Chapter 2
Minimum-Norm-Based Source Imaging Algorithms

2.1 Introduction

In this chapter, we describe the minimum-norm and related methods, which are classic algorithms for electromagnetic brain imaging [1, 2]. In this chapter, the minimum-norm method is first formulated based on the maximum-likelihood principle, and the properties of the minimum-norm solution are discussed. This discussion leads to the necessity of regularization when implementing the minimum-norm method. We discuss two different representative regularization methods: the L_2-norm regularization and the L_1-norm regularization. The minimum-norm method is, then, formulated based on Bayesian inference—Bayesian formulation providing a form of the minimum-norm method where the regularization is already embedded.

2.2 Definitions

In electromagnetic brain imaging, we use an array of sensors to obtain bioelectromagnetic measurements. We define the output of the mth sensor at time t as $y_m(t)$, and the column vector containing outputs from all sensors, such that

$$\boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_M(t) \end{bmatrix}, \tag{2.1}$$

where M is the total number of sensors. This column vector $\boldsymbol{y}(t)$ expresses the outputs of the sensor array, and it may be called the data vector or array measurement.

K. Sekihara and S.S. Nagarajan, *Electromagnetic Brain Imaging*,
DOI 10.1007/978-3-319-14947-9_2

A spatial location is represented by a three-dimensional vector $\boldsymbol{r}$: $\boldsymbol{r} = (x, y, z)$. A source vector is defined as a three-dimensional column vector $\boldsymbol{s}(\boldsymbol{r}, t)$:

$$\boldsymbol{s}(\boldsymbol{r}, t) = \begin{bmatrix} s_x(\boldsymbol{r}, t) \\ s_y(\boldsymbol{r}, t) \\ s_z(\boldsymbol{r}, t) \end{bmatrix}, \tag{2.2}$$

where $s_x(\boldsymbol{r}, t)$, $s_y(\boldsymbol{r}, t)$, and $s_z(\boldsymbol{r}, t)$ are the x, y, and z components. The physical nature of the source vector is the electromotive force generated by neuronal activities in the brain. Additional discussion regarding the nature of the sources is presented in Sect. A.1 in the Appendix. The magnitude of the source vector is denoted as a scalar $s(\boldsymbol{r}, t)$, and the orientation of the source is denoted as a three-dimensional unit vector $\boldsymbol{\eta}(\boldsymbol{r}) = [\eta_x(\boldsymbol{r}), \eta_y(\boldsymbol{r}), \eta_z(\boldsymbol{r})]^T$, where the superscript T indicates the matrix transpose. Then, the relationship

$$\boldsymbol{s}(\boldsymbol{r}, t) = s(\boldsymbol{r}, t)\boldsymbol{\eta}(\boldsymbol{r}) = s(\boldsymbol{r}, t) \begin{bmatrix} \eta_x(\boldsymbol{r}) \\ \eta_y(\boldsymbol{r}) \\ \eta_z(\boldsymbol{r}) \end{bmatrix} \tag{2.3}$$

holds.

2.3 Sensor Lead Field

We assume that a unit-magnitude source exists at $\boldsymbol{r}$. We denote the output of the mth sensor due to this unit-magnitude source as $l_m^x(\boldsymbol{r})$, $l_m^y(\boldsymbol{r})$, and $l_m^z(\boldsymbol{r})$ when the unit-magnitude source is directed in the x, y, and z directions, respectively. The column vectors $\boldsymbol{l}_x(\boldsymbol{r})$, $\boldsymbol{l}_y(\boldsymbol{r})$, and $\boldsymbol{l}_z(\boldsymbol{r})$ are defined as

$$\begin{aligned} \boldsymbol{l}_x(\boldsymbol{r}) &= [l_1^x(\boldsymbol{r}), l_2^x(\boldsymbol{r}), \ldots, l_M^x(\boldsymbol{r})]^T, \\ \boldsymbol{l}_y(\boldsymbol{r}) &= [l_1^y(\boldsymbol{r}), l_2^y(\boldsymbol{r}), \ldots, l_M^y(\boldsymbol{r})]^T, \\ \boldsymbol{l}_z(\boldsymbol{r}) &= [l_1^z(\boldsymbol{r}), l_2^z(\boldsymbol{r}), \ldots, l_M^z(\boldsymbol{r})]^T. \end{aligned}$$

These vectors express the sensor array sensitivity for a source located at $\boldsymbol{r}$ and directed in the x, y, and z directions. Using these column vectors, the sensitivity of the whole sensor array for a source at $\boldsymbol{r}$ is expressed using an $M \times 3$ matrix:

$$\boldsymbol{L}(\boldsymbol{r}) = [\boldsymbol{l}_x(\boldsymbol{r}), \boldsymbol{l}_y(\boldsymbol{r}), \boldsymbol{l}_z(\boldsymbol{r})]. \tag{2.4}$$

This matrix $\boldsymbol{L}(\boldsymbol{r})$ is called the lead-field matrix. We also define the lead-field vector, $\boldsymbol{l}(\boldsymbol{r})$, that expresses the sensitivity of the sensor array in a particular source direction $\boldsymbol{\eta}(\boldsymbol{r})$, such that

$$\boldsymbol{l}(\boldsymbol{r}) = \boldsymbol{L}(\boldsymbol{r})\boldsymbol{\eta}(\boldsymbol{r}). \tag{2.5}$$

The problem of estimating the sensor lead field is referred to as the bioelectromagnetic forward problem. Arguments on how to compute the sensor lead field are presented in Appendix A.

2.4 Voxel Source Model and Tomographic Source Reconstruction

Using the lead-field matrix in Eq. (2.4), the relationship between the sensor data, $\boldsymbol{y}(t)$, and the source vector, $\boldsymbol{s}(\boldsymbol{r}, t)$, is expressed as

$$\boldsymbol{y}(t) = \int_{\Omega} \boldsymbol{L}(\boldsymbol{r})\boldsymbol{s}(\boldsymbol{r}, t)\, d\boldsymbol{r}. \tag{2.6}$$

Here, $d\boldsymbol{r}$ indicates the volume element, and the integral is performed over a volume where sources are assumed to exist. This volume is called the source space, which is denoted Ω. Equation (2.6) expresses the relationship between the sensor outputs $\boldsymbol{y}(t)$ and the source distribution $\boldsymbol{s}(\boldsymbol{r}, t)$.

The bioelectromagnetic inverse problem is the problem of estimating the source-vector spatial distribution, $\boldsymbol{s}(\boldsymbol{r}, t)$, from the measurements, $\boldsymbol{y}(t)$. Here, we assume that we know the sensor lead field $\boldsymbol{L}(\boldsymbol{r})$, although our knowledge of the sensor lead field is to some degree imperfect because it must be estimated using an analytical model or numerical computations.

When estimating $\boldsymbol{s}(\boldsymbol{r}, t)$ from $\boldsymbol{y}(t)$, $\boldsymbol{s}(\boldsymbol{r}, t)$ is continuous in space, while $\boldsymbol{y}(t)$ is discrete in space. A common strategy here is to introduce voxel discretization over the source space. Let us define the number of voxels as N, and the locations of the voxels are denoted as $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N$. Then, the discrete form of Eq. (2.6) is expressed as:

$$\boldsymbol{y}(t) = \sum_{j=1}^{N} \boldsymbol{L}(\boldsymbol{r}_j)\boldsymbol{s}(\boldsymbol{r}_j, t) = \sum_{j=1}^{N} \boldsymbol{L}(\boldsymbol{r}_j)\boldsymbol{s}_j(t). \tag{2.7}$$

where the source vector at the jth voxel, $\boldsymbol{s}(\boldsymbol{r}_j, t)$, is denoted $\boldsymbol{s}_j(t)$ for simplicity. We introduce the augmented lead-field matrix over all voxel locations as

$$\boldsymbol{F} = [\boldsymbol{L}(\boldsymbol{r}_1), \boldsymbol{L}(\boldsymbol{r}_2), \ldots, \boldsymbol{L}(\boldsymbol{r}_N)], \tag{2.8}$$

which is an $M \times 3N$ matrix. We define a $3N \times 1$ column vector containing the source vectors at all voxel locations, $\boldsymbol{x}(t)$, such that

$$\boldsymbol{x}(t) = \begin{bmatrix} \boldsymbol{s}_1(t) \\ \boldsymbol{s}_2(t) \\ \vdots \\ \boldsymbol{s}_N(t) \end{bmatrix}. \tag{2.9}$$

Equation (2.7) is then rewritten as

$$\boldsymbol{y}(t) = [\boldsymbol{L}(\boldsymbol{r}_1), \boldsymbol{L}(\boldsymbol{r}_2), \ldots, \boldsymbol{L}(\boldsymbol{r}_N)] \begin{bmatrix} \boldsymbol{s}_1(t) \\ \boldsymbol{s}_2(t) \\ \vdots \\ \boldsymbol{s}_N(t) \end{bmatrix} = \boldsymbol{F}\boldsymbol{x}(t). \tag{2.10}$$

Here, since the augmented lead-field matrix $\boldsymbol{F}$ is a known quantity, the only unknown quantity is the $3N \times 1$ column vector, $\boldsymbol{x}(t)$. This vector $\boldsymbol{x}(t)$ is called the voxel source vector.

The spatial distribution of the source orientation, $\boldsymbol{\eta}(\boldsymbol{r})$, may be a known quantity if accurate subject anatomical information (such as high-precision subject MRI) can be obtained with accurate co-registration between the MRI coordinate and the sensor coordinate. In this case, the inverse problem is the problem of estimating the source magnitude, $s(\boldsymbol{r}, t)$, instead of the source vector, $\boldsymbol{s}(\boldsymbol{r}, t)$. Let us consider a situation in which the source orientations at all voxel locations are predetermined. Defining the orientation of a source at the jth voxel as $\boldsymbol{\eta}_j$, the lead field at the jth voxel is expressed as the column vector $\boldsymbol{l}_j$, which is obtained as $\boldsymbol{l}_j = \boldsymbol{L}(\boldsymbol{r}_j)\boldsymbol{\eta}_j$, according to Eq. (2.5). Thus, the augmented lead field is expressed as an $M \times N$ matrix $\boldsymbol{H}$ defined such that

$$\boldsymbol{H} = [\boldsymbol{L}(\boldsymbol{r}_1)\boldsymbol{\eta}_1, \boldsymbol{L}(\boldsymbol{r}_2)\boldsymbol{\eta}_2, \ldots, \boldsymbol{L}(\boldsymbol{r}_N)\boldsymbol{\eta}_N] = [\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_N], \tag{2.11}$$

whereby Eq. (2.10) can be reduced as follows:

$$\begin{aligned} \boldsymbol{y}(t) &= [\boldsymbol{L}(\boldsymbol{r}_1), \boldsymbol{L}(\boldsymbol{r}_2), \ldots, \boldsymbol{L}(\boldsymbol{r}_N)] \begin{bmatrix} \boldsymbol{s}_1(t) \\ \boldsymbol{s}_2(t) \\ \vdots \\ \boldsymbol{s}_N(t) \end{bmatrix} \\ &= [\boldsymbol{L}(\boldsymbol{r}_1), \boldsymbol{L}(\boldsymbol{r}_2), \ldots, \boldsymbol{L}(\boldsymbol{r}_N)] \begin{bmatrix} \boldsymbol{\eta}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\eta}_2 & \cdot & \vdots \\ \vdots & \cdot & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\eta}_N \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{bmatrix} \\ &= [\boldsymbol{L}(\boldsymbol{r}_1)\boldsymbol{\eta}_1, \boldsymbol{L}(\boldsymbol{r}_2)\boldsymbol{\eta}_2, \ldots, \boldsymbol{L}(\boldsymbol{r}_N)\boldsymbol{\eta}_N] \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{bmatrix} = \boldsymbol{H}\boldsymbol{x}(t). \end{aligned} \tag{2.12}$$

Thus, the voxel source vector $\boldsymbol{x}(t)$, in this case, is an $N \times 1$ column vector,

$$\boldsymbol{x}(t) = \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{bmatrix}, \tag{2.13}$$

in which the jth component of $\boldsymbol{x}(t)$ is $s_j(t)$, which is the scalar intensity at the jth voxel. In this book, the same notation $\boldsymbol{x}(t)$ is used to indicate either the $3N \times 1$ vector in Eq. (2.9) or the $N \times 1$ vector in Eq. (2.13), unless any confusion arises.

In summary, denoting the additive noise in the sensor data ε, the relationship between the sensor data $\boldsymbol{y}(t)$ and the voxel source vector $\boldsymbol{x}(t)$ is expressed as

$$\boldsymbol{y}(t) = \boldsymbol{F}\boldsymbol{x}(t) + \varepsilon, \tag{2.14}$$

where $\boldsymbol{x}(t)$ is a $3N \times 1$ column vector in Eq. (2.9). When voxels have predetermined orientations, using the augmented lead field matrix $\boldsymbol{H}$ in Eq. (2.11), the relationship between $\boldsymbol{y}(t)$ and $\boldsymbol{x}(t)$ is expressed as

$$\boldsymbol{y}(t) = \boldsymbol{H}\boldsymbol{x}(t) + \varepsilon, \tag{2.15}$$

where $\boldsymbol{x}(t)$ is an $N \times 1$ column vector in Eq. (2.13).

2.5 Maximum Likelihood Principle and the Least-Squares Method

When estimating the unknown quantity $\boldsymbol{x}$ from the sensor data $\boldsymbol{y}$, the basic principle is to interpret the data $\boldsymbol{y}$ as a realization of most probable events. That is, the sensor data $\boldsymbol{y}$ is considered the result of the most likely events. We call this the maximum likelihood principle. In this chapter, we first derive the maximum likelihood solution of the unknown source vector $\boldsymbol{x}$.

We assume that the noise distribution is Gaussian, i.e.,

$$\varepsilon \sim \mathcal{N}(\varepsilon|\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Namely, the noise in the sensor data is the identically and independently distributed Gaussian noise with a mean of zero, and the same variance σ^2. According to (C.1) in the Appendix, the explicit form of the noise probability distribution is given by

$$p(\varepsilon) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left[-\frac{1}{2\sigma^2}\|\varepsilon\|^2\right]. \tag{2.16}$$

Since the linear relationship in Eq. (2.14) holds, the probability distribution of the sensor data $\boldsymbol{y}(t)$ is expressed as

$$p(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left[-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2\right], \tag{2.17}$$

where the explicit time notation (t) is omitted from the vector notations $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ for simplicity.[1]

This $p(\boldsymbol{y})$ as a function of the unknown parameter $\boldsymbol{x}$ is called the likelihood function, and the maximum likelihood estimate $\widehat{\boldsymbol{x}}$ is obtained such that[2]

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\text{argmax}} \log p(\boldsymbol{y}), \tag{2.18}$$

where $\log p(\boldsymbol{y})$ is called the log-likelihood function. Using the probability distribution in Eq. (2.17), the log-likelihood function $\log p(\boldsymbol{y})$ is expressed as

$$\log p(\boldsymbol{y}) = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 + \mathcal{C}, \tag{2.19}$$

where $\mathcal{C}$ expresses terms that do not contain $\boldsymbol{x}$. Therefore, the $\boldsymbol{x}$ that maximizes $\log p(\boldsymbol{y})$ is equal to the one that minimizes $\mathcal{F}(\boldsymbol{x})$ defined such that

$$\mathcal{F}(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2. \tag{2.20}$$

That is, the maximum likelihood solution $\widehat{\boldsymbol{x}}$ is obtained using

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\text{argmin}} \mathcal{F}(\boldsymbol{x}) : \quad \text{where} \quad \mathcal{F} = \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2. \tag{2.21}$$

This $\mathcal{F}(\boldsymbol{x})$ in Eq. (2.20) is referred to as the least-squares cost function, and the method that estimates $\boldsymbol{x}$ through the minimization of the least-squares cost function is the method of least-squares.

2.6 Derivation of the Minimum-Norm Solution

In the bioelectromagnetic inverse problem, the number of voxels N, in general, is much greater than the number of sensors M. Thus, the estimation of the source vector $\boldsymbol{x}$ is an ill-posed problem. When applying the least-squares method to such an ill-posed problem, the problem arises that an infinite number of $\boldsymbol{x}$ could make the cost

[1] For the rest of this chapter, the explicit time notation is omitted from these vector notations, unless otherwise noted.

[2] The notation argmax indicates the value of $\boldsymbol{x}$ that maximizes $\log p(\boldsymbol{y})$ which is an implicit function of $\boldsymbol{x}$.

function equal to zero. Therefore, we cannot obtain an optimum solution of $\boldsymbol{x}$ based only on the least-squares method.

A general strategy for overcoming this problem is to integrate a "desired property" of the unknown parameter $\boldsymbol{x}$ into the estimation problem. That is, we choose $\boldsymbol{x}$ so as to maximize this "desired property," and also satisfy $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$. Quite often, a small norm of the solution vector is used as this "desired property," and in this case, the optimum estimate $\widehat{\boldsymbol{x}}$ is obtained using

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{x}\|^2 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}. \tag{2.22}$$

In the optimization above, the notation of "subject to" indicates a constraint, (i.e., the above optimization requires that the estimate $\widehat{\boldsymbol{x}}$ be chosen such that $\boldsymbol{x}$ minimizes $\|\boldsymbol{x}\|^2$ as well as satisfies $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$.) To solve the constraint optimization problem in Eq. (2.22), we use the method of Lagrange multipliers that can convert a constrained optimization problem to an unconstrained optimization problem. In this method, using an $M \times 1$ column vector $\boldsymbol{c}$ as the Lagrange multipliers, we define a function called the Lagrangian $\mathbb{L}(\boldsymbol{x}, \boldsymbol{c})$ such that

$$\mathbb{L}(\boldsymbol{x}, \boldsymbol{c}) = \|\boldsymbol{x}\|^2 + \boldsymbol{c}^T (\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}). \tag{2.23}$$

The solution $\widehat{\boldsymbol{x}}$ is obtained by minimizing $\mathbb{L}(\boldsymbol{x}, \boldsymbol{c})$ above with respect to $\boldsymbol{x}$ and $\boldsymbol{c}$—the solution $\widehat{\boldsymbol{x}}$ being equal to $\widehat{\boldsymbol{x}}$ obtained by solving the constrained optimization in Eq. (2.22).

To derive an $\boldsymbol{x}$ that minimizes Eq. (2.23), we compute the derivatives of $\mathbb{L}(\boldsymbol{x}, \boldsymbol{c})$ with respect to $\boldsymbol{x}$ and $\boldsymbol{c}$, and set them to be zero, giving

$$\frac{\partial \mathbb{L}(\boldsymbol{x}, \boldsymbol{c})}{\partial \boldsymbol{x}} = 2\boldsymbol{x} - \boldsymbol{F}^T \boldsymbol{c} = 0, \tag{2.24}$$

$$\frac{\partial \mathbb{L}(\boldsymbol{x}, \boldsymbol{c})}{\partial \boldsymbol{c}} = \boldsymbol{y} - \boldsymbol{F}\boldsymbol{x} = 0. \tag{2.25}$$

Using the equations above, we can derive

$$\widehat{\boldsymbol{x}} = \boldsymbol{F}^T \left(\boldsymbol{F}\boldsymbol{F}^T\right)^{-1} \boldsymbol{y}. \tag{2.26}$$

The solution in Eq. (2.26) is called the minimum-norm solution, which is well known as a solution for the ill-posed linear inverse problem.

2.7 Properties of the Minimum-Norm Solution

The minimum-norm solution is expressed as

$$\widehat{\boldsymbol{x}} = \boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^{-1}(\boldsymbol{F}\boldsymbol{x} + \boldsymbol{\varepsilon}) = \boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^{-1}\boldsymbol{F}\boldsymbol{x} + \boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^{-1}\boldsymbol{\varepsilon}. \tag{2.27}$$

The first term on the right-hand side is expressed as $E(\widehat{\boldsymbol{x}})$, which indicates the expectation of $\widehat{\boldsymbol{x}}$. This term represents how the solution deviates from its true value even in the noiseless cases . The second term indicates the influence of the noise $\boldsymbol{\varepsilon}$. The first term is rewritten as

$$E(\widehat{\boldsymbol{x}}) = \boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^{-1}\boldsymbol{F}\boldsymbol{x} = \boldsymbol{Q}\boldsymbol{x}, \tag{2.28}$$

where

$$\boldsymbol{Q} = \boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^{-1}\boldsymbol{F}. \tag{2.29}$$

Apparently, the first term is not equal to the true value $\boldsymbol{x}$, and the matrix $\boldsymbol{Q}$ in Eq. (2.29) expresses the relationship between the true value $\boldsymbol{x}$ and the estimated value $E(\widehat{\boldsymbol{x}})$.

Denoting the (i, j)th element of $\boldsymbol{Q}$ as $Q_{i,j}$, the jth element of $E(\widehat{\boldsymbol{x}})$, $E(\widehat{x}_j)$, is expressed as

$$E(\widehat{x}_j) = \sum_{k=1}^{N} Q_{j,k}x_k. \tag{2.30}$$

The above equation shows how each element of the true vector $\boldsymbol{x}$ affects the value of $E(\widehat{x}_j)$. That is, $Q_{j,k}$ expresses the amount of leakage of x_k into $\widehat{x}_j$ when $j \neq k$. If the weight $Q_{j,1}, \ldots, Q_{j,N}$ has a sharp peak at j, $\widehat{x}_j$ may be close to the true value x_j. If the weight has no clear peak or if the weight has a peak at j' that is different from j, $\widehat{x}_j$ may be very different from x_j. Because of such properties, the matrix $\boldsymbol{Q}$ is called the resolution matrix.

We next examine the second term, which expresses the noise influence. The noise influence is related to the singular values of $\boldsymbol{F}$. The singular value decomposition of $\boldsymbol{F}$ is defined as

$$\boldsymbol{F} = \sum_{j=1}^{M} \gamma_j \boldsymbol{u}_j \boldsymbol{v}_j^T, \tag{2.31}$$

where we assume that $M < N$, and the singular values are numbered in decreasing order. Using

$$\boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^{-1} = \sum_{j=1}^{N} \frac{1}{\gamma_j}\boldsymbol{v}_j\boldsymbol{u}_j^T, \tag{2.32}$$

we can express the second term in Eq. (2.27) as

$$\sum_{j=1}^{N} \frac{(\boldsymbol{u}_j^T\boldsymbol{\varepsilon})}{\gamma_j}\boldsymbol{v}_j. \tag{2.33}$$

The equation above shows that the denominator contains the singular values. Thus, if higher order singular values are very small and close to zero, the terms containing such small singular values amplify the noise influence, resulting in a situation where

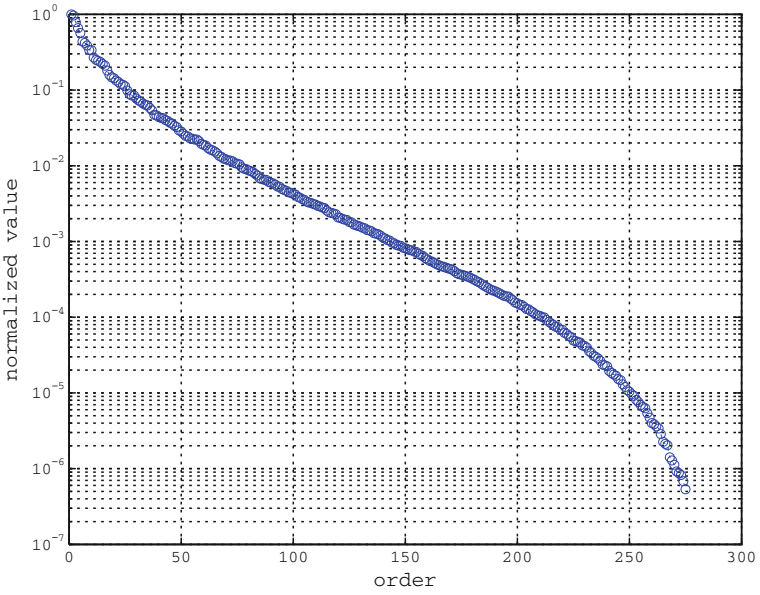

Fig. 2.1 Typical plot of the singular values of the lead field matrix $\boldsymbol{F}$. We assume the 275-channel CTF Omega whole-head MEG sensor system (VMS MedTech Ltd., BC, Canada). A typical location of the subject head relative to the whole head sensor array is assumed. An $8 \times 8 \times 10$ cm region is also assumed as the source space within the subject's head. The spherical homogeneous conductor model is used for computing the sensor lead field. The singular values are normalized with the maximum (i.e., the first) singular value

the second term is dominated in Eq. (2.27), and the minimum norm solution would contain large errors due to the noise.

A plot of a typical singular-value spectrum of the lead field matrix $\boldsymbol{F}$ is shown in Fig. 2.1. To obtain the plot, we used the sensor array of the 275-channel CTF Omega whole-head MEG sensor system (VMS MedTech Ltd., BC, Canada) and spherical homogeneous conductor model to compute the sensor lead field [3].[3] The plot shows that higher order singular values of the lead field matrix are very small. In Fig. 2.1, the ratio of the maximum and minimum singular values reaches the order of 10^{-7}. Therefore, the minimum-norm method in Eq. (2.26) generally produces results highly susceptible to the noise in the sensor data.

2.8 L_2-Regularized Minimum-Norm Solution

When a large amount of noise is overlapped onto the sensor data $\boldsymbol{y}$, if we seek a solution that satisfies $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$, the resultant solution $\boldsymbol{x}$ would be severely affected by the noise. In other words, when noise exists in the sensor data, it is more or less

[3] Computing the lead field using the spherical homogeneous conductor model is explained in Sect. A.2.4 in the Appendix.

meaningless to impose the constraint $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$, so, instead of using the optimization in Eq. (2.22), we should use

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{x}\|^2 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 \leq d, \tag{2.34}$$

where d is a positive constant. In Eq. (2.34), the condition $\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 \leq d$ does not require $\boldsymbol{F}\boldsymbol{x}$ to be exactly equal to $\boldsymbol{y}$, but allow $\boldsymbol{F}\boldsymbol{x}$ to be different from $\boldsymbol{y}$ within a certain range specified by d. Therefore, the solution $\widehat{\boldsymbol{x}}$ is expected to be less affected by the noise in the sensor data $\boldsymbol{y}$.

Unfortunately, there is no closed-form solution for the optimization problem in Eq. (2.34), because of the inequality constraint. Although we can solve Eq. (2.34) numerically, we proceed in solving it by replacing the inequality constraint with the equality constraint. This is possible because the solution of Eq. (2.34) generally exists on the border of the constraint. Thus, we can change the optimization problem in Eq. (2.34) to

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{x}\|^2 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 = d. \tag{2.35}$$

Since this is an equality-constraint problem, we can use the method of Lagrange multipliers. Using the Lagrange multiplier λ, the Lagrangian is defined as

$$\mathbb{L}(\boldsymbol{x}, \boldsymbol{c}) = \|\boldsymbol{x}\|^2 + \lambda \left(\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 - d \right). \tag{2.36}$$

Thus, the solution $\widehat{\boldsymbol{x}}$ is given as

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\ \mathbb{L}(\boldsymbol{x}, \boldsymbol{c}) = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left[\|\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 \right]. \tag{2.37}$$

In the above expression, we disregard the term $-\lambda d$, which does not affect the results of the minimization. Also, we can see that the multiplier λ works as a balancer between the L_2-norm[4] term $\|\boldsymbol{x}\|^2$ and the squared error term $\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2$.

To derive the solution of $\boldsymbol{x}$ that minimizes $\mathbb{L}(\boldsymbol{x}, \boldsymbol{c})$, we compute the derivative of $\mathbb{L}(\boldsymbol{x}, \boldsymbol{c})$ with respect to $\boldsymbol{x}$ and set it to zero, i.e.,

$$\begin{aligned} \frac{\mathbb{L}(\boldsymbol{x}, \boldsymbol{c})}{\partial \boldsymbol{x}} &= \frac{1}{\partial \boldsymbol{x}} \left(\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{x}^T \boldsymbol{F}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{F}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{F}^T \boldsymbol{F}\boldsymbol{x} + \xi \boldsymbol{x}^T \boldsymbol{x} \right) \\ &= -2\boldsymbol{F}^T \boldsymbol{y} + 2 \left(\boldsymbol{F}^T \boldsymbol{F} + \xi \boldsymbol{I} \right) \boldsymbol{x} = \boldsymbol{0}, \end{aligned} \tag{2.38}$$

where we use $1/\lambda = \xi$. We can then derive

$$\widehat{\boldsymbol{x}} = \left(\boldsymbol{F}^T \boldsymbol{F} + \xi \boldsymbol{I} \right)^{-1} \boldsymbol{F}^T \boldsymbol{y}. \tag{2.39}$$

[4] A brief summary of the norm of vectors is presented in Sect. C.4 in the Appendix.

Using the matrix inversion lemma in Eq. (C.92), we obtain

$$\widehat{\boldsymbol{x}} = \boldsymbol{F}^T \left(\boldsymbol{F}\boldsymbol{F}^T + \xi \boldsymbol{I} \right)^{-1} \boldsymbol{y}. \tag{2.40}$$

The solution in Eq. (2.40) is called the L_2-norm-regularized minimum-norm solution, or simply L_2-regularized minimum-norm solution.

Let us compute the noise influence term for the L_2-regularized minimum-norm solution. Using Eq. (2.31), we have

$$\boldsymbol{F}^T \left(\boldsymbol{F}\boldsymbol{F}^T + \xi \boldsymbol{I} \right)^{-1} = \sum_{j=1}^{N} \frac{\gamma_j}{\gamma_j^2 + \xi} \boldsymbol{v}_j \boldsymbol{u}_j^T, \tag{2.41}$$

and the L_2-regularized minimum-norm solution is expressed as

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \boldsymbol{F}^T \left(\boldsymbol{F}\boldsymbol{F}^T + \xi \boldsymbol{I} \right)^{-1} (\boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{F}^T \left(\boldsymbol{F}\boldsymbol{F}^T + \xi \boldsymbol{I} \right)^{-1} \boldsymbol{H}\boldsymbol{x} + \sum_{j=1}^{N} \frac{\gamma_j}{\gamma_j^2 + \xi} \boldsymbol{v}_j \boldsymbol{u}_j^T \boldsymbol{\varepsilon}. \end{aligned} \tag{2.42}$$

The second term, expressing the influence of noise, is

$$\sum_{j=1}^{N} \frac{\gamma_j (\boldsymbol{u}_j^T \boldsymbol{\varepsilon})}{\gamma_j^2 + \xi} \boldsymbol{v}_j. \tag{2.43}$$

In the expression above, the denominator contains the positive constant ξ, and it is easy to see that this ξ prevents the terms with smaller singular values from being amplified.

One problem here is how to choose an appropriate value for ξ. Our argument above only suggests that if the noise is large, we need a large ξ, but if small, a smaller ξ can be used. However, the arguments above do not lead to the derivation of an appropriate ξ. We will return to this problem in Sect. 2.10.2 where L_2-regularized minimum-norm solution is re-derived based on a Bayesian formulation, in which deriving the optimum ξ is embedded.

2.9 L_1-Regularized Minimum-Norm Solution

2.9.1 L_1-Norm Constraint

In the preceding section, we derived a solution that minimizes the L_2-norm of the solution vector $\boldsymbol{x}$. In this section, we argue for a solution that minimizes the L_1-norm of $\boldsymbol{x}$, which is defined in Eq. (C.64). The L_1-norm-regularized solution is obtained

using [4–6]

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{j=1}^{N} |x_j| \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 = d. \tag{2.44}$$

The only difference between the equation above and Eq. (2.35) is to minimize either L_2 norm $\|\boldsymbol{x}\|^2$ in Eq. (2.35) or L_1 norm, $\|\boldsymbol{x}\|_1 = \sum_j |x_j|$ in Eq. (2.44). Although it may look as if there is no significant difference between the two methods, the results of source estimation are significantly different. The L_1-norm regularization gives a "so-called" sparse solution, in which only few x_j have nonzero values and a majority of other x_j have values close to zero.

Using the method of Lagrange multipliers and following exactly the same arguments as in Sect. 2.8, the L_1-norm solution can be obtained by minimizing the cost function $\mathcal{F}$, i.e.,

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, \mathcal{F}: \quad \mathcal{F} = \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 + \xi \sum_{j=1}^{N} |x_j|, \tag{2.45}$$

where again ξ is a positive constant that controls the balance between the first and the second terms in the cost function above. Unfortunately, the minimization problem in Eq. (2.45) does not have a closed-form solution, so numerical methods are used here to obtain the solution $\widehat{\boldsymbol{x}}$.

2.9.2 Intuitive Explanation for Sparsity

Actually, it is not easy to provide an intuitive explanation regarding why the optimization in Eq. (2.44) or (2.45) causes a sparse solution. The straightforward (and intuitively clear) formulation to obtain a sparse solution should use the L_0-norm minimization, such that

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{j=1}^{N} \mathcal{T}(x_j) \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 = d, \tag{2.46}$$

where the function $\mathcal{T}(x)$ is defined in Eq. (C.65). In the above formulation, since $\sum_{j=1}^{N} \mathcal{T}(x_j)$ indicates the number of nonzero x_j, $\widehat{\boldsymbol{x}}$ is the solution that has the smallest number of nonzero x_j and still satisfies $\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 = d$. The optimization problem in Eq. (2.46) is known to require impractically long computational time. The optimization for the L_1-norm cost function in Eq. (2.44) approximates this L_0-norm optimization in Eq. (2.46) so as to obtain a sparse solution within a reasonable range of computational time [7].

The regularization methods mentioned above can be summarized to have a form of the cost functions expressed as

$$\mathcal{F} = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \xi\phi(\boldsymbol{x}). \tag{2.47}$$

The first term is the data-fitting term, and the second term $\phi(\boldsymbol{x})$ expresses the constraint, which has the following form for general L_p-norm cases ($0 \le p \le 1$):

$$\phi(\boldsymbol{x}) = \sum_{j=1}^{N} \mathcal{T}(x_j) \quad \text{for } L_0\text{-norm}, \tag{2.48}$$

$$\phi(\boldsymbol{x}) = \sum_{j=1}^{N} |x_j| \quad \text{for } L_1\text{-norm}, \tag{2.49}$$

$$\phi(\boldsymbol{x}) = \left[\sum_{j=1}^{N} x_j^p\right]^{1/p} \quad \text{for } L_p\text{-norm}. \tag{2.50}$$

The plots of $\phi(\boldsymbol{x})$ with respect to the x_j axis are shown in Fig. 2.2. In this figure, the four kinds of plots of $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|_p$ when $p = 0$, $p = 0.3$, $p = 0.7$, and $p = 1$

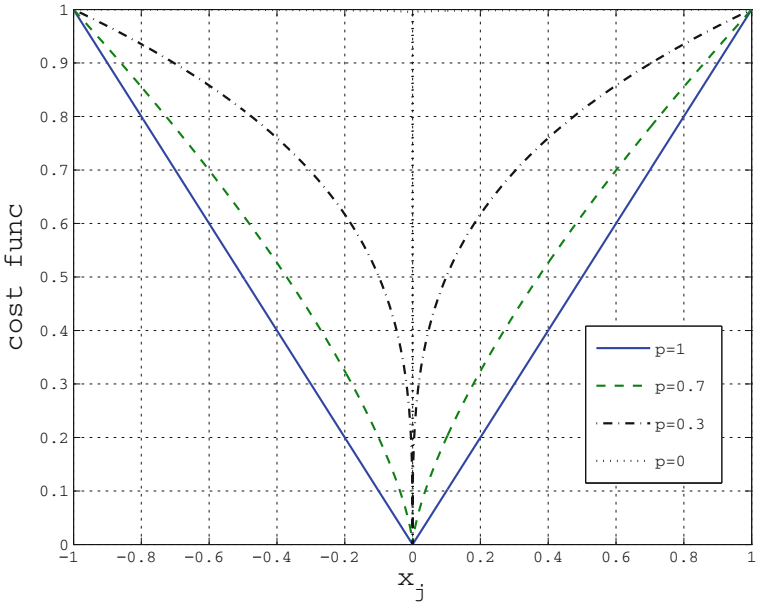

Fig. 2.2 Plots of objective function $\phi(\boldsymbol{x})$ defined in Eqs. (2.48)–(2.50) with respect to the x_j axis. The four cases of $p = 0$, $p = 0.3$, $p = 0.7$, and $p = 1$ are shown. The cases of $p = 0$, and $p = 1$ correspond to the L_0 and L_1 norm constraints (A brief summary of the norm of vectors is presented in Sect. C.4 in the Appendix.)

are shown. It can be seen in this figure that the L_0-norm constraint is approximated by the L_p-norm constraint, and as p becomes closer to 0, the L_p-norm provides a better approximation.

Let us see how the L_p-norm regularization causes sparse solutions when $0 \leq p \leq 1$. To do so, we consider a simplest estimation problem in which only two voxels exist and the voxels have source intensity of x_1 and x_2. We assume a noiseless measurement using a single-sensor; the sensor data being represented by a scalar y. The optimization for the L_1-norm solution is expressed in this case as

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left(|x_1| + |x_2|\right) \quad \text{subject to} \quad y = h_1 x_1 + h_2 x_2, \tag{2.51}$$

where $\widehat{\boldsymbol{x}} = (\widehat{x}_1, \widehat{x}_2)^T$, and h_1 and h_2 are the sensor lead field. For the sake of comparison, we also argue the L_2-norm regularization whose optimization is given as follows:

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left(x_1^2 + x_2^2\right)^{1/2} \quad \text{subject to} \quad y = h_1 x_1 + h_2 x_2. \tag{2.52}$$

The optimization process is depicted in Fig. 2.3. In Fig. 2.3a, the tetragon at the center represents the L_1-norm objective function, $|x_1| + |x_2| =$ constant. The broken line represents the x_1 and x_2 that satisfy the measurement equation $y = h_1 x_1 + h_2 x_2$. Thus, as a result of the optimization in Eq. (2.51), the x_1 and x_2 on the broken line that minimize $|x_1| + |x_2|$ should be chosen as the solution, i.e., the point (x_1, x_2) at which the tetragon touches the broken line is chosen as the solution for the optimization. Such solution is indicated by the small filled circle in Fig. 2.3a. In this solution, x_2 has a nonzero value but x_1 is zero, i.e., a sparse solution is obtained. It can be seen in this figure that in most cases, the point at which the tetragon touches the broken

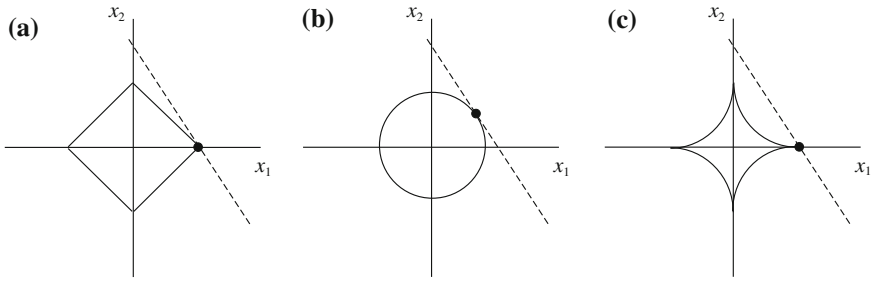

Fig. 2.3 The optimization process is depicted for the simple case in which a single sensor and two voxels exist. Source magnitudes at the voxels are represented by x_1 and x_2. The broken lines represent the x_1 and x_2 that satisfy the measurement equation, $y = h_1 x_1 + h_2 x_2$. The *filled black circles* indicate an example of the solution for each case. **a** L_1-norm regularization in Eq. (2.51). The tetragon at the center represents the L_1-norm objective function $|x_1| + |x_2| =$ constant. **b** L_2-norm regularization in Eq. (2.52). The circle at the center represents the L_2-norm objective function $x_1^2 + x_2^2 =$ constant. **c** L_p-norm regularization where $0 < p < 1$

line is likely to be located at one of its vertices, so a sparse solution is likely to be obtained.

Figure 2.3b shows the case of the L_2-norm minimization in Eq. (2.52). In this figure, the broken line again represents the x_1 and x_2 that satisfy the measurement equation $y = h_1 x_1 + h_2 x_2$, and the circle represents the L_2-norm objective function $x_1^2 + x_2^2 =$ constant. In this case, the x_1 and x_2 on the broken line that minimizes $x_1^2 + x_2^2$ should be chosen, and the resultant solution is (x_1, x_2) at which the circle touches the broken line. An example of such solution is indicated by the small filled circle. In this case, both x_1 and x_2 have nonzero values, and a non-sparse solution is likely to be obtained using L_2-norm regularization.

Finally, Fig. 2.3c shows a case of the general L_p norm minimization ($0 < p < 1$). An example of such solution is indicated by the small, filled circle. Using the general L_p norm regularization, the solution is more likely to be sparse than the case of the L_1-norm minimization. However, the computational burden for the general L_p norm minimization is so high that it is seldom used in practical applications.

2.9.3 Problem with Source Orientation Estimation

When applying the L_1-norm regularization to the bioelectromagnetic source localization, it has been known that the method fails in estimating correct source orientations. The reason for this is described as follows: The components of the solution vector $\boldsymbol{x}$ is denoted explicitly as

$$\boldsymbol{x} = \left[s_1^x, s_1^y, s_1^z, \ldots, s_j^x, s_j^y, s_j^z, \ldots, s_N^x, s_N^y, s_N^z \right]^T,$$

where s_j^x, s_j^y, s_j^z are the x, y, and z components of the source at the jth voxel. When the jth voxel has a source activity, it is generally true that s_j^x, s_j^y, s_j^z have nonzero values. However, when using the L_1 regularization, only one of s_j^x, s_j^y, s_j^z tends to have nonzero value, and others tend to be close to zero because of the nature of a sparse solution. As a result, the source orientation may be erroneously estimated.

To avoid this problem, the source orientation is estimated in advance using some other method [4] such as the L_2-norm minimum-norm method. Then, the L_1-norm method is formulated using the orientation-embedded data model in Eq. (2.15). That is, we use

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \mathcal{F} : \quad \mathcal{F} = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \xi \sum_{j=1}^{N} |x_j|. \tag{2.53}$$

In this case, the sparsity is imposed on the source vector magnitude, $s_1, s_2, \ldots, s_N$, and only a few of $s_1, s_2, \ldots, s_N$ have nonzero values, allowing for the reconstruction of a sparse source distribution.

2.10 Bayesian Derivation of the Minimum-Norm Method

2.10.1 Prior Probability Distribution and Cost Function

In this section, we derive the minimum-norm method based on Bayesian inference. As in Eq. (2.16), we assume that the noise $\boldsymbol{\varepsilon}$ is independently and identically distributed Gaussian, i.e.,

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \beta^{-1}\boldsymbol{I}), \tag{2.54}$$

where the precision β is used, which is the inverse of the noise variance, $\beta^{-1} = \sigma^2$. Thus, using Eq. (2.14), the conditional probability distribution of the sensor data for a given $\boldsymbol{x}$, $p(\boldsymbol{y}|\boldsymbol{x})$ is

$$p(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{\beta}{2\pi}\right)^{M/2} \exp\left[-\frac{\beta}{2}\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2\right]. \tag{2.55}$$

This conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ is equal to the likelihood $p(\boldsymbol{y})$ in the arguments in Sect. 2.5. Since $\boldsymbol{x}$ is a random variable in the Bayesian arguments, we use the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$, instead of $p(\boldsymbol{y})$.

Let us derive a cost function for estimating $\boldsymbol{x}$. Taking a logarithm of the Bayes's rule in Eq. (B.3) in the Appendix, we have

$$\log p(\boldsymbol{x}|\boldsymbol{y}) = \log p(\boldsymbol{y}|\boldsymbol{x}) + \log p(\boldsymbol{x}) + \mathcal{C}, \tag{2.56}$$

where $\mathcal{C}$ represents the constant terms. Neglecting $\mathcal{C}$, the cost function $\mathcal{F}(\boldsymbol{x})$ in general form is obtained as

$$\mathcal{F}(\boldsymbol{x}) = -2\log p(\boldsymbol{x}|\boldsymbol{y}) = \beta\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 - 2\log p(\boldsymbol{x}). \tag{2.57}$$

The first term on the right-hand side is a squared error term, which expresses how well the solution $\boldsymbol{x}$ fits the sensor data $\boldsymbol{y}$. The second term $-2\log p(\boldsymbol{x})$ is a constraint imposed on the solution. The above equation indicates that the constraint term in the cost function is given from the prior probability distribution in the Bayesian formulation. The optimum estimate of $\boldsymbol{x}$ is obtained by minimizing the cost function $\mathcal{F}(\boldsymbol{x})$.

2.10.2 L_2-Regularized Method

Let us assume the following Gaussian distribution for the prior probability distribution of $\boldsymbol{x}$,

$$p(\boldsymbol{x}) = \left(\frac{\alpha}{2\pi}\right)^{N/2} \exp\left[-\frac{\alpha}{2}\|\boldsymbol{x}\|^2\right]. \tag{2.58}$$

Substituting Eq. (2.58) into (2.57), we get the cost function

$$\mathcal{F}(\boldsymbol{x}) = \beta \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 + \alpha \|\boldsymbol{x}\|^2. \tag{2.59}$$

The cost function in Eq. (2.59) is the same as the cost function in Eq. (2.37), assuming $\lambda = \beta/\alpha$. Thus, the solution obtained by minimizing this cost function is equal to the solution of the L_2-norm regularized minimum-norm method introduced in Sect. 2.8.

To obtain the optimum estimate of $\boldsymbol{x}$, we should compute the posterior distribution. In this case, the posterior is known to have a Gaussian distribution because $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x})$ are both Gaussian, and the mean and the precision matrix of this posterior distribution is derived as in Eqs. (B.24) and (B.25). Substituting $\boldsymbol{\Phi} = \alpha \boldsymbol{I}$ and $\boldsymbol{\Lambda} = \beta \boldsymbol{I}$ into these equations, we have

$$\boldsymbol{\Gamma} = \alpha \boldsymbol{I} + \beta \boldsymbol{F}^T \boldsymbol{F}, \tag{2.60}$$

$$\bar{\boldsymbol{x}}(t) = \left(\boldsymbol{F}^T \boldsymbol{F} + \frac{\alpha}{\beta} \boldsymbol{I} \right)^{-1} \boldsymbol{F}^T \boldsymbol{y}(t). \tag{2.61}$$

The Bayesian solution which minimizes the cost function in Eq. (2.59) is given in Eq. (2.61). This solution is the same as Eq. (2.39). Comparison between Eqs. (2.61) and (2.39) shows that the regularization constant is equal to α/β, which is the inverse of the signal-to-noise ratio of the sensor data. This is in accordance with the arguments in Sect. 2.8 that when the sensor data contains larger amounts of noise, a larger regularization constant must be used.

The optimum values of the hyperparameters α and β can be obtained using the EM algorithm, as described in Sect. B.5.6. The update equations for the hyperparameters are:

$$\widehat{\alpha}^{-1} = \frac{1}{3N} \left[\frac{1}{K} \sum_{k=1}^{K} \bar{\boldsymbol{x}}^T(t_k) \bar{\boldsymbol{x}}(t_k) + \mathrm{tr}\left(\boldsymbol{\Gamma}^{-1} \right) \right], \tag{2.62}$$

$$\widehat{\beta}^{-1} = \frac{1}{M} \left[\frac{1}{K} \sum_{k=1}^{K} \|\boldsymbol{y}(t_k) - \boldsymbol{F}\bar{\boldsymbol{x}}(t_k)\|^2 + \mathrm{tr}\left(\boldsymbol{F}^T \boldsymbol{F} \boldsymbol{\Gamma}^{-1} \right) \right]. \tag{2.63}$$

Here, we assume that multiple K time-point data is available to determine α and β.

The Bayesian minimum-norm method is summarized as follows. First, $\boldsymbol{\Gamma}$ and $\bar{\boldsymbol{x}}(t_k)$ are computed using Eqs. (2.60) and (2.61) with initial values set to α and β. Then, the values of α and β are updated using (2.62) and (2.63). Using the updated α and β, the values of $\boldsymbol{\Gamma}$ and $\bar{\boldsymbol{x}}(t_k)$ are updated using Eqs. (2.60) and (2.61). These procedures are repeated and the resultant $\bar{\boldsymbol{x}}(t_k)$ is the optimum estimate of $\boldsymbol{x}(t_k)$.

The EM iteration may be stopped by monitoring the marginal likelihood, which is obtained using Eq. (B.29) as

$$\log p(\boldsymbol{y}(t_1), \ldots, \boldsymbol{y}(t_K)|\alpha, \beta) = -\frac{1}{2} K \log |\boldsymbol{\Sigma}_y| - \frac{1}{2} \sum_{k=1}^{K} \boldsymbol{y}^T(t_k) \boldsymbol{\Sigma}_y^{-1} \boldsymbol{y}(t_k), \tag{2.64}$$

where according to Eq. (B.30), $\boldsymbol{\Sigma}_y$ is expressed as

$$\boldsymbol{\Sigma}_y = \beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{F}\boldsymbol{F}^T. \tag{2.65}$$

If the increase of the likelihood in Eq. (2.64) with respect to the iteration count becomes very small, the iteration may be stopped.

2.10.3 L_1-Regularized Method

The method of L_1-norm regularization can also be derived based on the Bayesian formulation. To derive the L_1-regularization. we use the Laplace distribution as the prior distribution

$$p(\boldsymbol{x}) = \prod_{j=1}^{N} \frac{1}{2b} \exp\left[-\frac{1}{b}|x_j|\right]. \tag{2.66}$$

Then, using Eq. (2.57), (and replacing $\boldsymbol{F}$ with $\boldsymbol{H}$), the cost function is derived as

$$\mathcal{F}(\boldsymbol{x}) = \beta\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + 2b\sum_{j=1}^{N}|x_j|, \tag{2.67}$$

which is exactly equal to Eq. (2.53), if we set $\xi = 2b/\beta$.

Another formulation for deriving the L_1-regularized method is known. It uses the framework of the sparse Bayesian learning described in Chap. 4. In Chap. 4, assuming the Gaussian prior,

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \prod_{j=1}^{N}\mathcal{N}(x_j|0,\alpha_j^{-1}) = \prod_{j=1}^{N}\left(\frac{\alpha_j}{2\pi}\right)^{1/2}\exp\left[-\frac{\alpha_j}{2}x_j^2\right], \tag{2.68}$$

we derive the marginal likelihood for the hyperparameter $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]$, $p(\boldsymbol{y}|\boldsymbol{\alpha})$, using,

$$p(\boldsymbol{y}|\boldsymbol{\alpha}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{\alpha})d\boldsymbol{x}, \tag{2.69}$$

and eventually derive the Champagne algorithm. However, instead of implementing Eq. (2.69), there is another option in which we compute the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ using

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha} = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}), \tag{2.70}$$

where

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}. \tag{2.71}$$

The estimate $\widehat{\boldsymbol{x}}$ is, then, obtained by

$$\widehat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\text{argmax}}\, p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}).$$

To compute $p(\boldsymbol{x})$ using Eq. (2.71), we need to specify the hyperprior $p(\boldsymbol{\alpha})$. However, we usually have no such information and may use noninformed prior $p(\boldsymbol{\alpha}) = \text{const}$. Substituting this flat prior into Eq. (2.71), we have

$$p(\boldsymbol{x}) \propto \int p(\boldsymbol{x}|\boldsymbol{\alpha})d\boldsymbol{\alpha}.$$

However, the integral in the above equation is difficult to compute. The formal procedure to compute $p(\boldsymbol{x})$ in this case is to first assume the Gamma distribution for the hyperprior $p(\boldsymbol{\alpha})$, such that

$$p(\boldsymbol{\alpha}) = \prod_{j=1}^{N} p(\alpha_j) = \prod_{j=1}^{N} \Gamma(a)^{-1} b^a (\alpha_j)^{a-1} e^{-b\alpha_j}. \tag{2.72}$$

Then, $p(\boldsymbol{x})$ in Eq. (2.71) is known to be obtained as Student t-distribution, such that [8]

$$\begin{aligned} p(x_j) &= \int p(x_j|\alpha_j)p(\alpha_j)d\alpha_j \\ &= \int \left(\frac{\alpha_j}{2\pi}\right)^{1/2} \exp\left(-\frac{\alpha_j}{2}x_j^2\right) \frac{b^a}{\Gamma(a)} \left(\alpha_j\right)^{a-1} e^{-b\alpha_j} d\alpha_j \\ &= \frac{b^a \Gamma(a+\frac{1}{2})}{\sqrt{2\pi}\Gamma(a)} \left(b + \frac{x_j^2}{2}\right)^{-(a+\frac{1}{2})}. \end{aligned} \tag{2.73}$$

We then assume that $a \to 0$ and $b \to 0$, (which is equivalent to making $p(\boldsymbol{\alpha})$ a noninformed prior,) $p(x_j)$ then becomes

$$p(x_j) \to \frac{1}{|x_j|} \quad \text{i.e.} \quad p(\boldsymbol{x}) \to \prod_{j=1}^{N} \frac{1}{|x_j|}. \tag{2.74}$$

Using Eq. (2.57), the cost function, in this case, is derived as

$$\mathcal{F}(\boldsymbol{x}) = \beta\|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|^2 + \sum_{j=1}^{N} \log|x_j|. \tag{2.75}$$

This Eq. (2.75) is not exactly equal to the L_1-norm cost function, since the constraint term is not equal to $\sum_{j=1}^{N} |x_j|$ but has form of $\sum_{j=1}^{N} \log |x_j|$. Since these constraint terms have similar properties, the solution obtained by minimizing this cost function has a property very similar to the L_1-norm-regularized minimum-norm solution. Related arguments are found in Chap. 6.

References

1. M.S. Hämäläinen, R.J. Ilmoniemi, Interpreting measured magnetic fields of the brain: estimates of current distributions. Technical Report TKK-F-A559, Helsinki University of Technology (1984)
2. M.S. Hämäläinen, R.J. Ilmoniemi, Interpreting magnetic fields of the brain: minimum norm estimates. Med. Biol. Eng. Comput. **32**, 35–42 (1994)
3. J. Sarvas, Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Phys. Med. Biol. **32**, 11–22 (1987)
4. K. Uutela, M. Hämäläinen, E. Somersalo, Visualization of magnetoencephalographic data using minimum current estimate. NeuroImage **10**, 173–180 (1999)
5. B.D. Jeffs, Maximally sparse constrained optimization for signal processing applications. Ph.D. thesis, University of Southern California (1989)
6. K. Matsuura, Y. Okabe, Multiple current-dipole distribution reconstructed by modified selective minimum-norm method, in *Biomag 96*, (Springer, Heidelberg, 2000), pp. 290–293
7. R. Tibshirani, Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B (Methodological) **58**(1), 267–288 (1996)
8. M.E. Tipping, Sparse Bayesian learning and relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)

http://www.springer.com/978-3-319-14946-2

Electromagnetic Brain Imaging
A Bayesian Perspective
Sekihara, K.; Nagarajan, S.S.
2015, XIV, 270 p. 32 illus., 27 illus. in color., Hardcover
ISBN: 978-3-319-14946-2