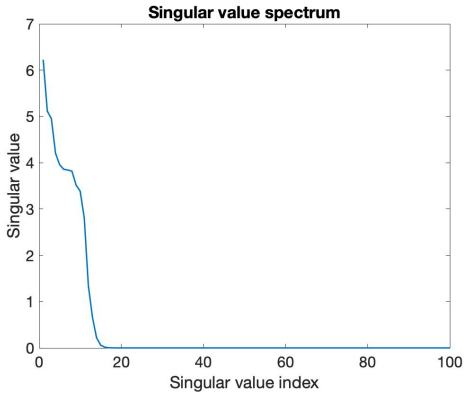
(b) Baseline location



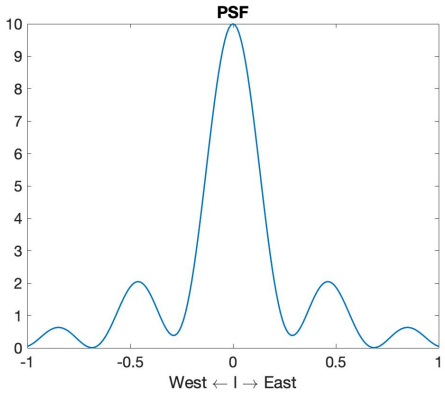
(c) Source



(d) Covariance matrix



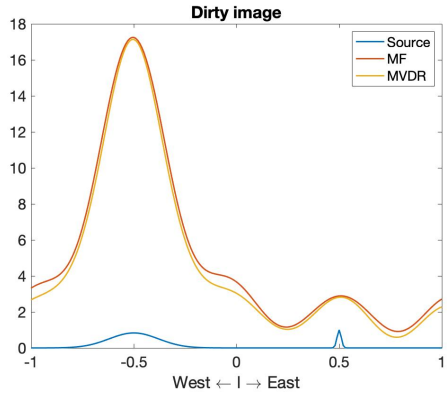
(e) Singular value spectrum



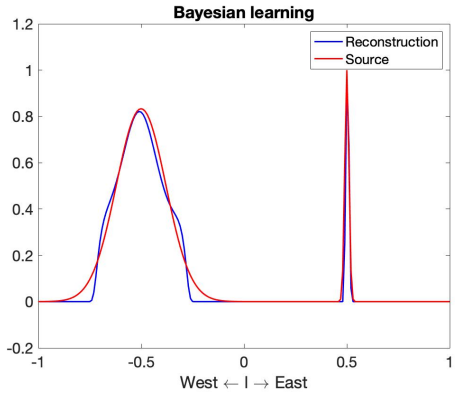
(f) PSF

Figure 6.1: One-dimensional simulation

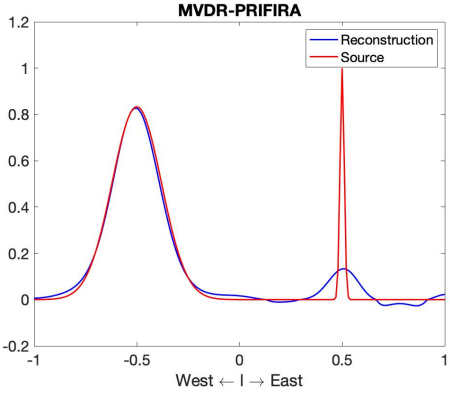
$Q = 84681$ image pixels. We only consider single frequency sub-band and single snapshot. The antenna placement is shown in Figure 6.3(a). The observing frequency is set to 58.975 MHz. The frequency-dependent uv coverage is shown in Figure 6.3(b). Each pixel value in the test image is normalized by the maximal pixel value. The test image



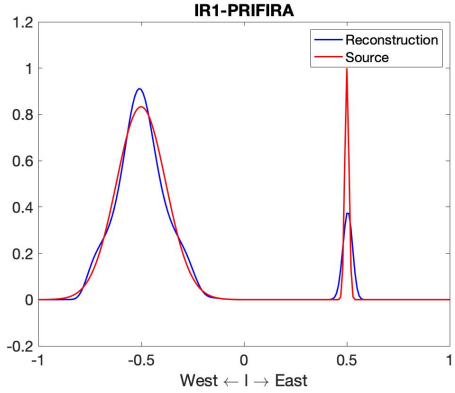
(a) Dirty image



(b) Bayesian learning



(c) MVDR-PRIFIRA



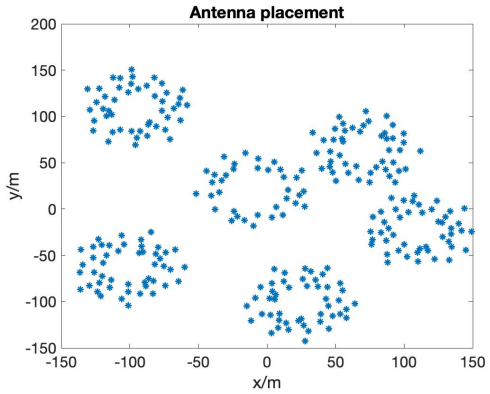
(d) IR1-PRIFIRA

Figure 6.2: One-dimensional reconstruction

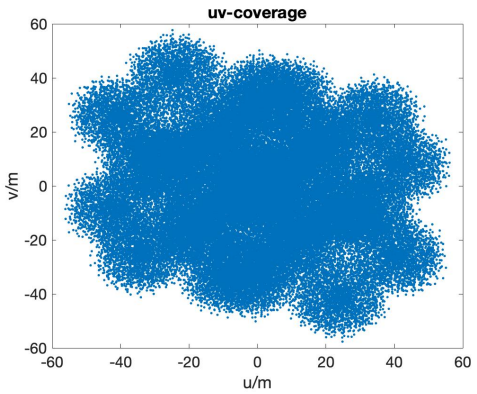
is shown in Figure 6.3(c). The receiver noise is white Gaussian noise and the covariance is set to $\boldsymbol{R}_n = 4\boldsymbol{I}$. This noise level is set according to the noise level of a real data set as introduced in [29]. The sample number is $N = 10^5$. The covariance matrix $\hat{\boldsymbol{R}}$ is shown in Figure 6.3(d). The figure of covariance matrix demonstrates that the noise $\boldsymbol{n}$ is dominant which means that we can assume a Gaussian distribution of the noise $\boldsymbol{e}$. The PSF of this imaging system is shown in Figure 6.3(e). Dirty image and MVDR dirty image are shown in Figure 6.3(f) and Figure 6.3(g), respectively. Comparing the structures of test image with the main beam of PSF, we find that there are both extended sources and point sources in the test image. Due to the limited aperture size, leakage effect will be conspicuous in the direct back-projected reconstruction, thus, we can see many side lobes in the PSF and the dirty image. Both MF beamforming and MVDR beamforming overestimate the pixel values.

6.2.1 Reconstruction Results and Performance Summary

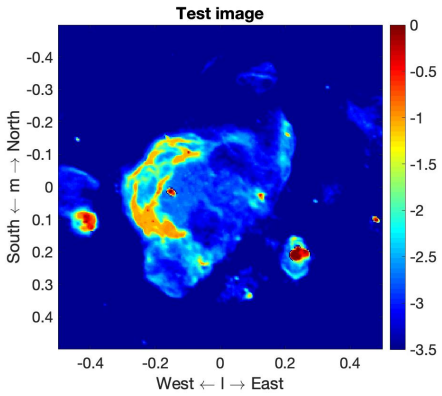
We compare the reconstruction result of our Bayesian learning methods with the results of following state-of-the-art methods: the CLEAN with a Gaussian main beam , the



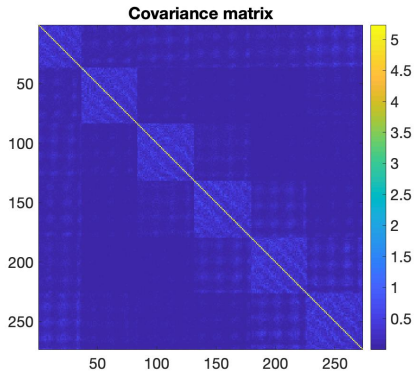
(a) Antenna placement



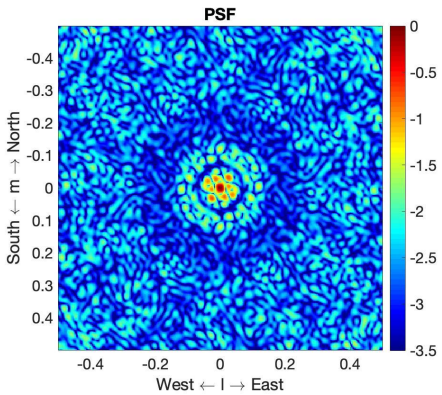
(b) uv-coverage



(c) Test image



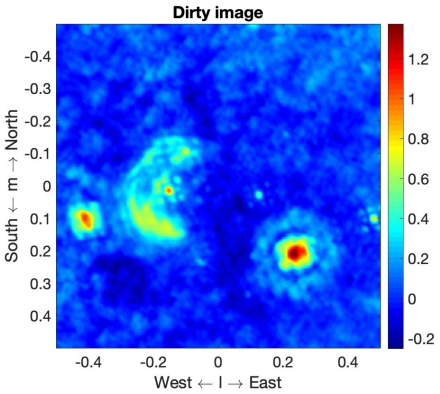
(d) Covariance matrix



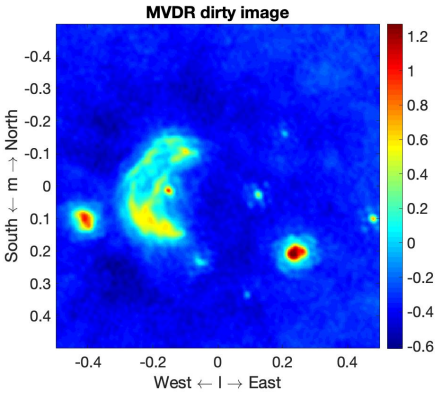
(e) PSF

ADMM with an ℓ_1-regularization, the SARA, the Richard-Lucy, the LSQR, the MVDR-PRIFIRA and the IR1-PRIFIRA. The reconstruction results are shown in Figure 6.4 in logarithmic scale.

Figure 6.4(a) and Figure 6.4(b) show that Richardson-Lucy and ADMM both cause many point-like residuals in the background, which makes it difficult to distinguish the point sources from the artifacts. Figure 6.4(c) shows that LSQR leads to many



(f) Dirty image



(g) MVDR dirty image

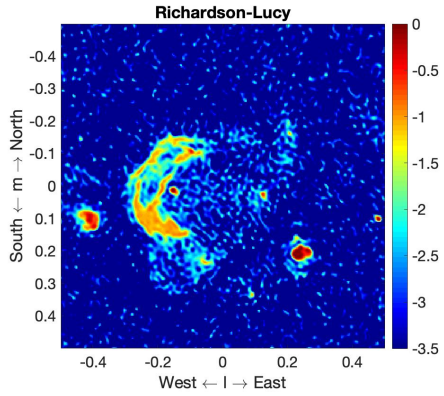
Figure 6.3: Two-dimensional simulation

	Iteration number	Reconstruction time (s)	$\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_1$	$\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2$
Bayesian learning	2	138.69	90.45	2.10
ADMM	115	128.18	93.26	1.85
CLEAN	5000	70.46	118.69	2.79
LSQR	4	5.77	367.33	3.67
MVDR-PRIFIRA	3	7.62	102.24	2.54
IR1-PRIFIRA	13	32.62	82.67	2.20
SARA	200	372.57	105.33	2.86
Richardson-Lucy	200	353.72	80.27	2.23

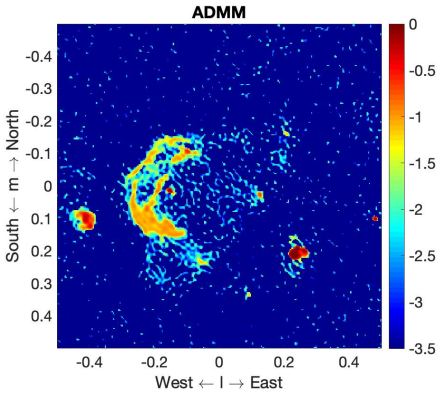
Table 6.1: Performance summary

side lobes in the reconstructed image, thus the weak sources are mixed with the side lobes. Figure 6.4(d) shows that the reconstructed image of CLEAN is composed of points, which means that the CLEAN is not suitable for the reconstruction of large-scale sources. Figure 6.4(e) shows that SARA causes many artifacts in the background which are similar to the weak extended sources. Figure 6.4(f) shows that MVDR-PRFIRA relatively well reconstructs the extended sources, but underestimates some point sources and leads to a smooth reconstruction. Figure 6.4(g) shows that IR1-PRIFIRA results in a sharper and sparser reconstruction than MVDR-PRIFIRA. Figure 6.4(h) shows that our Bayesian learning method causes a sharp and sparse reconstruction. We can still see a shape of the weak extended sources but the sources are composed of many discrete parts.

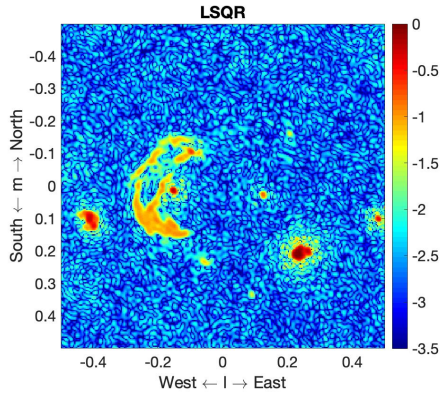
In addition to observing the reconstructed images directly, we can evaluate the reconstruction quality in terms of ℓ_1 error and ℓ_2 error. We can make a performance summary table similar to what Naghibzadeh did in [7]. Table 6.1 shows the iteration number, the reconstruction time, the ℓ_1 and ℓ_2 error. To be noted, the results of Bayesian learning and PRIFIRA we compare here are not the convergent results. The results are attained when the corresponding stopping rules are satisfied.



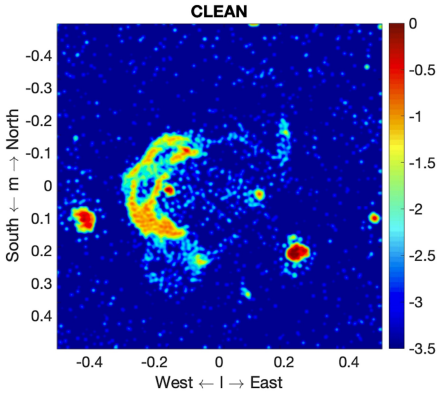
(a) Richardson-Lucy



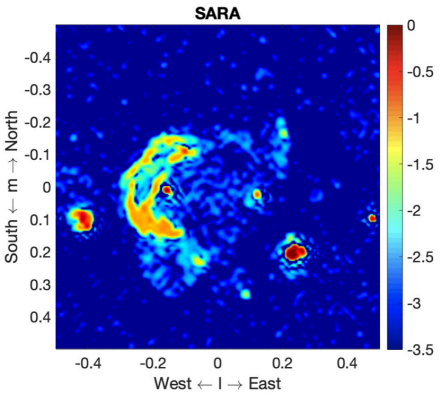
(b) ADMM



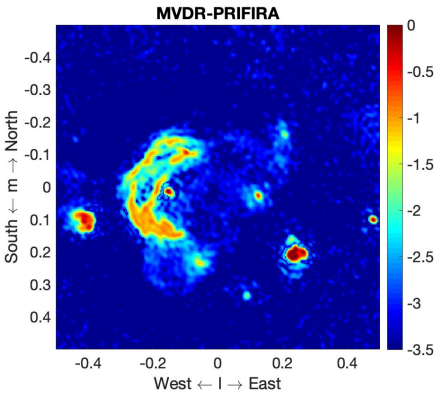
(c) LSQR



(d) CLEAN



(e) SARA

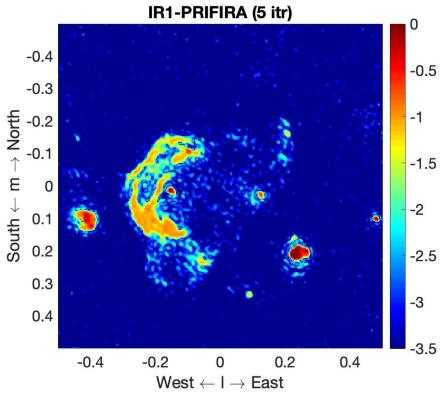


(f) MVDR-PRIFIRA

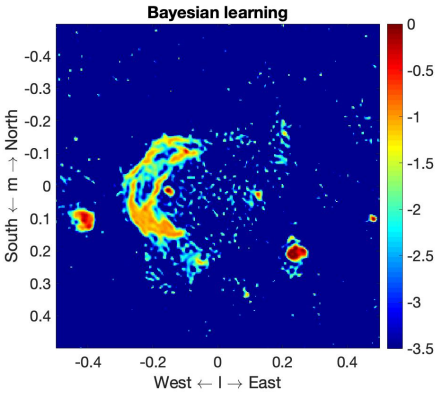
6.2.2 Further Analysis

As we mentioned in the previous section, the Bayesian learning result we compare with other methods is not a convergent result. In this section, we will show the convergent result and make further analysis about the result.

We initialize $\boldsymbol{\beta}$ to zero first. The corresponding reconstruction results of 2 iterations

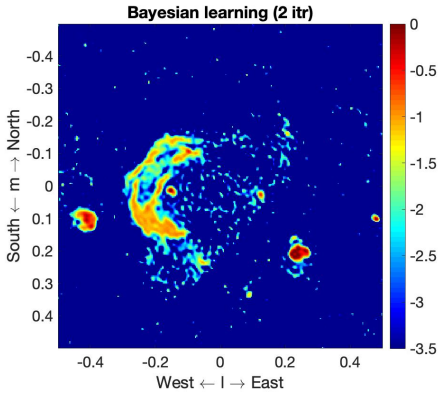


(g) IR1-PRIFIRA

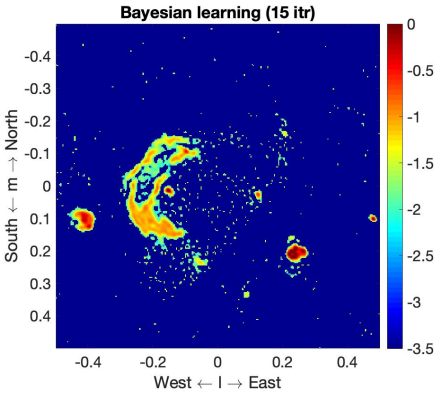


(h) Bayesian learning

Figure 6.4: Two-dimensional reconstruction



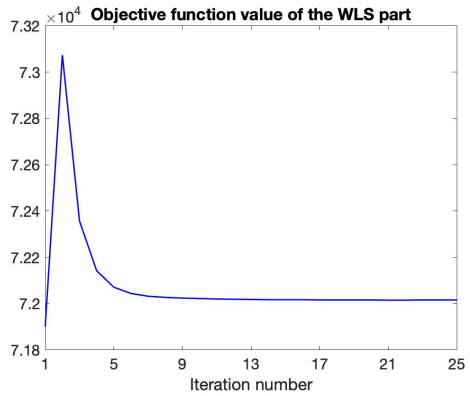
(a) 2 iterations



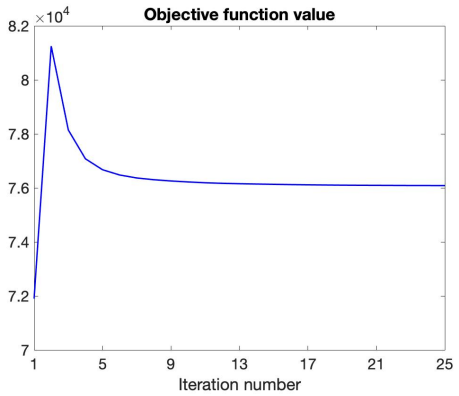
(b) 15 iterations

Figure 6.5: Bayesian learning reconstruction with zero initial regularization parameters

(not convergent) and 15 iterations (convergent) are shown in Figure 6.5. We also show the value of the objective function $\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2 + \sum_{i=1}^{Q} \beta_i |\sigma_i|$ and the value of the WLS part $\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2$ of each iteration in Figure 6.6. From Figure 6.5, we can see that the result of 15 iterations is sparser than the result of 2 iterations. Figure 6.6 illustrates that the estimation result converges after around 9 iterations, and the objective function reaches the highest value at the second iteration. To be noted, the initial values of the regularization parameters $\boldsymbol{\beta}$ are zero. Therefore, the solution of the first iteration is the solution of the WLS problem. If we compare the objective value of the WLS part $\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\boldsymbol{\sigma})\|_2^2$ with the expectation of the square norm of the whitened noise $\mathrm{E}\{\|\boldsymbol{\Gamma}\boldsymbol{e}\|_2^2\}$ as we mentioned in Section 5.2, we find these two values are the closest at the second iteration. Therefore, the solution moves from the WLS solution to a regularized solution that is as close to the data as the noise level permits, and keeps moving away from fitting to the data. As the iteration number increases, more and more importance is put on the regularization term and the solution



(a) Objective function value of the WLS part



(b) Objective function value

Figure 6.6: Objective function value with zero initial regularization parameters

become sparser and sparser until convergence. Underfitting may occur in this process, thus we involve stopping rule to prevent this. The stopping rule can be regarded as another regularization term besides the ℓ_1-regularization term. ℓ_1-regularization term helps us select some solutions from all the possible solutions, and the stopping rule makes the further selection.

6.3 Conclusions

We tested the Bayesian learning method by one-dimensional and two-dimensional simulations and compared the reconstruction results with other imaging methods. We further showed how the stopping rule controlled the solution.

7 Conclusions and Future Work

7.1 Conclusions

Radio interferometric imaging is to reconstruct the sky image over the field of view based on the measured correlation data. By employing the array processing framework, a linear data model is attained. Accordingly, radio interferometric imaging is regarded as a linear inverse problem. However, due to the limited aperture size and sparse sampling, this inverse problem is ill-posed. To obtain a unique and stable solution, side information should be offered, here we involved regularization. Based on the sparsity assumption and prior knowledge of the noise, we reformulated the problem into a weighted least square problem with ℓ_1-regularization. We employed Bayesian framework to solve this problem. We proposed our Bayesian learning method to learn the regularization parameters from data and to update the image estimate correspondingly until the convergence of the regularization parameters. Our Bayesian learning method is based on an EM algorithm framework. We employed an optimization method based on constructing auxiliary functions and an variational approximation method to compute the distribution in the E-step of the EM algorithm. We involved a stopping rule to prevent the solution from being too sparse to fit to the data. The reconstruction results attained by our Bayesian learning method are comparable with other existing radio astronomy imaging methods.

Compared to conventional ℓ_1-regularization with only one regularization parameter, our regularization is generalized to the case that each variable is associated with one regularization parameter. Compared to common approaches such as heuristics and cross-validation to determine regularization parameters, we used the Bayesian learning to infer the parameters from data by maximizing the posterior distribution of the regularization parameters. Compared to the efficient imaging method PRIFIRA in which the exact objective function obtaining the final solution is hard to be determined, our Bayesian learning method can clearly reveal the objective function which the final solution comes from. Moreover, in PRIFIRA, the nonnegativity of image pixel values is not taken into account, but in our Bayesian learning method, we involve nonnegative constraints.

Our Bayesian learning method provides a good option for sparse radio astronomy image models. It converge very quickly to the sparse optimal solution. Moreover, it is allowed to stop at a earlier point to control the sparsity of the solution. It can recover point sources very well and profile some extended sources. To be noted, it can cause discontinuity of extended sources and cannot recover weak extended sources. These may be overcome by involving mixture prior models on the pixel values.

7.2 Future Work

This Bayesian learning method can also be tested on real data. However, due to the absence of ground truth, the reliability of the reconstruction should be evaluated by a proper method.

We set the stopping rule based on the noise level. However, this may be unreliable due to the uncertainty of scaling. A more reliable stopping rule can be proposed.

In our formulation, each variable is associated with one regularization parameter, which implies that the image pixel values follow independent zero-mean Laplace distributions. In the future work, the mean of the Laplace distribution can be discussed and generalized.

Moreover, instead of the independent Laplace distributions, other prior distributions such as Chi-square distribution can be assumed. We can even propose a mixture of different distributions to recover both point sources and extended sources properly.

Due to the existence of large-scale matrix, the computation of our Bayesian learning method is not fast sufficiently. Strategies to speed up the computation is required for future work. One possibility is to combine the Krylove-based method with our Bayesian learning method but the nonnegativity constraints should be involved.

$\tilde{\boldsymbol{r}}$

$\hat{\boldsymbol{\sigma}}$

$+\tau\mathcal{R}(\boldsymbol{\sigma})$

$$\|\boldsymbol{\Gamma}(\tilde{\boldsymbol{r}} - \boldsymbol{M}\hat{\boldsymbol{\sigma}})\|_2^2 + \sum_{i=1}^{Q} \beta_i |\hat{\sigma}_i| \tag{7.1}$$

Bibliography

[1] S. Naghibzadeh. "Image formation for future radio telescopes". PhD thesis. Delft University of Technology, 2018.

[2] J. Högbom. "ReferencesAperture Synthesis with a Non-Regular Distribution of Interferometer Baselines". In: *Astronomy and Astrophysics Supplement* 15 (1974), p. 3417.

[3] B. G. Clark. "An efficient implementation of the algorithm 'CLEAN'". In: *Astronomy and Astrophysics* 89.3 (Sept. 1980), pp. 377–378.

[4] D. G. Steer, P. E. Dewdney, and M. R. Ito. "Enhancements to the deconvolution algorithm 'CLEAN'". In: *Astronomy and Astrophysics* 137.2 (Aug. 1984), pp. 159–1645.

[5] T. J. Cornwell amnd K. F. Evans. "A simple maximum entropy deconvolution algorithm". In: *Astronomy and Astrophysics* 141.1 (Feb. 1985), pp. 77–83.

[6] R. E. Carrillo, J. D. McEwen, and Y. Wiaux. "Sparsity Averaging Reweighted Analysis (SARA): a novel algorithm for radio-interferometric imaging". In: *Monthly Notices of the Royal Astronomical Society* 426 (2 Oct. 2012), pp. 1223–1234.

[7] S. Naghibzadeh and A.-J. van der Veen. In: *Monthly Notices of the Royal Astronomical Society* 479 (4 Oct. 2018), pp. 5638–5656.

[8] A Leshem and A.-J. van der Veen. "Radio-astronomical imaging in the presence of strong radio interference". In: *IEEE Transactions on Information Theory* 46.5 (2000), pp. 1730–1747.

[9] A. M. Sardarabadi. "Covariance matching techniques for radio astronomy calibration and imaging". PhD thesis. Delft University of Technology, 2016.

[10] A. R. Thompson, J. M. Moran, and G. W. Swenson Jr. *Interferometry and Synthesis in Radio Astronomy (Third Edition)*. Wiley-VCH, 2017.

[11] A.-J. van der Veen and G. Leus. *Signal Processing for Communications*. Delft, 2005.

[12] A.-J. van der Veen and S. J. Wijnholds. "Signal Processing Tools for Radio Astronomy". In: *Handbook of Signal Processing Systems*. Ed. by Shuvra S. Bhattacharyya et al. New York, NY: Springer New York, 2013, pp. 421–463.

[13] B. Ottersten, P. Stocia, and R. Roy. In: *Digital Signal Processing* 8 (3 July 1998), pp. 185–210.

[14] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. Philadelphia, PA: SIAM, 1997.

[15] D. Sundararajan. "Aliasing and Leakage". In: *Fourier Analysis—A Signal Processing Approach*. Singapore: Springer Singapore, 2018, pp. 159–178.

[16] T. L. Wilson, K. Rohlfs, and S. Hüttemeister. *Tools of Radio Astronomy*. Springer, 2009.

[17] A. M. Sardarabadi, A. Leshem, and A.J. van der Veen. "Radio astronomical image formation using constrained least squares and Krylov subspaces". In: *Astronomy and Astrophysics* (Oct. 2015).

[18] D. Calvetti and E. Somersalo. “Inverse problems: From regularization to Bayesian inference”. In: *WIREs Computational Statistics* 10 (3 June 2018).
[19] M. E. Tipping. “Sparse Bayesian Learning and the Relevance Vector Machine”. In: *Machine Learning Research* 1 (Jan. 2001). Ed. by A. Smola, pp. 211–244.
[20] S. S. Chen, D. L. Donoho, and M. A. Saunders. “Atomic decomposition by basis pursuit”. In: *Scientific Computing* 20 (1 Aug. 1998), pp. 33–61.
[21] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
[22] D. MacKay. “Laplace’s Method”. In: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003, pp. 341–342.
[23] F. Sha et al. “Multiplicative Updates for Nonnegative Quadratic Programming”. In: *Neural Computation* 19 (Jan. 2007), pp. 2004–2031.
[24] D. D. Lee and Y. Lin. “Bayesian L_1-Norm Sparse Learning”. In: *Proceedings of the 2006 IEEE International Conference Acoustics, Speech and Signal Processing*. Vol. 5. May 2006, pp. V605–V608.
[25] M. J. Wainwright and M. I. Jordan. “Graphical models, exponential families, and variational inference”. In: *Foundation and Trends in Machine Learning* 1 (2008), pp. 1–305.
[26] M. P. Wand et al. “Mean Field Variational Bayes for Elaborate Distributions”. In: *International Society for Bayesian Analysis*. 4 4. 2011, pp. 1–48.
[27] V. A. Morozov. “The error principle in the solution of operational equations by the regularization method”. In: *U.S.S.R. Comput. Math. Math. Phys.* 8 (2 1968), pp. 63–87.
[28] J. N. Chengalur. “Two Element Interferometers”. In: *Radio Astronomy*, p. 3.
[29] S. J. Wijnholds. “Fish-Eye Observing with Phased Array Radio Telescopes”. PhD thesis. Delft University of Technology, 2010.

EM Algorithm

A

To create the lower bound, $\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$, a distribution $Z(\boldsymbol{\sigma})$ over the hidden variables $\boldsymbol{\sigma}$ is defined. With this distribution, the logarithm of the marginal likelihood can be rewritten as

$$\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta}) = \ln \int Z(\boldsymbol{\sigma}) \frac{p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta})}{Z(\boldsymbol{\sigma})} d\boldsymbol{\sigma}. \tag{A.1}$$

For any concave function such as logarithm function, the Jensen's inequality holds as

$$f(E\{x\}) \geq E\{f(x)\}. \tag{A.2}$$

The equality hold when f is affine or x is constant. According to (A.1) and (A.2), the lower bound is presented as

$$\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta}) \geq \int Z(\boldsymbol{\sigma}) \ln \frac{p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta})}{Z(\boldsymbol{\sigma})} d\boldsymbol{\sigma}. \tag{A.3}$$

The EM algorithm works iteratively by two steps: (i) E-step finds a tight lower bound of $\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$; (ii) M-step forces $\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta})$ to increase by maximizing the lower bound.

A.1 E-step

In the kth iteration, to ensure that the lower bound is tight, the lower bound should be equal to $\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta}^{(k-1)})$. The equality between $\ln p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta}^{(k-1)})$ and the lower bound holds if $\frac{p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta}^{(k-1)})}{Z^{(k)}(\boldsymbol{\sigma})}$ is a constant. Therefore, $Z^{(k)}(\boldsymbol{\sigma})$ is chosen as

$$\begin{aligned} Z^{(k)}(\boldsymbol{\sigma}) &= \frac{p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta}^{(k-1)})}{\int p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta}^{(k-1)}) d\boldsymbol{\sigma}} \\ &= \frac{p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta}^{(k-1)})}{p(\tilde{\boldsymbol{r}}|\boldsymbol{\beta}^{(k-1)})} \\ &= p(\boldsymbol{\sigma}|\tilde{\boldsymbol{r}}, \boldsymbol{\beta}^{(k-1)}). \end{aligned} \tag{A.4}$$

A.2 M-step

After we find $Z^{(k)}(\boldsymbol{\sigma})$, we should maximize the lower bound over $\boldsymbol{\beta}$ as

$$\begin{aligned}
\boldsymbol{\beta}^{(k)} &= \arg\max_{\boldsymbol{\beta}} \int Z^{(k)}(\boldsymbol{\sigma}) \ln \frac{p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta})}{Z^{(k)}(\boldsymbol{\sigma})} d\boldsymbol{\sigma} \\
&= \arg\max_{\boldsymbol{\beta}} \int Z^{(k)}(\boldsymbol{\sigma}) \ln p(\boldsymbol{\sigma}, \tilde{\boldsymbol{r}}|\boldsymbol{\beta}) d\boldsymbol{\sigma} \\
&= \arg\max_{\boldsymbol{\beta}} \int Z^{(k)}(\boldsymbol{\sigma}) \ln [p(\tilde{\boldsymbol{r}}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma}|\boldsymbol{\beta})] d\boldsymbol{\sigma} \\
&= \arg\max_{\boldsymbol{\beta}} \int Z^{(k)}(\boldsymbol{\sigma}) \left[\sum_{i=1}^{Q} (\ln \beta_i - \beta_i |\sigma_i|) \right] d\boldsymbol{\sigma}
\end{aligned} \tag{A.5}$$

To maximize the final resulting cost function in (A.5), we take the derivative of the cost function over β_i as

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_i} = \int Z^{(k)}(\boldsymbol{\sigma}) \left(\frac{1}{\beta_i} - |\sigma_i| \right) d\boldsymbol{\sigma}. \tag{A.6}$$

Setting (A.6) equal to zero, we obtain

$$\beta_i = \frac{1}{\int Z^{(k)}(\boldsymbol{\sigma}) |\sigma_i| d\boldsymbol{\sigma}}. \tag{A.7}$$

Convergence Analysis of the Optimization Method Based on Auxiliary Function

B

We demonstrate here that this auxiliary-function-based optimization method converges monotonically to the global minimum of $f(\boldsymbol{\sigma})$ referring to [23].

B.1 Monotonic Convergence

The auxiliary function is constructed around the current estimate of the minimizer. Subsequently, the next estimate is attained by minimizing this auxiliary function, which can be regarded as setting an upper bound on the cost function $f(\boldsymbol{\sigma})$. Letting this iteration continues, we will find a stationary point or a local minimum of $f(\boldsymbol{\sigma})$. If those two properties stated in Section 5.1.2 hold, this process would not increase the value of $f(\boldsymbol{\sigma})$ due to:

$$\begin{aligned} f(\sigma^{(0)}) = g(\sigma^{(0)}, \sigma^{(0)}) \geq g(\sigma^{(1)}, \sigma^{(0)}) \geq f(\sigma^{(1)}) \\ f(\sigma^{(1)}) = g(\sigma^{(1)}, \sigma^{(1)}) \geq g(\sigma^{(2)}, \sigma^{(1)}) \geq f(\sigma^{(2)}) \\ \vdots \end{aligned} \tag{B.1}$$

Therefore, what we should prove here is that the property $f(\boldsymbol{\sigma}) \leq g(\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}})$ holds, which is equivalent to

$$\boldsymbol{\sigma}^T \boldsymbol{Q} \boldsymbol{\sigma} \leq \sum_i \frac{(\boldsymbol{Q}\tilde{\boldsymbol{\sigma}})_i}{\tilde{\sigma}_i} \sigma_i^2. \tag{B.2}$$

We involve a diagonal matrix $\boldsymbol{K}$ of which each entry is

$$K_{ij} = \delta_{ij} \frac{(\boldsymbol{Q}\tilde{\boldsymbol{\sigma}})_i}{\tilde{\sigma}_i}, \tag{B.3}$$

where δ_{ij} is the Kronecker delta function. If $(\boldsymbol{K} - \boldsymbol{Q})$ is positive semidefinite, the inequality (B.2) holds. For any vector $\boldsymbol{x} \in \mathbb{R}^{Q \times 1}$, we can apply a coordinate-wise

scaling as $x_i\tilde{\sigma}_i$, and then

$$\begin{aligned}(\boldsymbol{x}\odot\tilde{\boldsymbol{\sigma}})^T(\boldsymbol{K}-\boldsymbol{Q})(\boldsymbol{x}\odot\tilde{\boldsymbol{\sigma}}) &= \sum_{i,j} x_i\tilde{\sigma}_i(K_{ij}-Q_{ij})x_j\tilde{\sigma}_j \\ &= \sum_{i,j}[\delta_{ij}(\boldsymbol{Q}\tilde{\boldsymbol{\sigma}})_i x_i x_j\tilde{\sigma}_j - Q_{ij}x_i x_j\tilde{\sigma}_i\tilde{\sigma}_j] \\ &= \sum_{i,j}(Q_{ij}\tilde{\sigma}_i\tilde{\sigma}_j x_i^2 - Q_{ij}x_i x_j\tilde{\sigma}_i\tilde{\sigma}_j) \\ &= \frac{1}{2}\sum_{i,j} Q_{ij}\tilde{\sigma}_i\tilde{\sigma}_j(x_i-x_j)^2 \\ &\geq 0.\end{aligned} \tag{B.4}$$

Therefore, $(\boldsymbol{K}-\boldsymbol{Q})$ is positive semidefinite, which means that inequality (B.2) holds as we expect.

B.2 Global Convergence

A mapping $\mathcal{M}$ is defined in this iterative optimization procedure, $\mathcal{M}:\boldsymbol{\sigma}^{(k)}\rightarrow\boldsymbol{\sigma}^{(k+1)}$. A sequence of estimates are generated as $\{\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)},\cdots\}$. The corresponding sequence of cost function value $\{f(\boldsymbol{\sigma}^{(1)}),f(\boldsymbol{\sigma}^{(2)}),\cdots\}$ is monotonically decreasing and is bounded below the global minimum of $f(\boldsymbol{\sigma})$. When the iteration number k increase to infinity, the sequence of $\{f(\boldsymbol{\sigma}^{(1)}),f(\boldsymbol{\sigma}^{(2)}),\cdots\}$ converges. What we should prove here is that it converge to the global minimum.

We briefly introduce the proof procedure here and the details can be seen in [23]. It is demonstrated firstly that any limit point of the sequence $\{\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)},\cdots\}$ must be a fixed point of the mapping $\mathcal{M}$. It is shown secondly that there is no convergent subsequences of $\mathcal{M}$ with spurious fixed points as limit points. Finally, the result of subsequence convergence is extended and it is proved that the sequence $\{\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)},\cdots\}$ converges to the global minimizer.

Minimization of the KL-Divergence

C

Here we introduce how to obtain the minimization problem (5.20) by minimizing the KL-divergence as (5.19).

In (5.19),

$$\begin{aligned}
&\int_{\boldsymbol{\sigma}_I \succeq 0} \hat{Z}_I(\boldsymbol{\sigma}_I) \ln \hat{Z}_I(\boldsymbol{\sigma}_I) d\boldsymbol{\sigma}_I \\
&= \sum_{i \in I} \int_{\sigma_i \geq 0} \hat{Z}_i(\sigma_i) \ln \hat{Z}_i(\sigma_i) d\sigma_i \\
&= \sum_{i \in I} \int_0^{\infty} (-\ln \mu_i - \frac{\sigma_i}{\mu_i}) \frac{e^{-\frac{\sigma_i}{\mu_i}}}{\mu_i} d\sigma_i \\
&= -\sum_{i \in I} (\ln \mu_i + 1)
\end{aligned} \tag{C.1}$$

and

$$\begin{aligned}
&\int_{\boldsymbol{\sigma}_I \succeq 0} \hat{Z}_I(\boldsymbol{\sigma}_I) \ln Z_I(\boldsymbol{\sigma}_I) d\boldsymbol{\sigma}_I \\
&= \int_{\boldsymbol{\sigma}_I \succeq 0} \hat{Z}_I(\boldsymbol{\sigma}_I) \ln \frac{exp\Big\{ -\left(\boldsymbol{Q}\boldsymbol{\sigma}^{MP} + \boldsymbol{b}\right)_I^T \boldsymbol{\sigma}_I - \frac{1}{2}\boldsymbol{\sigma}_I^T \boldsymbol{Q}_{II} \boldsymbol{\sigma}_I \Big\}}{\mathcal{Z}_I} d\boldsymbol{\sigma}_I \\
&= -\ln \mathcal{Z}_I - \int_{\boldsymbol{\sigma}_I \succeq 0} \hat{Z}_I(\boldsymbol{\sigma}_I) \left[\left(\boldsymbol{Q}\boldsymbol{\sigma}^{MP} + \boldsymbol{b}\right)_I^T \boldsymbol{\sigma}_I + \frac{1}{2}\boldsymbol{\sigma}_I^T \boldsymbol{Q}_{II} \boldsymbol{\sigma}_I \right] d\boldsymbol{\sigma}_I \\
&= -\ln \mathcal{Z}_I - \sum_{i \in I} \left(\boldsymbol{Q}\boldsymbol{\sigma}^{MP} + \boldsymbol{b}\right)_i \int_0^{\infty} \hat{Z}_i(\sigma_i) \sigma_i d\sigma_i \\
&\quad - \frac{1}{2} \sum_{i \in I, j \in I} Q_{ij} \int_0^{\infty} \int_0^{\infty} \hat{Z}_i(\sigma_i) \hat{Z}_j(\sigma_j) \sigma_i \sigma_j d\sigma_i d\sigma_j \\
&= -\ln \mathcal{Z}_I - \sum_{i \in I} \left(\boldsymbol{Q}\boldsymbol{\sigma}^{MP} + \boldsymbol{b}\right)_i \mu_i - \frac{1}{2} \sum_{i \in I, j \in I} Q_{ij} \sigma_i \sigma_j,
\end{aligned} \tag{C.2}$$

where $\mathcal{Z}_I$ is the normalization to ensure that $\int_{\boldsymbol{\sigma}_I \succeq 0} Z_I(\boldsymbol{\sigma}_I) d\boldsymbol{\sigma}_I = 1$, and μ_i is the mean of σ_i under the distribution $\hat{Z}_i(\sigma_i)$. Combining (C.1) and (C.2), we can obtain the minimization problem as (5.20).