

Journal of the American Statistical Association Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Estimating the Number of Sources in Magnetoencephalography Using Spiked Population Eigenvalues

Zhigang Yao, Ye Zhang, Zhidong Bai & William F. Eddy

To cite this article: Zhigang Yao, Ye Zhang, Zhidong Bai & William F. Eddy (2018) Estimating the Number of Sources in Magnetoencephalography Using Spiked Population Eigenvalues, Journal of the American Statistical Association, 113:522, 505-518, DOI: <u>10.1080/01621459.2017.1341411</u>

To link to this article: <u>https://doi.org/10.1080/01621459.2017.1341411</u>



View supplementary material \square

Accepted author version posted online: 10 Jul 2017. Published online: 08 Jun 2018.



Submit your article to this journal 🗹

Article views: 274



View Crossmark data

Estimating the Number of Sources in Magnetoencephalography Using Spiked Population Eigenvalues

Zhigang Yao^a, Ye Zhang ^{ob}, Zhidong Bai ^{oc}, and William F. Eddy^d

^aDepartment of Statistics and Applied Probability, National University of Singapore, Singapore; ^bSchool of Science and Technology, Örebro University, Örebro, Sweden; ^cKLASMOE and School of Mathematics and Statistics, Northeast Normal University, Changchun, China; ^dDepartment of Statistics, Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

Magnetoencephalography (MEG) is an advanced imaging technique used to measure the magnetic fields outside the human head produced by the electrical activity inside the brain. Various source localization methods in MEG require the knowledge of the underlying active sources, which are identified by a priori. Common methods used to estimate the number of sources include principal component analysis or information criterion methods, both of which make use of the eigenvalue distribution of the data, thus avoiding solving the time-consuming inverse problem. Unfortunately, all these methods are very sensitive to the signal-to-noise ratio (SNR), as examining the sample extreme eigenvalues does not necessarily reflect the perturbation of the population ones. To uncover the unknown sources from the very noisy MEG data, we introduce a framework, referred to as the intrinsic dimensionality (ID) of the optimal transformation for the SNR rescaling functional. It is defined as the number of the spiked population eigenvalues of the associated transformed data matrix. It is shown that the ID yields a more reasonable estimate for the number of sources than its sample counterparts, especially when the SNR is small. By means of examples, we illustrate that the new method is able to capture the number of signal sources in MEG that can escape PCA or other information criterion-based methods. Supplementary materials for this article are available online.

1. Introduction

Thousands of synchronized neurons give rise to macroscopic oscillations, which can be observed in the electroencephalogram (EEG). Meanwhile, the electric currents generated by those synchronized neurons induce extremely weak magnetic fields (10-100 femto-Tesla). Measuring magnetic fields of this magnitude is a great challenge. The recent development of superconducting quantum interference devices (SQUIDs) makes it possible to detect these extremely weak magnetic signals. Magnetoencephalography (MEG) is a noninvasive imaging technique that measures the magnetic fields produced by neuronal activity in the brain, with many coils placed around the head (Cohen 1968). The fundamental difference between MEG and other functional imaging modalities, such as positron emission tomography (PET) or functional magnetic resonance imaging (fMRI), is that the neuronal activity is measured directly by MEG, in the sense that the signal it measures is a direct consequence of brain activity, but indirectly by PET or fMRI, which measure the blood flow changes or other vascular phenomena. Because of its impressive temporal resolution (better than 1 ms, compared to 1 sec for fMRI, or up to 1 min for PET), MEG measures brain activity without any time delay, and this makes MEG a near optimal tool for studying the brain.

In MEG, the sources are mainly the electric currents generated by the activated neurons in the human cortex, and they

are usually formulated as mathematical point current dipoles. A brief description of the mathematical modeling and the corresponding forward and inverse problems in MEG can be found in the supplementary materials. The main challenge posed by MEG is to determine the location of electrical activity within the brain from the induced magnetic fields around the skull. This is a typical ill-posed problem since (i) one can always construct a large number of nontrivial dipoles that have a vanishing magnetic field outside the head. (ii) the identification process is usually unstable, that is, a small amount of noise in the measurement data can lead to enormous errors in the estimates. To tackle the ill-posedness of the inverse MEG problem, a sequence of regularization algorithms have been extensively exploited during the past two decades: see Hamalainen and Ilmoniemi (1994), Uutela, Hamalainen, and Somersalo (1999), Vanni and Uutela (2000), Pulvermuller, Shtvrov, and Ilmoniemi (2003), Mattout et al. (2006), and references therein. However, almost all of them are highly restricted by the predefined number of dipoles or the prior distribution assumed in the data (Schmidt, George, and Wood 1999). A fundamental challenge here is that it is almost impossible to pick up how many active dipoles will be needed in advance. Other approaches, also known as "imaging methods," represented by the multiple signal classification (MUSIC; see, e.g., Mosher and Leahy 1998), attempt to estimate the number of dipoles (less than 10) by separating the signal

CONTACT Zhigang Yao Zhigang.yao@nus.edu.sg Department of Statistics and Applied Probability, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 117546.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*. (B) Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

© 2018 American Statistical Association

ARTICLE HISTORY

Received May 2016 Revised April 2017

KEYWORDS

Brain imaging; Eigenthresholding; Intrinsic dimensionality; Inverse MEG problem; Spiked eigenvalues



Check for updates

subspace and the noise subspace, that is, they decompose the eigenvalues of the sample space of the MEG data into the eigenvalues of the signal subspace and noise subspace, following the assumption that the lead field vector at each source location is orthogonal to the noise. By assuming all dipoles are uncorrelated, the beamformer-inspired approaches (Van Veen, Joseph, and Hecox 1992) solve a more "tangent" problem, that is, avoiding the need to estimate the *prior*. Beamformers are essentially spatial filters that suppress the dipoles in a number of selected locations, blocking out the signal originated elsewhere. Most recently, several works based on time-varying dipoles (Yao and Eddy 2014) have been proposed, where it is suggested that the number of varying dipoles is estimated in a dynamic fashion.

Though the underlying number of dipoles is generally unknown and thus impossible to verify, there has been increasing demand for a reliable estimation. Due to the nature of the inverse problem, estimating the number of dipoles in MEG can pose unconventional challenges. The same type of problem can be traced back to the problem of detecting the number of signal sources in the presence of noise, or its *intrinsic dimensionality*, in a multiple channels of time series (Ligget 1973; Schmidt 1986). The heuristics are as follows: if the noise is Gaussian, then the number of the dipoles is related to the multiplicity of the smallest eigenvalues of the sample covariance matrix, and in this case the key to estimating the number of dipoles is to find a threshold that separates the *spiked* eigenvalues, the extreme eigenvalues, from the bulk eigenvalues of sample covariance matrix, or, more generally, to perform a sequence of hypothesis tests on the same eigenvalues against the threshold (Bartlett 1954; Lawley 1956). Centering around making use of the distribution of eigenvalues of the sample covariance matrix, there are two schools of thoughts in the literature: (1) pursuing the number of signal sources from principal component analysis (PCA), independent component analysis (ICA) (Green et al. 2001) and factor analysis (Malinowski 1977a, b); (2) estimating the solution based on certain information criterion, such as Akaike information criterion (AIC) and minimum description length (MDL). This being said, separating the spiked eigenvalues from the sample covariance does not necessarily give an accurate estimation for the number of sources. The reasons are two-fold: (1) the sample eigenvalues are not consistent estimators of their population counterparts, and hence any test based on sample eigenvalues inevitably reveals considerable deviation from the truth, and (2) the unknown noise structure, which, although it can be estimated, amplifies the difficulty of estimating the transition function (to be defined in Section 2) between the sample eigenvalues and the population eigenvalues. The first one is well illustrated by considering an extreme case (see Bai 1999 for details): assuming the population covariance $\sigma^2 I_p$ with finite fourth moment and $p/n \rightarrow \gamma \in (0, 1)$ (*n* is the size of samples), then the empirical spectral distribution of the sample covariance converges to the Marčenko–Pastur law $F_{\gamma}(dx)$

$$F_{\gamma}(dx) = \frac{1}{2\pi x \gamma \sigma^2} \sqrt{(b_{\gamma} - x)(x - a_{\gamma})} dx, \ a_{\gamma} \le x \le b_{\gamma}$$

where $a_{\gamma} = \sigma^2 (1 - \sqrt{\gamma})^2$, $b_{\gamma} = \sigma^2 (1 + \sqrt{\gamma})^2$. This means, with high probability, the *k*-largest eigenvalues (or *k*-smallest eigenvalues) of the sample covariance matrix converge to the

spectrum bounds a_{γ} (or b_{γ}), rather than the population eigenvalue σ^2 in the null case. The second reason outlined above implies that the inverse of transition function does not generally work, unless there are only a few distinct spiked eigenvalues, which is usually the case in real applications (in MEG, we expect only a few spiked eigenvalues). A crucial problem, particularly in MEG, is that the magnitude of the unknown noise aggregates the estimation error of the population spikes under the non-Gaussian model, which fails most of existing methods, such as PCA. This calls for a reexamination of the effect caused on the sample extreme eigenvalues by perturbation on the population covariance matrix, the essence of which involves developing new estimates for the population eigenvalues and a justification for its consistency in Gaussian and non-Gaussian spiked model.

The goal of this article is to bridge the estimation of the number of sources and the behavior of the spiked eigenvalues, on the population level. The remainder of the article is structured as follows: in Section 2, we demonstrate how the problem can be transformed into one of estimating the extreme eigenvalues of the population matrix, instead of directly investigating the limiting behavior of sample spikes. This differentiates our method from other thresholding methods. The relation of the spiked sample eigenvalues and the spiked population eigenvalues is then presented. A new estimator for the population spiked eigenvalues is given, with an algorithm estimating the intrinsic dimensionality under an optimal signal-to-noise ratio transformation. In Section 3, by means of simulated examples and the *empty-room* data, we illustrate that our method is able to capture the number of sources under various SNRs, with hundreds of channels or more. We also compare our approach with other methods. Finally, a real MEG dataset is tested in Section 4, and a short discussion and concluding remarks are given in Section 5.

2. Methodology

In a typical MEG experiment, the magnetic field **B** is sampled on a finite number (say *K*) of sensors, with each one measuring one component (radial direction) of the magnetic field, namely, **B**_z; if $\mathbf{e} = (0, 0, 1)$, a unit vector, is used to find **B**_z, the *z* component of **B** can be obtained by $\mathbf{B}_z = \mathbf{B} \cdot \mathbf{e}$. Nevertheless, for simplicity, we will ignore the subscript *z* in **B**_z from now on. Therefore, the general framework of the MEG model becomes

$$\mathbf{Y}(t) = \mathbf{B}(t) + \mathbf{e}(t) = \mathbf{G}\mathbf{Q}(t) + \mathbf{e}(t), \quad (1)$$

where $\mathbf{Y}(t) = [Y_1(t), Y_2(t), \dots, Y_K(t)]^\top$ is the $K \times 1$ vector representing the observed magnetic field by K sensors at time t. The design matrix $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_N]$ is of size $K \times$ 3N matrix with each submatrix \mathbf{G}_i of $K \times 3$ being the corresponding magnetic field observed across sensors generated by a unit dipole at a given location $\mathbf{r}_i(1 \le i \le N)$. The $\mathbf{Q}(t) =$ $[\mathbf{Q}_1^\top(t), \mathbf{Q}_2^\top(t), \dots, \mathbf{Q}_N^\top(t)]^\top$ is the time course vector representing the strength and moments of all N dipoles, with each submatrix $\mathbf{Q}_i(t)$ of 3×1 being the time course of ith dipole at time t. Considering the entire time course of $\mathbf{Q}(t)$, we denote the covariance matrix of $\mathbf{B}(t)$ (or $\mathbf{G}\mathbf{Q}(t)$) as \mathbf{R}_s . The vector $\mathbf{e}(t) = [e_1(t), e_2(t), \dots, e_K(t)]^\top$ accounts for the presence of additive noise in the MEG data.

2.1. Optimal Signal-to-Noise Ratio (SNR) and Intrinsic Dimensionality (ID)

Before discussing the method of determining the number of sources in MEG, introduce a so-called SNR rescaling functional $\mathcal{I}_{\mathbf{R},\mathbf{R}_{u}}$, such that for any matrix $\mathbf{X} \in \mathbb{R}^{K \times K}$

$$\mathcal{I}_{\mathbf{R},\mathbf{R}_n}(\mathbf{X}) = \frac{\|\mathbf{X}^\top \mathbf{R} \mathbf{X}\|_2}{\|\mathbf{X}^\top \mathbf{R}_n \mathbf{X}\|_2},\tag{2}$$

where **R** and **R**_n are the covariance matrix of the observed data $\mathbf{Y}(t)$ and noise $\mathbf{e}(t)$, respectively, and $\|\mathbf{X}\|_2 = \sqrt{\lambda_{\max}(\mathbf{X}^{\top}\mathbf{X})}$ is a standard operation norm for a matrix $\mathbf{X} \in \mathbb{R}^{K \times K}$. Such a norm is invariant with respect to any unitary matrix transformation, that is, for any unitary matrix U_1 , U_2 the equality $\|U_1\mathbf{X}U_2\|_2 = \|\mathbf{X}\|_2$ holds. Therefore, the transformation can be employed for the signal and noise simultaneously. As a ratio between the noise and signal, a high value of $\mathcal{I}_{\mathbf{R},\mathbf{R}_n}$ means a small effect caused by noise.

The following theorem finds an optimal transformation for which the SNR rescaling functional $\mathcal{I}_{\mathbf{R},\mathbf{R}_n}$ attains its maximum. The proof can be found in the Appendix.

Theorem 1. Suppose that matrix **R** is nonsingular. Let Φ_n and Λ_n be eigenvectors and eigenvalues of noise covariance \mathbf{R}_n . Let $\lambda_{\max} > 0$ be the maximal eigenvalue of the matrix \mathbf{R}_{adj} , where

$$\mathbf{R}_{\mathrm{adj}} = \mathbf{W}_n^\top \mathbf{R} \mathbf{W}_n, \qquad \mathbf{W}_n = \mathbf{\Phi}_n \mathbf{\Lambda}_n^{-1/2}. \tag{3}$$

Then

$$\max_{\mathbf{X}\in\mathbb{R}^{K\times K}}\mathcal{I}_{\mathbf{R},\mathbf{R}_n}(\mathbf{X})=\lambda_{\max}$$

and

$$\mathbf{X}_{1}^{\text{opt}} = \mathbf{W}_{n},\tag{4}$$

$$\mathbf{X}_{2}^{\text{opt}} = \mathbf{\Phi}_{n} \mathbf{\Lambda}_{n}^{-1/2} \mathbf{\Phi}_{\text{adj}}$$
(5)

are two maximums of functional $\mathcal{I}_{\mathbf{R},\mathbf{R}_n}(\cdot)$, that is, $\mathcal{I}_{\mathbf{R},\mathbf{R}_n}(\mathbf{X}_i^{\text{opt}}) = \lambda_{\max}$ for i = 1, 2. Here, the columns of matrix $\mathbf{\Phi}_{\text{adj}}$ are the eigenvectors of \mathbf{R}_{adj} .

If we assume the signal and noise are uncorrelated as in (1), then the covariance matrix \mathbf{R} simply breaks down to

$$\mathbf{R}=\mathbf{R}_{s}+\mathbf{R}_{n},$$

where \mathbf{R}_s is the signal covariance matrix. Suppose we know the noise covariance matrix \mathbf{R}_n , then a whitening process can be applied to transform \mathbf{R} and \mathbf{R}_n

$$\mathbf{W}_{n}^{\top}\mathbf{R}\mathbf{W}_{n} = \mathbf{W}_{n}^{\top}\mathbf{R}_{s}\mathbf{W}_{n} + \mathbf{I},$$
(6)

where \mathbf{W}_n is defined in (3).

Theorem 1 implies that, with $\mathbf{X} = \mathbf{X}_2^{\text{opt}}$ defined in (5), we can rewrite

$$\mathbf{X}^{\top}\mathbf{R}\mathbf{X} = \mathbf{\Lambda}_{\mathrm{adj}} = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_M, \lambda_{M+1}, \dots, \lambda_K)$$

by noting that $\mathbf{R}_{adj} = \mathbf{W}_n^{\top} \mathbf{R} \mathbf{W}_n$ and the fact that

$$\mathbf{\Phi}_{\mathrm{adj}}^{\top}\mathbf{R}_{\mathrm{adj}}\mathbf{\Phi}_{\mathrm{adj}} = \mathbf{\Lambda}_{\mathrm{adj}}.$$

By Equation (6), it is not difficult to show that $\lambda_k \ge 1, k = 1, \ldots, K$. Denote by *M* the number of nonzero eigenvalues of \mathbf{R}_s . Then, we can split $\{\lambda_k\}_{k=1}^K$ into two groups: let $\{\lambda_k\}_{k=1}^M$ be

the first group (we call it the spiked group), where all eigenvalues are strictly larger than the unit; the second group, which is called the bulk group, contains only the unit eigenvalues, that is, $\{\lambda_k \equiv 1\}_{k=M+1}^K$. Obviously, the spiked group can be regarded as the contribution of the associated eigenvalues of both \mathbf{R}_s and \mathbf{R}_n , whereas the ones in the bulk group originate from the eigenvalues of \mathbf{R}_n after the whitening process. Therefore, the intrinsic dimensionality (ID) of the data can be determined by counting the number of distinct eigenvalues of \mathbf{R}_{adi} in the spiked group. In the next subsection, we confirm that the ID is an invariant, therefore, it can be used to estimate the number of sources in the MEG problem. However, with limited knowledge of \mathbf{R}_n , one would mainly rely on an estimate of \mathbf{R}_n , say $\widehat{\mathbf{R}}_n$. The challenge is that the distribution of the eigenvalues of \mathbf{R}_{adj} are no longer tractable, in the sense that neither does the thresholding apply nor are the sample eigenvalues of $\widehat{\mathbf{R}}_{adj}$ good estimates of their population counterparts. Rather than using the spiked sample eigenvalues, we propose a new method to estimate the spiked population eigenvalues of \mathbf{R}_{adj} , based on which the inference of the dimensionality is made (see Theorem 3 in Section 2.2).

Definition 1. $\mathbf{X}_{i}^{\text{opt}}$ (i = 1, 2) defined in (4) and (5), are called optimal transformations ($\mathbf{X}_{1}^{\text{opt}}$ and $\mathbf{X}_{2}^{\text{opt}}$ are essentially equivalent; we will be only using $\mathbf{X}_{1}^{\text{opt}}$) for the SNR-functional (2). Moreover, $\mathbf{R}_{\text{adj}} = (\mathbf{X}_{1}^{\text{opt}})^{\top} \mathbf{R} \mathbf{X}_{1}^{\text{opt}} \equiv \mathbf{W}_{n}^{\top} \mathbf{R} \mathbf{W}_{n}$ is called an optimal associated covariance matrix (with respect to the optimal transformation $\mathbf{X}_{1}^{\text{opt}}$). Accordingly, $\widehat{\mathbf{R}}_{\text{adj}} = \widehat{\mathbf{W}}_{n}^{\top} \widehat{\mathbf{R}} \widehat{\mathbf{W}}_{n}$, with an estimated $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{W}}_{n}$ (named as a quasi-optimal transformation), is called a quasi-optimal associated covariance matrix.

An the end of this subsection, we indicate that it is reasonable to use the quasi-optimal associated covariance matrix $\widehat{\mathbf{R}}_{adj}$ instead of the optimal one \mathbf{R}_{adj} , if a good noise estimation has been captured. Theorem 2 (its proof can be found in the Appendix) also implies that the transformation used in Theorem 1 is a stable operator.

Theorem 2. Let $\widehat{\mathbf{R}}_n$ be an ω -estimator of \mathbf{R}_n such that

$$\|\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}_n\| \le \omega, \quad \|\widehat{\mathbf{\Lambda}}_n^{-1/2} - \mathbf{\Lambda}_n^{-1/2}\| \le \omega, \tag{7}$$

where $\{\widehat{\Phi}_n, \widehat{\Lambda}_n\}$ and $\{\Phi_n, \Lambda_n\}$ are eigenvectors and eigenvalues of $\widehat{\mathbf{R}}_n$ and \mathbf{R}_n , respectively. $\|\cdot\|$ denotes any type of norm of a matrix. Moreover, assume that there exists a constant *C* such that $\|\mathbf{\Lambda}_n^{-1/2}\| \leq C$. Then

$$\|\widehat{\mathbf{R}}_{\mathrm{adj}} - \mathbf{R}_{\mathrm{adj}}\| = O(\omega).$$

2.2. Determining the Intrinsic Dimensionality (ID)

In this section, we present the details to determine the ID from the quasi-optimal associated covariance matrix $\widehat{\mathbf{R}}_{adj}$. We indicate that, from this section, instead of using the original data $\mathbf{Y}(t)$, we will only deal with the data after the quasi-optimal transformation, introduced in the previous section, that is, the new data structure $\mathbf{Z}(t) = \widehat{\mathbf{W}}_{n}^{\top} \mathbf{Y}(t)$.

2.2.1. Definition and Assumption

Definition 2. Let \mathbf{V}_K be a population covariance matrix of dimension K and assume that the empirical spectral distribution (ESD) H_T of \mathbf{V}_K tends to a proper probability distribution H, called the limiting spectral distribution (LSD), as the dimension $K \to \infty$. An eigenvalue λ of \mathbf{V}_K is called a spiked eigenvalue if $\lambda \notin \Gamma_H$, where Γ_H denotes the support of H. Otherwise, it is called a bulk eigenvalue.

Remark: To avoid possible confusion when the eigenvalues vary with *K*, this definition can be modified as $d(\lambda(\mathbf{V}_K), \Gamma_H) > \delta_0$ for *spiked eigenvalues* and otherwise for *bulk eigenvalues*, where δ_0 is a prechosen positive constant and $d(\cdot)$ is a distance function, that is, $d(\lambda(\mathbf{V}_K), \Gamma_H) = \inf_{x \in \lambda(\mathbf{V}_K), y \in \Gamma_H} |x - y|$.

We generate the spiked covariance model in Bai and Ding (2012). Define $Z(1), Z(2), \ldots, Z(T)$ iid sample vectors drawn from the *K*-dimensional population with mean vector μ and covariance matrix

$$\mathbf{V}_{K} = \mathcal{O}^{\top} \operatorname{diag}(\lambda_{1} I_{m_{1}}, \lambda_{2} I_{m_{2}}, \dots, \lambda_{L} I_{m_{L}}, I_{K-M}) \mathcal{O}, \qquad (8)$$

with unit bulk eigenvalues and *L* distinct spiked eigenvalues $\{\lambda_l\}_{l=1}^L$, with respective multiplicities m_1, \ldots, m_L , satisfying $\lambda_1 > \lambda_2 > \cdots > \lambda_L$. Let *M* be a fixed constant. Denote $\sum_{l=1}^L m_l = M < K$, and assume finite fourth moment $E \|\mathbf{Z}(t)\|^4 < \infty$ ($1 \le t \le T$). Here I_{m_l} is an identity matrix of size m_l . Without losing generality, we assume $\boldsymbol{\mu} = 0$.

Definition 3. The number L will be called the intrinsic dimensionality.

The intrinsic dimensionality *L* will be determined by Algorithm 1. Here, throughout, we assume that

- (A1) There exists an orthogonal K-dimensional matrix \mathcal{O} such that $\mathbf{Z}(t) = \mathcal{OZ}(t), t = 1, \ldots, T$ are the sample vectors and sequence $\mathcal{Z}(t)$ can be split as $\mathcal{Z}(t) = (\mathcal{Z}_{M}^{\top}(t), \mathcal{Z}_{K-M}^{\top}(t))^{\top}$ according to their dimensions M and K M. Moreover, $\mathcal{Z}_{M}(t)$ and $\mathcal{Z}_{K-M}(t)$ are independent.
- (A2) *K* and *T* are related so that $K/T \rightarrow \gamma \in (0, 1)$ as $T \rightarrow \infty$.
- (A3) $\lambda_M \lambda_{M+1} > \sqrt{K/T}$, that is, the gap between the spiked and bulk eigenvalues is larger than a critical value $\sqrt{K/T}$.
- (A4) There exists a positive number ϵ_0 such that $\min_{i=1}^{L-1} (\lambda_i \lambda_{i+1}) \ge \epsilon_0$.

Assumption (A1) guarantees the extension of the spiked covariance model in Bai and Ding (2012) to the model with a nonblock structure (i.e., a dense V_K). This is the well-known "source condition" in regularization theory. Assumption (A3) is needed, since the BBP phase transition in the spiked population model exists, which says that only when the population spike is larger than a critical value, will its corresponding sample counterpart have a different asymptotic behavior from the null case (see Baik, Ben Arous, and Péché (2005) for details). In practice, the value of *M* will be estimated by Algorithm 3. Assumption (A4) sets the lower bound of the minimum eigen-gap of the spiked population eigenvalues.

2.2.2. Estimating Spiked Eigenvalues

To estimate the spiked eigenvalue of the population matrix, define

$$\underline{s}_{T,k} = \frac{\gamma_T - 1}{\hat{\lambda}_{T,k}} + \frac{1}{T} \sum_{j \notin J_l} \frac{1}{\hat{\lambda}_{T,j} - \hat{\lambda}_{T,k}}, \quad k \in J_l,$$
(9)

where $\gamma_T = K/T$, $\hat{\lambda}_{T,k}$ (k = 1, ..., K) is the *k*th sample spiked eigenvalues of the quasi-optimal associated covariance matrix $\widehat{\mathbf{R}}_{adj} = \widehat{\mathbf{W}}_n^\top \widehat{\mathbf{R}} \widehat{\mathbf{W}}_n$, and $J_l = \{s_l + 1, ..., s_l + m_l\}$ is denoted as the index set of the *l*th population spiked eigenvalue λ_l , where $s_l = \sum_{i=1}^{l-1} |J_i|$.

As we can see in (9), the value of $\underline{s}_{T,k}$ consists of two parts. The first part is the contribution of the *k*th eigenvalue of the sample matrix $\widehat{\mathbf{R}}_{adj}$ with the factor $\gamma_T - 1$, and the second part is the contribution of the remaining eigenvalues with the factor of the sample size *T*.

The following theorem provides an estimator of the eigenvalues of a population covariance matrix from the sample covariance matrix, which helps us to estimate the ID.

Theorem 3 (Theorem 3.1, Bai and Ding 2012). If λ_l is a distant population spike with multiplicity m_l , then, under the existence of the fourth moment of underlying distributions,

$$-1/\underline{s}_{T,k} \to \lambda_l$$
 (10)

as $T \to \infty$ almost surely, for any $k \in J_l$.

Note that the limiting property of the estimators $\underline{s}_{T,k}$ relies on the true set J_l . However, the exact set of J_l is not accessible in most of the cases. We will change the condition of summation in (9) by replacing $\{j \notin J_l\}$ with the set $\{j : |\hat{\lambda}_{T,j} - \hat{\lambda}_{T,k}| > \varepsilon'\}$. Moreover, theoretically, for any $k \in J_l$, $-1/\underline{s}_{T,k}$ can be considered as an estimator of the spiked population eigenvalue λ_l . In this article, using the "averaging" technique, instead of choosing one of $\{\underline{s}_{T,k}\}$ we use the following "averaged" one

$$\underline{s}_{T}^{(l)} = \frac{\gamma_{T} - 1}{\hat{\lambda}_{T}^{(l)}} + \frac{1}{T} \cdot \sum_{\{j: |\hat{\lambda}_{T,j} - \hat{\lambda}_{T}^{(l)}| > \varepsilon'\}} \frac{1}{\hat{\lambda}_{T,j} - \hat{\lambda}_{T}^{(l)}}$$
(11)

with

$$J_l = \{j : |\hat{\lambda}_{T,j} - \hat{\lambda}_{T,k_l}| \le \epsilon, \ k_l < j \le K\}, \ J_0 = \emptyset, \quad (12)$$

$$k_{l} = 1 + \sum_{i=1}^{l-1} |J_{i}|, \qquad \hat{\lambda}_{T}^{(l)} = \frac{1}{|J_{l}|} \sum_{k \in J_{l}} \hat{\lambda}_{T,k},$$
(13)

where ε' in (11) and ϵ in (12) are two thresholds. Note that ε' can be any arbitrary small number decided by users. This is to remove some eigenvalues closest to $\hat{\lambda}_{T,k}$ in (9) (or $\hat{\lambda}_{T}^{(l)}$ in (13)). This modification will not affect the consistency of the estimator, as stated in Bai and Ding (2012). We set $\varepsilon' \equiv 1\% \cdot \hat{\lambda}_{T,1}$ throughout this article. The threshold ϵ , however, is a tuning parameter. It controls the radius of the sample eigenvalues necessary for inclusion to estimate the corresponding spiked population eigenvalue.

It is not difficult to show that $-1/\underline{s}_T^{(l)} \to \lambda_l$ as $T \to \infty$, and it is a better estimator in comparison with $\underline{s}_{T,k}$ (at least from the numerical point of view, see a demonstration at the end of this subsection).

We first present a high-level description of the algorithm— Algorithm 1. The element of noise estimation in Step 2 will be briefly discussed in the next section. Readers may refer to the supplementary materials for more details. Apart from the spiked population eigenvalues, the bulk population eigenvalues will be estimated simultaneously with tuning parameter δ . The thresholds of δ and ϵ in Step 4 are to be determined by a self-exploited procedure. This will be explained through Algorithms 2 and 3. A complete integration of Algorithms 1–3 can be found in the supplementary materials.

Algorithm 1 A high-level algorithm determining the ID

- 1. Obtain the quasi-optimal associated covariance matrix \mathbf{R}_{adj} from data covariance \mathbf{R} . If the data have a negligible noise, set $\widehat{\mathbf{R}}_{adj} = \widehat{\mathbf{R}}$ and go to Step 3; otherwise, go to the next step.
- 2. Given any estimated $\widehat{\mathbf{R}}_n$, calculate the quasi-optimal transformation \mathbf{W}_n and obtain the corresponding quasioptimal associated covariance matrix $\widehat{\mathbf{R}}_{adj}$.
- 3. Calculate all eigenvalues of $\widehat{\mathbf{R}}_{adj}$: $\{\widehat{\lambda}_{T,k}\}_{k=1}^{K}$ and sort them in descending order.
- 4. Choose appropriate thresholds δ and ϵ .
- 5. Estimate the spiked population eigenvalues $\{\lambda_i\}_{i=1}^L$ using (10)-(13) in Theorem 3 from the sample spiked eigenvalues { $\hat{\lambda}_{T,k}$: $\hat{\lambda}_{T,k} > \delta, k = 1, \ldots, K$ }.

Output: $\{\tilde{\lambda}_i\}_{i=1}^L$ are the estimated population spiked eigenvalues and the ID equals L.

2.2.3. Determining Optimal Thresholds δ and ϵ

Theoretically, the thresholds δ and ϵ can be learned by the following optimization problem:

$$\{\delta, \epsilon\} = \underset{\delta, \epsilon}{\operatorname{argmin}} \left\| \frac{1}{V} \sum_{i=1}^{V} \mathbf{Z}_{i}(\delta, \epsilon) - \bar{\mathbf{Z}} \right\|_{F}^{2}, \quad (14)$$

where $\mathbf{Z} = [\mathbf{Z}(1), \dots, \mathbf{Z}(T)]$ denotes the transformed MEG data, $\|\cdot\|_F$ means the Frobenius norm of a matrix, and $\mathbf{Z}_i(\delta, \epsilon)$ is a random sample distributed with the mean vector $\mathbf{Z} =$ $\frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}(t)$ and covariance matrix $\hat{\mathbf{V}}_{K}$ by the estimated spiked population eigenvalues $\tilde{\lambda}_l$'s from Algorithm 1. Alternatively, one may use $\widehat{\mathbf{V}}'_{K} = \widehat{\mathcal{O}}^{\top} \widehat{\mathbf{V}}_{K} \widehat{\mathcal{O}}$, where $\widehat{\mathcal{O}}$ comes from the spectral decomposition of the sample covariance matrix $\widehat{\mathbf{R}} = \widehat{\mathcal{O}}^{\top} \mathbf{\Lambda} \widehat{\mathcal{O}}$. It is enough to use V_K for optimizing ϵ in (14), since V_K contains nearly all information about the eigenvalues. Note that the estimated bulk population eigenvalues can be set as the average of the sample bulk eigenvalues, that is,

$$\tilde{\lambda}_0 = \frac{1}{|\Gamma|} \sum_{\hat{\lambda}_{T,k} \in \Gamma} \hat{\lambda}_{T,k},\tag{15}$$

where set $\Gamma := \{\hat{\lambda}_{T,k} : \hat{\lambda}_{T,k} \le \delta, k = 1, \dots, K\}$ contains all bulk sample eigenvalues of $\widehat{\mathbf{R}}_{adj}$ and $|\Gamma|$ denotes as the cardinality of set Γ.

A detailed implementation of learning process (14) is presented in Algorithms 2 and 3, where Algorithm 2 presents an iterative method of selecting the optimal threshold ϵ , and Algorithm 3 shows an efficient approach for choosing an appropriate threshold δ .

Algorithm 2 Determining the optimal threshold ϵ

Input: Set the iteration index j = 1 and provide an appropriate threshold δ and an initial threshold $\epsilon = \epsilon_0$.

- 1. Estimated the spiked population eigenvalues $\{\tilde{\lambda}_i\}_{i=1}^L$ from the spiked sample eigenvalues { $\hat{\lambda}_{T,k}$: $\hat{\lambda}_{T,k} > \delta$, k =1, ..., K} with threshold ϵ .
- 2. Generate V random samples $\{\mathbf{Z}_i\}_{i=1}^V$ distributed with the mean vector Z and covariant spiked population eigenvalues $\{\tilde{\lambda}_i\}_{i=1}^L$. the relative discrepancy mean vector \mathbf{Z} and covariance matrix $\widehat{\mathbf{V}}_K$ by the estimated
- 3. Compute $\Delta_i =$ $\|\frac{1}{V}\sum_{i=1}^{V}\mathbf{Z}_{i}(\delta,\epsilon)-\bar{\mathbf{Z}}\|_{F}/\|\bar{\mathbf{Z}}\|_{F}.$
- 4. If $\epsilon > \hat{\lambda}_{T,1} \hat{\lambda}_{T,K}$, output the result; otherwise, update $\epsilon = \epsilon + \epsilon_0$, j = j + 1 and go back to Step 1.

Output: The optimal threshold $\tilde{\epsilon}$ is the one with the minimum relative discrepancy Δ_i .

Remark: (a) In Algorithm 2, ϵ_0 can be any appropriate value. Obviously, the smaller ϵ_0 , the better estimated ϵ . However, small ϵ_0 leads to more iterations in the algorithm. If there are no other good choices, we suggest using $\epsilon_0 = \hat{\lambda}_{T,K}$, the smallest sample eigenvalue. (b) By (12), ϵ is the radius of set J_l , which determines the estimation of population spiked eigenvalue $\hat{\lambda}_{T,l}$. The stopping criteria in Step 4 describes the maximal possible value of the radius of J_l . $\epsilon = \hat{\lambda}_{T,1} - \hat{\lambda}_{T,K}$ means the extreme case, when there is only one spiked eigenvalue and one bulk eigenvalue. All sample eigenvalues except the smallest one are used to estimate the population spiked eigenvalue. If the data covariance matrix has several distinguishable spiked eigenvalues, we may use the alternative stopping criteria " $\epsilon > p \cdot \hat{\lambda}_{T,1}$," where p should be chosen from case to case. Usually, p should be greater than 30%. Obviously, the smaller *p* is, the fewer iterations it requires. Note that we have used " $\epsilon > 40\% \cdot \hat{\lambda}_{T,1}$." The two criteria provide the same result in our simulation and real data application.

Algorithm 3 Choosing an appropriate threshold δ

- 1. Set index i = K and $\delta = \hat{\lambda}_{T,K}$.
- 2. If condition $\hat{\lambda}_{T,i-1} \hat{\lambda}_{T,i} < \epsilon_0/2$ holds, go to the next step; otherwise, output the result.
- 3. Set i = i 1 and $\delta = \hat{\lambda}_{T,i}$. If i < 2, output the error information "The spiked eigenvalues model cannot be employed;" otherwise, go back to Step 2.

Output: $\tilde{\delta} = \delta$ is an appropriate threshold.

Remark: In Algorithm 3, the initial value of δ is suggested to be $\lambda_{T,K}$. The algorithm separates the bulk and spiked eigenvalues and therefore it is an approach to estimate the number of spiked eigenvalues M.

2.2.4. A Demonstration of Algorithms 1, 2, and 3

We first generate a random (Gaussian) sample matrix $(300 \times$ 6000, that is, K = 300 and T = 6000) with a mean zero and a covariance matrix

$$\mathbf{V}_{300} = \text{diag}(\lambda_1 I_{20}, \lambda_2 I_{10}, \lambda_3 I_{40}, \lambda_4 I_{30}, I_{200}), \qquad (16)$$



Figure 1. A demonstration of Algorithms 1–3 with different distributions in Step 2 of Algorithm 2. " \circ " represents the sample eigenvalues and " \Box " represents the true spiked population eigenvalues. " \times ", "+," and "*"—the estimated spiked population eigenvalues with normal distribution, uniform distribution, and *t*-distribution, respectively.

where the $\lambda_1 = 20$, $\lambda_2 = 17$, $\lambda_3 = 10$, $\lambda_4 = 7$ are the four true spiked population eigenvalues. Then, the eigenvalues of the sample covariance matrix $\hat{\mathbf{R}}$ (in "o") are displayed in Figure 1. In the same figure, the exact spiked population eigenvalues (in "□") and the estimated spiked population eigenvalues (in "×," "+," and "*") by different sampling distributions in Step 2 of Algorithm 2 are also highlighted. The learning process of selecting the optimal threshold ϵ with the initial guess $\epsilon_0 = 10\% \cdot \hat{\lambda}_{T,1}$ is displayed in Figure 2 (the Gaussian distribution is used in Step 2 of Algorithm 2). The result of the learning process is the histogram (c) with the smallest relative discrepancy, that is, the optimal threshold $\tilde{\epsilon} = 3 \cdot \epsilon_0 = 30\% \cdot \hat{\lambda}_{T,1}$. The result also suggests that the Algorithm 1 recovers the true population spiked eigenvalues well in terms of number and magnitude; specifically, the optimal estimated spiked population eigenvalues are (20.42, 16.85, 10.53, 7.9), which implies that the number of spiked eigenvalues equals four.

However, there could be certain variation in the estimated population spiked eigenvalues with different sampling distributions used in tuning ϵ in Algorithm 2. We observe that in Table 1, with different initial guesses ϵ_0 in the first step and distributions in Step 2 of Algorithm 2, we have obtained only slightly different spiked population eigenvalues. However, the number of estimated spiked population eigenvalues is very stable with respect to these parameters. The supplementary materials contain more numerical experiments, including several replicates of the same setup, and a case of eight dipoles. In all experiments, the number of estimated spiked population eigenvalues is strictly equal to the truth, although the actual estimated spiked eigenvalues may differ due to the randomness. Therefore, it is reasonable to use it for estimating the sources. Hence, we propose that

(A5) The number of sources in MEG equals the intrinsic dimensionality of the data.

2.3. Estimation of Noise Covariance

Consider the transform $\mathbf{Y}(t) \mapsto \frac{1}{\sqrt{T}} \Omega_n \mathbf{Y}(t)$, used to suppress the effect of noise from the model (1). If $\Omega_n = I$, the above transformation is the so-called brute-force transformation (BT). Denote the concentration matrix $\Omega_n = \mathbf{R}_n^{-1}$. If we have a reasonably good estimator $\hat{\Omega}_n$ (or $\hat{\mathbf{R}}_n$), then $\mathbf{Y}(t) \mapsto \frac{1}{\sqrt{T}} \hat{\Omega}_n^{1/2} \mathbf{Y}(t)$; this



Figure 2. The learning process of selecting the optimal threshold ϵ with the initial guess $\epsilon_0 = 10\% \cdot \hat{\lambda}_{7,1}$ and stopping criteria $\epsilon > 40\% \cdot \hat{\lambda}_{7,1}$ (a) Results after the first iteration. $\Delta_1 = 1.67$. (b) Results after the second iteration. $\Delta_2 = 1.42$. (c) Results after the third iteration. $\Delta_3 = 1.13$. (d) Results after the forth iteration. $\Delta_4 = 1.26$. In each histogram, all the double arrows remain the same length (except the last one, since all remaining eigenvalues belong to a class with a smaller threshold), which is the value of the threshold ϵ in the current iteration. Moreover, the number above each double arrow labels the group of the corresponding spiked eigenvalue, which contains the sample eigenvalues within a distance of the double arrow length. "×" with labels 1', 2', ... representing the estimated population spiked eigenvalues, corresponding groups with labels 1, 2,

Table 1. A demonstration of Algorithm 2 with different initial guesses ϵ_0 and sampling distributions used in Step 2 of Algorithm 2. Each vector with four elements in the table denotes the estimated spiked population eigenvalues (SPEs) with the corresponding initial guess ϵ_0 and the distribution.

Estimated SPEs	Gaussian	Uniform	<i>t</i> -distribution	
$\begin{split} \epsilon_0 &= 5\% \cdot \hat{\lambda}_{T,1} \\ \epsilon_0 &= 10\% \cdot \hat{\lambda}_{T,1} \\ \epsilon_0 &= 15\% \cdot \hat{\lambda}_{T,1} \end{split}$	(19.81,17.09,9.89,6.78)	(20.60,16.74,10.38,7.50)	(19.15,16.28,9.64,6.40)	
	(20.42,16.85,10.53,7.9)	(19.13,16.43,10.91,8.71)	(18.65,15.85,9.32,5.22)	
	(20.43,17.64,11.03,8.93)	(19.42,18.32,9.66,5.95)	(17.84,15.14,8.90,6.13)	

is the whitening transformation (WT) that we have used in (6). This being said, challenges remain on how to obtain a good $\hat{\Omega}_n$. Our heuristic is that if we can estimate \mathbf{R}_n reasonably well, particularly the diagonal of the \mathbf{R}_n , the resultant denoised covariance $\hat{\mathbf{R}}_{adj}$ can be used as a good candidate for finding spiked eigenvalues. In this article, we mainly consider $\hat{\Omega}_n = \hat{\mathbf{R}}_n^{-1}$, given an estimate $\hat{\mathbf{R}}_n$. Note that the estimation might vary case by case; only WT is considered. We are now in position to discuss three algorithms of noise estimation accommodating the three kinds of correlations that possibly exist in the data: inner-sensor, intersensor, and the combined one.

The first noise estimation algorithm to use is Fourier transform. The idea is to estimate the noise variance on its frequency domain while the MEG data is sampled in the time domain. The second algorithm is based on residual analysis (Roger and Arnold 1996). We estimate the inverse of noise variance instead of the original one. The last algorithm aims to estimate the noise covariance by thresholding methods (Bickel and Levina 2008), which has been recently employed on MEG data by Zhang and Su (2015). Our updated approach results in a nondiagonal noise covariance matrix. A detailed description of the above three noise estimation algorithms can be found in supplementary materials.

2.4. AIC, MDL, and Malinowski's Method

The information-theoretical criteria AIC (Wax and Kailath 1995) and MDL (Schwarz 1978) will be used for comparison. Determining the number of signals is equivalent to finding the number of free parameters in the model (1). For a fair comparison, we will evaluate AIC and MDL based on the eigenvalues of the $\widehat{\mathbf{R}}_{adj}$,

$$\begin{split} \text{AIC}(N) &= -2\log\left(\frac{\prod_{j=N+1}^{K}\hat{\lambda}_{T,j}^{\frac{1}{K-N}}}{\frac{1}{K-N}\prod_{j=N+1}^{L}\hat{\lambda}_{T,j}}\right)^{(K-N)T} + 2N(2K-N)\\ \text{MDL}(N) &= -\log\left(\frac{\prod_{j=N+1}^{L}\hat{\lambda}_{T,j}^{\frac{1}{K-N}}}{\frac{1}{K-N}\prod_{j=N+1}^{K}\hat{\lambda}_{T,j}}\right)^{(K-N)T} \\ &+ \frac{1}{2}N(2K-N)\log T, \end{split}$$

where *N* is the number of free parameters. In our case, *N* refers to the number of signal sources. If the noise is independent and identically distributed, the problem of finding the number of signal sources can be achieved by minimizing,

Number of sources =
$$\underset{N}{\operatorname{argmin}}$$
 AIC(N)
Number of sources = $\underset{N}{\operatorname{argmin}}$ MDL(N).

Malinowski's method (Malinowski 1977b), a popular factor analysis method, is also used here for comparison, where an empirical indicator function (EIF) (Malinowski 1977a) is introduced as a criterion

EIF(N) =
$$\frac{\left(\sum_{j=N+1}^{K} \hat{\lambda}_{T,j}\right)^{1/2}}{T^{1/2} (K-N)^{3/2}}$$

and the number of sources is estimated by

Number of sources =
$$\underset{N}{\operatorname{argmin}}$$
 EIF(N).

Each of the AIC, MDL and EIF methods tend to overestimate the number of signal sources, since they rely on the independence and normality assumption. In Section 3, their performances are compared with our method.

3. Simulation Study

3.1. Computer Simulation

Before running our algorithms on a real dataset, we tested a simplified case. In this example, we created a channel-level MEG data using a dipole configuration with four dipoles at specified locations in the head. The location and moments parameters of these simulated dipoles are summarized in Table 2. We simulated 128 electrodes (magnetometers) by randomly placing them on the upper part of the unit sphere with a radius of 100 mm. The head was modeled by a concentric three-sphere volume conductor. The radii of three spheres in the conductor is, respectively, 88, 92, and 100mm, with its corresponding conductivity 1, 1/80, 1. The geometrical information and the simulated signal are visualized in Figure 3. Note that the associated parameters for each dipole, such as the locations, did not vary during the simulation. In other words, each dipole contributed a different but constant signal at the same sensor. However, to work with time-varying dipoles, we applied a different frequency to the magnitudes of each dipole so that we could create a distinct time course for each dipole. The time course of each dipole with

Table 2. Illustration of dipole simulation. The location of each dipole (a total of four) is expressed in terms of spherical coordinates (r, θ, ϕ) , where r is radial distance, θ is inclination, and ϕ is azimuth. m_1 and m_2 are the dipole moment parameters. s is the strength parameter of a dipole.

Dipole index	1	2	3	4
<i>r</i> (mm)	0	10	— 10	40
ϕ	0.5	0.1	- 0.5	- 0.3
θ	3	0.1	- 0.3	0.3
<i>m</i> ₁	1	0	0.3	1
m'	0	0.5	0.4	0.7
s (mA)	0	0.5	0.2	0



Figure 3. The sensor map and the simulated signal. (a) The simulated MEG sensors on the skull. (b) The simulated magnetic signal generated by the four dipoles at various sensors (vertical view).

unique frequency was modeled by either a sine or cosine function, as follows:

$$\sin\left(\frac{2\pi}{1/10}t\right), \quad \cos\left(\frac{2\pi}{1/15}t\right), \quad \sin\left(\frac{2\pi}{1/20}t - \frac{\pi}{4}\right),$$
$$\cos\left(\frac{2\pi}{1/30}t - \frac{\pi}{4}\right),$$

where $1 \le t \le T$. The pure magnetic signal produced by each dipole at each sensor was calculated using the Biot–Savart equation. The total length (*T*) of each trial is 1000 timesteps. The magnetometer data were obtained by adding up the contributions from each dipole and the simulated noise across all sensors. To work with different noise levels, we control

$$SNR = \sigma_{signal}^2 / \sigma_{noise}^2$$

where σ_{signal}^2 and σ_{noise}^2 are the variance of the signal and the noise within each trial, respectively. Here, we essentially have $\mathbf{X} = I$ and $\text{SNR} = \mathcal{I}_{R,R_n}(I)$.

We generated five trials of the data, as follows. In each trial, we calculated the variance of pure signal $\sigma_{\text{signal},k}^2$ at each sensor k ($1 \le k \le K$). Then, for a given SNR (1, 0.1, 0.01, 0.001, 0.0001), we set the corresponding variance of noise by $\sigma_{\text{noise},k}^2 = \sigma_{\text{signal},k}^2/\text{SNR}$. Furthermore, by adding Gaussian noise components with the calculated variance to the simulated data, we obtained the noised signal with an expected SNR in each trail. Finally, we averaged the corresponding noised signals over the five trials. Our method, based on the spiked population eigenvalue (SPE), was tested against PCA and other methods such as AIC, MDL, and EIF on this averaged data.

The performance of each method is summarized in Table 3, in terms of the estimated number of dipoles. We can see that

SPE successfully recovers the correct number of dipoles by estimating the number of spiked population eigenvalues, and it outperforms all other methods regardless of the SNR levels. PCA seems to pick up a number of dipoles which is more or less accurate when the SNR is large, but it tends to detect more dipoles for small SNRs. This phenomenon was expected, as the sample eigenvalues of the covariance matrix become increasingly unreliable as a measure of the number of dipoles when the SNR decreases. To illustrate this, we have zoomed in on the distribution of the sample eigenvalues (in "o") and the spiked population eigenvalues (in "*") in Figure 4, for SNR = 1, 0.1, 0.01, 0.001, where the sample eigenvalues of the covariance matrix \mathbf{R} and the quasi-optimal transformed covariance matrix \mathbf{R}_{adj} are displayed, respectively. From Figure 4, we also see that PCA uses the sample eigenvalues that become closer and closer in magnitude, either from $\widehat{\mathbf{R}}$ or $\widehat{\mathbf{R}}_{adi}$, and thus it tends to overestimate the number of dipoles. The spiked population eigenvalues of \mathbf{R}_{adj} estimated by SPE are superimposed, accordingly. It is suggested that, as the SNR decreases, the role of spiked population eigenvalues becomes dramatically more significant. In particular, when the SNR = 0.001, it is impossible to separate the sample eigenvalues, while the SPE still finds the right number of dipoles. Both the AIC and MDL largely overestimate the dipoles across all SNRs, while the number of dipoles estimated by EIF shows a reasonable range that covers the right number of dipoles but still overestimates. More experiments under the same setup can be found in the supplementary materials, to support the accuracy of the performance of SPE.

3.2. Hunting Unknown Signal in the MEG Room

For real MEG data on human subjects, we cannot clarify the accuracy of our methods, since the truth of how many sources may be present in the data is unknown. However, it would still

Table 3. Comparison of results from PCA, AIC, MDL, and EIF and spiked population eigenvalues (SPE) with simulated data. The first column shows the signal-to-noise ratios. The second to eighth columns are the estimated number of dipoles from the simulated data, where the true number of dipoles is four.

SNR	PCA (0.9/0.8/0.7)	AIC	MDL	EIF	SPE (FFT)	SPE (RS)	SPE (TH)
Noise $= 0$	3/3/3	120–127	120–127	4–127	4	4	4
1	21/3/2	120-127	120–127	3–19	4	4	4
0.1	69/48/34	122–127	120–127	1–11	4	4	4
0.01	76/57/44	121–127	120–127	1–11	4	4	4
0.001	78/59/46	120-127	120–127	1–10	4	4	4
0.0001	78/59/46	121–127	121–127	1–13	4	4	4



Figure 4. Illustration of spiked eigenvalues estimated from the covariance matrix $\hat{\mathbf{R}}$ and the adjusted covariance matrix $\hat{\mathbf{R}}_{adj}$ at four different SNRs (totally 4 × 4 subplots). Row 1: in each subplot (from left to right), the estimated spiked population eigenvalues are labeled as "*" and the sample eigenvalues are labeled as "°" at SNR of 1, 0.1, 0.01, and 0.001. The Rows 2–4 give the same information for $\hat{\mathbf{R}}_{adj}$, estimated, respectively, by Fourier transform, residual analysis, and thresholding method.

be quite interesting to see the performance of our methods on specific MEG data where we do know the truth. In the following analysis, a dataset from an empty MEG room will be used, that is, there is no subject in the MEG room. To our knowledge, all the devices in the room that might cause electric potential were turned off, but one device was constantly producing energy around 60 Hz. The magnetic field distribution was recorded by a 306-channel system. A small portion of the dataset, 5000 ms long with only 102 channels (magnetometers), was used in our analysis. Conservatively speaking, there was only one source (60 Hz), or at least one, with high frequency in our data. Our attempt was to verify the existence of this high-frequency source, and to estimate the number of active sources in the room, using our proposed method on this data. The magnitude of the raw data (Figure 5(a)) in the empty room is in the range of -1.6×10^4 fT to 1.3×10^4 fT. We can see that the white lines are equally distant in the modulus plot (Figure 5(b)) of complex Fourier coefficients truncated to 2000 for raw data. This is a clear indication of a periodic source at about 60 Hz in the data. To see if the number of sources that our method detects includes the 60 Hz one, it is necessary for us to run the same analysis in an environment when the 60 Hz is not available. This means we need to filter the 60 Hz signal from the raw data. In fact, we filtered all frequencies above 50 Hz. Figure 6(a) shows the modulus plot of the Fourier coefficients after filtering all frequencies above 50 Hz; all the white lines associated with 60 Hz, 120 Hz, 180 Hz, and so on, disappear. The image after filtering (shown in Figure 6(b)) is reconstructed



Figure 5. Raw data (empty room) and the modulus plot of the data after the Fourier transform. (a) The grayscale plot of the raw data of 5000 ms. The horizontal axis is time (milliseconds); the vertical axis is the channel number (102 channels in total). (b) The modulus plot of the complex-valued Fourier coefficients of the raw data. The horizontal axis is time (milliseconds); the vertical axis is frequency.



Figure 6. Raw data of the empty room after filtering. (a) The modulus plot of the complex-valued Fourier coefficients after filtering all frequencies above 50 Hz truncated to 1000. The horizontal axis is time (milliseconds); the vertical axis is frequency. (b) The grayscale plot of the real part of inverse Fourier transform after filtering coefficients, that is, the data after all coefficients above 50 Hz are zeroed out.

by the inverse Fourier transform of the real part after filtering. We do not show the imaginary part of the filtered inverse transformed data, because the figures are all nearly zero (less than 10×10^{-11} fT).

Note that the values of the measurements are very small in the empty room, and this is particularly true for the data after filtering. To avoid possible round-off error due to floating-point arithmetic, we proceed to normalize the original covariance matrix $\widehat{\mathbf{R}}$ by

$$\widehat{\mathbf{R}}^* = \phi(K) \cdot \frac{\widehat{\mathbf{R}}}{\|\widehat{\mathbf{R}}\|_2}$$

where *K* is the number of sensors (K = 128 in this experiment) and function ϕ describes the contribution of the dimension to the problem. In practice, one can choose $\phi(K) = 1$, *K*, K^2 (we set $\phi(K) = K$ here).

Remark: (1) For the noise estimation, we use the corresponding normalized data by $\mathbf{Y}_{\text{norm}}(t) = \sqrt{K/\|\mathbf{\widehat{R}}\|_2} \cdot \mathbf{Y}(t)$. (2) When the noise effect is negligible, no optimal SNR transformation is needed since a bad noise estimation may even deteriorate the accuracy of the result.

A summary of the performance of the different methods (PCA, AIC, MDL, EIF, and SPE) applied on this data (before filtering and after filtering) is shown in Table 4. Both the AIC and MDL underestimate the number of signal sources, while the EIF overestimates the number of signal sources. All are not able to tell the difference in the number of sources before and after filtering. Both PCA and SPE provide a reasonable estimate of the number of sources: four before filtering and three after filtering for PCA, three before filtering and two after filtering for SPE, while PCA tends to pick up more sources. This leads us to believe that there is at least one (or even two) other active sources (< 50 Hz) that exist in the MEG room.

 Table 4.
 Comparison of results from PCA, AIC, MDL, EIF, and SPE with the raw data of the empty room.

	PCA	AIC	MDL	EIF	SPE (FFT/RS/TH)
Before filtering	4	1	1	2–31	3
After filtering	3	1–2	1–2	3–30	2

4. Brain-Controlled Interfaces Data

The real data analysis reports results of finding varying brain sources in a Brain-Controlled Interfaces (BCI) experiment. The data consist of 28,000 ms recordings collected at the Center for Advanced Brain Magnetic Source Imaging (CABMSI) at Presbyterian University Hospital in Pittsburgh. In the first part of the experiment, the subjects were asked to imagine performing the "center-out" task using the wrist (imagined movement task), and in the second part, the subjects controlled a 2D cursor using the wrist to perform the center-out task following a visual target (overt movement task). The subject, as illustrated in Figure 7 (the time scale in the real experiment can be different), holds the 2D cursor and waits for the cursor to go to the center. To complete the trial successfully, the cursor stays at the center for a short period until the peripheral target appears. The cursor needs to move out to the target and stays there for another short period. The target changes color when hit by the cursor, and disappears when the holding period has finished.

The analysis is based on the raw MEG data from 102 magnetometers, without performing spatial filtering or smoothing. The goal is to investigate the dynamics of the sources in the data.



Figure 7. A BCI experiment chart. The bottom trace shows the speed profile of the cursor from a representative trial, and the dotted lines delimit the premovement/planning period.



Figure 8. Estimation of the number of signal sources in the BCI data. The horizontal axis is time (seconds); the vertical axis is number of estimated sources. (a) the noise estimation is based on the Fourier transform; the blue curve shows the estimated number of sources for each moving window with a length of 2000 ms; the estimated number of sources at each equidistant partition (2000 ms) of the data is also shown as a diamond in red. (b) and (c) display the same information, but with the noise estimated by the residual analysis and thresholding methods, respectively. (d) displays the same result with no noise estimation.

Figure 8 presents the evolution of the number of possible active sources estimated during the experiment. As can be expected, there tend to be more active sources during the imagined movement task period than overt movement task period. This reveals a possible delay of the observed magnetic signal reaching its peak. This can also be partially explained by the fact that the subject was engaged in the imagined movement task by the experimenter, that is, catch trials were inserted. The data also serve to show that there are still some active sources presented after the movement, which might come from the holding period of the subject preparing for the next task.

To illustrate the effect of partition of the data on the estimation, a dual-scale point of view is used: (1) the equidistant partition of the data, (2) a moving window scheme. For the equidistant partition, we equally divide the data into several chunks, each with a length of 2000 ms; for the moving window, we choose the window of the same length (2000 ms) and move the window from 0 ms at the rate of 600 ms per move. The numbers of the sources estimated by these two methods are depicted in Figure 8, respectively. The effects of choosing which partition to use are not so apparent in this particular data. In this case, we can see that there is no significant change to the estimated sources by moving the window, suggesting that the estimated number of sources at each separate time interval seems good enough to represent the dynamics of the number of sources. This provides us with some flexibility in using the dual-scale point of view when presenting the number of sources. The difference with respect to the number of sources lying in-between the overt and imagined movements shown in our analysis has been consistent with that in Wang et al. (2010), where the main finding is to decode the intended movement direction in the absence of overt movement using MEG.

We further examine the effect of noise estimation on the number of sources. Interestingly, Figure 8(a)-8(c) shows that noise estimation has generally detected additional active sources rather than no noise estimation (Figure 8(d)). Since the data have been normalized beforehand, any change of the number of the estimated spiked eigenvalues would imply a change of the sources. This phenomenon is expected as there might be some hidden eigenvalues from the signal part that may not be easily revealed when the noise presents. We note that, during the time period of 24,000-25,999 ms, the eigenvalues before and after noise estimation (i.e., Fourier transform) are (102, 3.5, 0.8, 0.1, ...) (see dotted line in Figure 8(a)) and (102, 10.5, 3.3, 0.4, ...) (see dotted line in Figure 8(d)), which suggests that the eigenvalue 10.5 is possibly from a potential source. Similar conclusions can be drawn for the other two methods of noise estimation. This further indicates that the optimal SNR transformation is helpful in estimating the number of sources.

5. Discussion

The determination of the number of signal sources in the MEG data is a very challenging problem. Due to the noisy characteristics of the MEG data, effective methods for dealing with this problem are lacking. Conventional approaches such as PCAbased methods or methods involving information criteria are essentially not helpful in deciding the dimensionality of the data. The difficulty lies in the fact that those approaches simply use the sample eigenvalue distribution, where the sample eigenvalues are not consistent estimators of the population counterparts when the dimension of the data is proportional to the sample size. As is the case with the MEG data, the sample eigenvalues are still mixtures of the signal and noise in the data, and it is quite hard to detect the energy that such a signal contributes to the eigenvalues, compared with the noise.

With the aim of proposing a framework that allows the flexible estimation of the intrinsic dimensionality of the data in MEG, we introduced the notation of spiked covariance model. We showed the importance of estimating the eigenvalues of the population covariance matrix, and the difference if this method from only using the sample counterparts. The spiked eigenvalues were found to be interpretable as a measure of the number of signal sources in the MEG data. Depending on how the SNR varies, the spiked covariance model was seen to be more reliable in estimating the number of signal sources. In this sense, it can be thought of as a significant improvement on the methods which use the sample eigenvalue distribution.

We show that the optimality of the eigenvalue distribution is achieved on the transformed data matrix rather than the original data matrix. However, the optimal SNR transformation requires a reasonably acceptable accuracy of the noise estimation. There have always been difficulties in defining noise in neurological experiments; in particular, for the MEG data, the noise structure is very complicated. This article is mainly focused on two noise structures: the independent noise (Fourier method), and the correlated noise (residual method, thresholding method), where in both situations, we aim to recover the main diagonal of the noise covariance matrix at the sensors. In the examples we considered, the estimated covariance matrices under the optimal SNR yielded visually apparent population eigenvalues that were remarkably different from the sample eigenvalues. From case to case, both methods seemed to capture particularly relevant underlying influential eigenvalues of the data, while the sample eigenvalues were seen as quite noisy. In fact, we noted that the spiked eigenvalues estimated from either method did not have much of an effect on the actual number of the spiked eigenvalues. This points to the potential for future work, on a further investigation of the noise estimation and its influence on the estimated spiked eigenvalues.

The performance of different methods on estimating the number of signal sources was examined in a simulated example, with varying SNRs. In all cases, the spiked population eigenvalues estimated gave quite robust results consistent with the true number of the sources; based on the same transformed covariance data matrix, the PCA as well as AIC, MDL, and EIF approaches only worked reasonably well for just a few cases, when the SNR was large, and failed for most situations. We also attempted to hunt for unknown sources existing in a shielded MEG room. We confirmed the existence of a single 60 Hz source in the room. In addition, another one or two potential sources were detected. One advantage of using our method is that we could possibly detect the hidden signal sources that are different in frequency, which can be further used in identifying high-frequency oscillation in the evaluation of epilepsy or other presurgical operations. In the BCI data, because the number of sources may change over time, the associated data covariance in the model would change as well. This fact, together the necessity of noise estimation incorporated in the optimal SNR transformation, suggests that a dynamic implementation of the estimation is preferable. We have shown that it is not necessary to estimate the number of sources from the entire data at once.

but rather that attempting to estimate it for each time point sequentially can also produce very stable results.

In conclusion, we have been trying to find a way of estimating the number of signal sources in the MEG data. Though the number of spiked population eigenvalues appears to be a useful means of guiding the practitioners' choice on the sources before further applying localization methods, it is certainly not the only choice. The issue of assessing which criterion is the most informative one to decide on the intrinsic dimensionality of the data deserves further scrutiny. Our method can serve as a reference.

Appendix A: Proof of Theorem 1

Proof. For any fixed $\mathbf{X} \in \mathbb{R}^{K \times K}$, there exists a positive number *C* such that $\|\mathbf{X}\|_2 \leq C$. Define

$$\theta_1 = 1 - \frac{2}{C(\lambda_{\max} \|\mathbf{R}_n\|_2 + \|\mathbf{R}\|_2)} \text{ and } \theta_0 = (1 - \theta_1) \cdot C \|\mathbf{R}\|_2 + 1.$$
(A 1)

Note that $1 = \theta_0 - (1 - \theta_1) \cdot C \|\mathbf{R}\|_2$ and $\theta_1 < 1$, we obtain

$$\|\mathbf{X}^{T}\mathbf{R}\mathbf{X} - \mathbf{X}\|_{2} \leq \|\mathbf{X}^{T}\mathbf{R}\mathbf{X}\|_{2} + \|\mathbf{X}\|_{2} = \|\mathbf{X}^{T}\mathbf{R}\mathbf{X}\|_{2} + (\theta_{0} - (1 - \theta_{1}) \cdot C\|\mathbf{R}\|_{2}) \|\mathbf{X}\|_{2} \leq \|\mathbf{X}^{T}\mathbf{R}\mathbf{X}\|_{2} + \theta_{0}\|\mathbf{X}\|_{2} - (1 - \theta_{1}) \cdot \|\mathbf{X}\|_{2} \cdot \|\mathbf{R}\|_{2} \cdot \|\mathbf{X}\|_{2} \leq \|\mathbf{X}^{T}\mathbf{R}\mathbf{X}\|_{2} + \theta_{0}\|\mathbf{X}\|_{2} - (1 - \theta_{1}) \cdot \|\mathbf{X}^{T}\mathbf{R}\mathbf{X}\|_{2} = \theta_{0}\|\mathbf{X}\|_{2} + \theta_{1}\|\mathbf{X}^{T}\mathbf{R}\mathbf{X}\|_{2},$$

which implies that

$$\begin{aligned} \|\mathbf{X}^T \mathbf{R} \mathbf{X}\|_2 &\leq \|\mathbf{X}^T \mathbf{R} \mathbf{X} - \mathbf{X}\|_2 + \|\mathbf{X}\|_2 \leq \theta_0 \|\mathbf{X}\|_2 + \theta_1 \|\mathbf{X}^T \mathbf{R} \mathbf{X}\|_2 + \|\mathbf{X}\|_2 \\ &= (1 + \theta_0) \|\mathbf{X}\|_2 + \theta_1 \|\mathbf{X}^T \mathbf{R} \mathbf{X}\|_2. \end{aligned}$$

Hence, we obtain

$$\|\mathbf{X}^{T}\mathbf{R}\mathbf{X}\|_{2} \leq \frac{1+\theta_{0}}{1-\theta_{1}}\|\mathbf{X}\|_{2}.$$
(A.2)

Furthermore, define (note that **R** is nonsingular by the assumption)

$$\theta_2 = -1 - \lambda_{\max} \frac{\|\mathbf{R}_n\|_2}{\|\mathbf{R}\|_2}.$$
 (A.3)

By definitions of θ_0 , θ_1 , and θ_2 in (A.1) and (A.3), respectively, we have

$$(1-\theta_2)C\|\mathbf{R}_n\|_2 + \frac{(1+\theta_0)(1+\theta_2)}{(1-\theta_1)\lambda_{\max}} = 0,$$

which implies that

$$(1 - \theta_2) \| \mathbf{X}^T \mathbf{R}_n \mathbf{X} \|_2 + \frac{(1 + \theta_0)(1 + \theta_2)}{(1 - \theta_1)\lambda_{\max}} \| \mathbf{X} \|_2$$

$$\leq (1 - \theta_2) C \| \mathbf{X} \|_2 \| \mathbf{R}_n \|_2 + \frac{(1 + \theta_0)(1 + \theta_2)}{(1 - \theta_1)\lambda_{\max}} \| \mathbf{X} \|_2$$

$$= \left\{ (1 - \theta_2) C \| \mathbf{R}_n \|_2 + \frac{(1 + \theta_0)(1 + \theta_2)}{(1 - \theta_1)\lambda_{\max}} \right\} \| \mathbf{X} \|_2 = 0.$$

Then, using the following inequalities

$$\begin{split} \|\mathbf{X}^{T}\mathbf{R}_{n}\mathbf{X} - \mathbf{X}\|_{2} &\leq \|\mathbf{X}^{T}\mathbf{R}_{n}\mathbf{X}\|_{2} + \|\mathbf{X}\|_{2} \\ &= \theta_{2}\|\mathbf{X}^{T}\mathbf{R}_{n}\mathbf{X}\|_{2} + (1 - \theta_{2})\|\mathbf{X}^{T}\mathbf{R}_{n}\mathbf{X}\|_{2} \\ &+ \left(1 - \frac{(1 + \theta_{0})(1 + \theta_{2})}{(1 - \theta_{1})\lambda_{\max}}\right)\|\mathbf{X}\|_{2} \\ &+ \frac{(1 + \theta_{0})(1 + \theta_{2})}{(1 - \theta_{1})\lambda_{\max}}\|\mathbf{X}\|_{2} \\ &\leq \left(1 - \frac{(1 + \theta_{0})(1 + \theta_{2})}{(1 - \theta_{1})\lambda_{\max}}\right)\|\mathbf{X}\|_{2} + \theta_{2}\|\mathbf{X}^{T}\mathbf{R}_{n}\mathbf{X}\|_{2} \\ &+ \left\{(1 - \theta_{2})C\|\mathbf{R}_{n}\|_{2} + \frac{(1 + \theta_{0})(1 + \theta_{2})}{(1 - \theta_{1})\lambda_{\max}}\right\}\|\mathbf{X}\|_{2} \\ &= \left(1 - \frac{(1 + \theta_{0})(1 + \theta_{2})}{(1 - \theta_{1})\lambda_{\max}}\right)\|\mathbf{X}\|_{2} + \theta_{2}\|\mathbf{X}^{T}\mathbf{R}_{n}\mathbf{X}\|_{2}, \end{split}$$

we deduce that $\|\mathbf{X}^T \mathbf{R}_n \mathbf{X} - \mathbf{X}\|_2 \le \theta_3 \|\mathbf{X}\|_2 + \theta_2 \|\mathbf{X}^T \mathbf{R}_n \mathbf{X}\|_2$, where $\theta_3 = (1 - \frac{(1+\theta_0)(1+\theta_2)}{(1-\theta_1)\lambda_{\max}})$. This implies

$$\begin{split} \|\mathbf{X}^T \mathbf{R}_n \mathbf{X}\|_2 &\geq \|\mathbf{X}\|_2 - \|\mathbf{X}^T \mathbf{R}_n \mathbf{X} - \mathbf{X}\|_2 \\ &\geq \|\mathbf{X}\|_2 - \theta_3 \|\mathbf{X}\|_2 - \theta_2 \|\mathbf{X}^T \mathbf{R}_n \mathbf{X}\|_2 \\ &= (1 - \theta_3) \|\mathbf{X}\|_2 - \theta_2 \|\mathbf{X}^T \mathbf{R}_n \mathbf{X}\|_2, \end{split}$$

which leads to the inequality

$$\|\mathbf{X}^{T}\mathbf{R}_{n}\mathbf{X}\|_{2} \geq \frac{1-\theta_{3}}{1+\theta_{2}}\|\mathbf{X}\|_{2}.$$
 (A.4)

Since $\frac{1-\theta_3}{1+\theta_2} = \frac{1+\theta_0}{(1-\theta_1)\lambda_{max}} > 0$, combine (A.2) and (A.4) to get

$$\mathcal{I}_{\mathbf{R},\mathbf{R}_n}(\mathbf{X}) = \frac{\|\mathbf{X}^T \mathbf{R} \mathbf{X}\|_2}{\|\mathbf{X}^T \mathbf{R}_n \mathbf{X}\|_2} \leq \frac{\frac{1+\theta_0}{1-\theta_1}}{\frac{1-\theta_2}{1+\theta_2}} = \lambda_{\max},$$

which yields the boundedness of $\mathcal{I}_{\mathbf{R},\mathbf{R}_n}$.

Finally, let us show that transformations $\mathbf{X}_{i}^{\text{opt}}$ (i = 1, 2) in (4)–(5) are the required solutions. Denote by I_{K} a *K*-dimensional identity matrix. By noting that $\mathbf{\Phi}_{n}$ and $\mathbf{\Phi}_{\text{adj}}$ are unitary matrices, we obtain

$$\begin{aligned} \mathcal{I}_{\mathbf{R},\mathbf{R}_{n}}(\mathbf{X}_{2}^{\text{opt}})^{T}\mathbf{R}\mathbf{X}_{2}^{\text{opt}}\|_{2} &= \frac{\|\mathbf{\Phi}_{\text{adj}}^{T}\mathbf{\Lambda}_{n}^{-1/2}\mathbf{\Phi}_{n}^{T}\mathbf{R}\mathbf{\Phi}_{n}\mathbf{\Lambda}_{n}^{-1/2}\mathbf{\Phi}_{\text{adj}}\|_{2}}{\|\mathbf{\Phi}_{\text{adj}}^{T}\mathbf{\Lambda}_{n}^{-1/2}\mathbf{\Phi}_{n}^{T}\mathbf{R}_{n}\mathbf{\Phi}_{n}\mathbf{\Lambda}_{n}^{-1/2}\mathbf{\Phi}_{\text{adj}}\|_{2}} \\ &= \frac{\|\mathbf{\Lambda}_{n}^{-1/2}\mathbf{\Phi}_{n}^{T}\mathbf{R}\mathbf{\Phi}_{n}\mathbf{\Lambda}_{n}^{-1/2}\|_{2}}{\|\mathbf{\Lambda}_{n}^{-1/2}\mathbf{\Phi}_{n}^{T}\mathbf{R}_{n}\mathbf{\Phi}_{n}\mathbf{\Lambda}_{n}^{-1/2}\|_{2}} \left(= \frac{\|\mathbf{W}_{n}^{T}\mathbf{R}\mathbf{W}_{n}\|_{2}}{\|\mathbf{W}_{n}^{T}\mathbf{R}_{n}\mathbf{W}_{n}\|_{2}} = \mathcal{I}_{\mathbf{R},\mathbf{R}_{n}}(\mathbf{X}_{1}^{\text{opt}}) \right) \\ &= \frac{\|\mathbf{\Lambda}_{\text{adj}}\|_{2}}{\|\mathbf{I}_{K}\|_{2}} = \lambda_{\text{max}}, \end{aligned}$$

which yields the required result and we prove the theorem completely. $\hfill \Box$

Appendix B: Proof of Theorem 2

Proof. By condition (7) and the boundedness of $\|\mathbf{\Lambda}_n^{-1/2}\|$ we have

$$\begin{split} \|\widehat{\boldsymbol{\Phi}}_{n}\widehat{\boldsymbol{\Lambda}}_{n}^{-1/2} - \boldsymbol{\Phi}_{n}\boldsymbol{\Lambda}_{n}^{-1/2}\| \\ &\leq \|\widehat{\boldsymbol{\Phi}}_{n}\widehat{\boldsymbol{\Lambda}}_{n}^{-1/2} - \widehat{\boldsymbol{\Phi}}_{n}\boldsymbol{\Lambda}_{n}^{-1/2}\| + \|\widehat{\boldsymbol{\Phi}}_{n}\boldsymbol{\Lambda}_{n}^{-1/2} - \boldsymbol{\Phi}_{n}\boldsymbol{\Lambda}_{n}^{-1/2}\| \\ &\leq \|\widehat{\boldsymbol{\Phi}}_{n}\| \cdot \|\widehat{\boldsymbol{\Lambda}}_{n}^{-1/2} - \boldsymbol{\Lambda}_{n}^{-1/2}\| + \|\widehat{\boldsymbol{\Phi}}_{n} - \boldsymbol{\Phi}_{n}\| \cdot \|\boldsymbol{\Lambda}_{n}^{-1/2}\| \leq (1+C)\boldsymbol{\omega} \end{split}$$

which implies that $\|\widehat{\mathbf{W}}_n - \mathbf{W}_n\| = \|\widehat{\mathbf{\Phi}}_n\widehat{\mathbf{\Lambda}}_n^{-1/2} - \mathbf{\Phi}_n\mathbf{\Lambda}_n^{-1/2}\| \le (1 + C)\omega.$

Similarly, it is easy to obtain the following inequality

$$\|\widehat{\mathbf{\Lambda}}_n^{-1/2}\| \le \|\widehat{\mathbf{\Lambda}}_n^{-1/2} - \mathbf{\Lambda}_n^{-1/2}\| + \|\mathbf{\Lambda}_n^{-1/2}\| \le \omega + C.$$

Then, using the above inequalities and definitions of R_{adj} and $\widehat{R}_{\text{adj}},$ we conclude that

$$\begin{split} \|\widehat{\mathbf{R}}_{adj} - \mathbf{R}_{adj}\| &= \|\widehat{\mathbf{W}}_n^T \widehat{\mathbf{R}} \widehat{\mathbf{W}}_n - \mathbf{W}_n^T \mathbf{R} \mathbf{W}_n\| \\ &\leq \|\widehat{\mathbf{W}}_n^T \widehat{\mathbf{R}} \widehat{\mathbf{W}}_n - \widehat{\mathbf{W}}_n^T \widehat{\mathbf{R}} \mathbf{W}_n\| + \|\widehat{\mathbf{W}}_n^T \widehat{\mathbf{R}} \mathbf{W}_n - \widehat{\mathbf{W}}_n^T \mathbf{R} \mathbf{W}_n\| \\ &+ \|\widehat{\mathbf{W}}_n^T \mathbf{R} \mathbf{W}_n - \mathbf{W}_n^T \mathbf{R} \mathbf{W}_n\| \\ &\leq \|\widehat{\mathbf{W}}_n^T \widehat{\mathbf{R}}\| \cdot \|\widehat{\mathbf{W}}_n - \mathbf{W}_n\| + \|\widehat{\mathbf{W}}_n^T\| \cdot \|\widehat{\mathbf{R}} - \mathbf{R}\| \\ &\cdot \|\mathbf{W}_n\| + \|\widehat{\mathbf{W}}_n^T - \mathbf{W}_n^T\| \cdot \|\mathbf{R} \mathbf{W}_n\| \\ &= \left(\|\widehat{\mathbf{W}}_n^T \widehat{\mathbf{R}}\| + \|\mathbf{R} \mathbf{W}_n\|\right) \|\widehat{\mathbf{W}}_n - \mathbf{W}_n\| + \|\widehat{\mathbf{W}}_n\| \\ &\cdot \|\mathbf{W}_n\| \cdot \|\widehat{\mathbf{R}} - \mathbf{R}\| \\ &\leq \left(\|(\widehat{\mathbf{\Phi}}_n \widehat{\mathbf{A}}_n^{-1/2})^T \widehat{\mathbf{R}}\| + \|\mathbf{R}(\mathbf{\Phi}_n \mathbf{A}_n^{-1/2})\|\right) (1 + C)\omega \\ &+ \|\widehat{\mathbf{\Phi}}_n \widehat{\mathbf{A}}_n^{-1/2}\| \cdot \|\mathbf{\Phi}_n \mathbf{A}_n^{-1/2}\|\omega \\ &\leq \left(\|\widehat{\mathbf{\Phi}}_n\| \cdot \|\widehat{\mathbf{A}}_n^{-1/2}\| \cdot \|\mathbf{R}\| + \|\mathbf{R}\| \cdot \|\mathbf{\Phi}_n\| \cdot \|\mathbf{A}_n^{-1/2}\|\right) \\ &\times (1 + C)\omega \\ &+ \|\widehat{\mathbf{\Phi}}_n\| \cdot \|\widehat{\mathbf{A}}_n^{-1/2}\| \cdot \|\mathbf{\Phi}_n\| \| \cdot \|\mathbf{A}_n^{-1/2}\| \cdot \omega \\ &\leq ((\omega + C)\|\mathbf{R}\| + \|\mathbf{R}\|C) (1 + C)\omega + (\omega + C)C\omega \\ &= (2C(1 + C)\|\mathbf{R}\| + C^2)\omega + ((1 + C)\|\mathbf{R}\| + C)\omega^2 \\ &= O(\omega), \end{split}$$

which yields the required result.

Supplementary Materials

A brief description of the forward and inverse problems in MEG, a complete integration of Algorithms 1–3 in Section 2.2, and detailed descriptions of the three algorithms for noise estimation can be found in the online supplementary materials.

Acknowledgments

The authors thank Dr. Anto Bagic for collecting the empty room MEG data and Professor Rob Kass for the BCI data. The authors' warm thanks also go to three referees and the associate editor for a number of constructive comments.

Funding

This research is supported by Yao's MOE Tier 1 (R-155-000-154-133) and Tier 2 (R-155-000-184-112) Grant Award.

ORCID

Ye Zhang () http://orcid.org/0000-0003-4023-6352 Zhidong Bai () http://orcid.org/0000-0002-5300-5513

References

Bai, Z. (1999), "Methodologies in Spectral Analysis of Large-Dimensional Random Matrices, A Review," *Statistica Sinica*, 9, 611–677. [506]

- Bai, Z., and Ding, X. (2012), "Estimation of Spiked Eigenvalues in Spiked Models," *Random Matrices: Theory and Applications*, 1, 1150011. [508]
- Baik, J., Ben Arous, G., and Péché, S. (2005), "Phase Transition of the Largest Eigenvalue for Nonnull Complex Sample Covariance Matrices," *Annals* of Probability, 33, 1643–1697. [508]
- Bartlett, M. S. (1954), "A Note on the Multiplying Factors for Various χ² Approximations," *Journal of the Royal Statistical Society*, Series B, 16, 296–298. [506]
- Bickel, P., and Levina, E. (2008), "Covariance Regularization by Thresholding," Annals of Statistics, 36, 2577–2604. [511]
- Cohen, D. (1968), "Magnetoencephalography: Evidence of Magnetic Fields Produced by Alpha Rhythm Currents," *Science*, 161, 784–786. [505]
- Green, A., Berman, M., Switzer, P., and Craig, M. (2001), "A Transformation for Ordering Multispectral Data in Terms of Imagequality With Implications for Noise Removal," *IEEE Transactions on Geoscience and Remote Sensing*, 26, 65–74. [506]
- Hamalainen, M. S., and Ilmoniemi, R. J. (1994), "Interpreting Magnetic Fields of the Brain: Minimum Norm Estimates," *Medical & Biological Engineering & Computing*, 32, 35–42. [505]
- Lawley, D. N. (1956), "Tests of Significance for the Latent Roots of Covariance and Correlation Matrices," *Biometrica*, 43, 128–136. [506]
- Ligget, W. S. (1973), "Passive Sonar: Fitting Models to Multiple Time Series," in *Signal Processing*, eds. J. W. R. Griffith, P. L. Stocklin, and C. van Schoneveld, New York: Academic Press, pp. 327–345. [506]
- Malinowski, E. R. (1977a), "Determination of the Number of Factors and the Experimental Error in a Data Matrix," *Analytical Chemistry*, 49, 612–617. [506,511]
- —— (1977b), "Theory of Error in Factor Analysis," *Analytical Chemistry*, 49, 606–612. [511]
- Mattout, J., Phillips, C., Penny, W., Rugg, M., and Friston, K. (2006), "MEG Source Localization Under Multiple Constraints: An Extended Bayesian Framework," *NeuroImage*, 30, 753–767. [505]
- Mosher, J., and Leahy, R. (1998), "Recursive MUSIC: A Framework for EEG and MEG Source Localization," *IEEE Transactions on Biomedical Engineering*, 45, 1342–1354. [505]

- Pulvermuller, F., Shtyrov, Y., and Ilmoniemi, R. (2003), "Spatiotemporal Dynamics of Neural Language Processing: An MEG Study Using Minimum-Norm Current Estimates," *NeuroImage*, 20, 1020–1025. [505]
- Roger, R. E., and Arnold, J. F. (1996), "Reliably Estimating the Noise in AVIRIS Hyperspectral Images," *International Journal of Remote Sensing*, 17, 1951–1962. [511]
- Schmidt, D. M., George, J. S., and Wood, C. C. (1999), "Bayesian Inference Applied to the Electromagnet Inverse Problem," *Human Brain Mapping*, 7, 195–212. [505]
- Schmidt, R. (1986), "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions on Antennas and Propagation*, 34, 276–280. [506]
- Schwarz, G. (1978), "Estimating the Dimension of a Model," Annals of Statistics, 6, 461–464. [511]
- Uutela, K., Hamalainen, M. S., and Somersalo, E. (1999), "Visualization of Magnetoencephalographic Data Using Minimum Current Estimates," *NeuroImage*, 10, 173–180. [505]
- Vanni, S., and Uutela, K. (2000), "Foveal Attention Modulates Responses to Peripheral Stimuli," *Journal of Neurophysiology*, 83, 2443–2452. [505]
- Van Veen, B., Joseph, J., and Hecox, K. (1992), "Localization of Intra-Cerebral Sources of Electrical Activity via Linearly Constrained Minimum Variance Spatial Filtering," in *Proceedings of IEEE Workshop on Statistical Signal and Array Processing*, pp. 526–529. [506]
- Wang, W., Sudre, G. P., Xu, Y., Kass, R. E., Collinger, J. L., Degenhart, A. D., Bagic, A. I., and Weber, D. J. (2010), "Decoding and Cortical Source Localization for Intended Movement Direction With MEG," *Journal of Neurophysiology*, 104, 2451–2461. [515]
- Wax, M., and Kailath, T. (1995), "Detection of Signals by Information Theoretic Criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33, 387–392. [511]
- Yao, Z., and Eddy, W. F. (2014), "A Statistical Approach to the Inverse Problem in Magnetoencephalography," *Annals of Applied Statistics*, 8, 1119–1144. [506]
- Zhang, J., and Su, L. (2015), "Temporal Autocorrelation-Based Beamforming With MEG Neuroimaging Data," *Journal of the American Statistical Association*, 110, 1375–1388. [511]