Journal of Machine Learning Research 1 (2015) 2099-47 Submitted 10/2012; Published 12/2015

On Linearly Constrained Minimum Variance Beamforming

Jian Zhang JZ79@KENT.AC.UK
Chao Liu CL304@KENT.AC.UK
School of Mathematics, Statistics and Actuarial Science
University of Kent
Canterbury, Kent CT2 7NF, UK

Editor: Xiaotong Shen

Abstract

Beamforming is a widely used technique for source localization in signal processing and neuroimaging. A number of vector-beamformers have been introduced to localize neuronal activity by using magnetoencephalography (MEG) data in the literature. However, the existing theoretical analyses on these beamformers have been limited to simple cases, where no more than two sources are allowed in the associated model and the theoretical sensor covariance is also assumed known. The information about the effects of the MEG spatial and temporal dimensions on the consistency of vector-beamforming is incomplete. In the present study, we consider a class of vector-beamformers defined by thresholding the sensor covariance matrix, which include the standard vector-beamformer as a special case. A general asymptotic theory is developed for these vector-beamformers, which shows the extent of effects to which the MEG spatial and temporal dimensions on estimating the neuronal activity index. The performances of the proposed beamformers are assessed by simulation studies. Superior performances of the proposed beamformers are obtained when the signal-to-noise ratio is low. We apply the proposed procedure to real MEG data sets derived from five sessions of a human face-perception experiment, finding several highly active areas in the brain. A good agreement between these findings and the known neurophysiology of the MEG response to human face perception is shown.

Keywords: MEG neuroimaging, vector-beamforming, sparse covariance estimation, source localization and reconstruction

1. Introduction

MEG is a non-invasive imaging technique that records brain activity with high temporal resolution. Postsynaptic current flow within the dendrites of active neurons generates a magnetic field that can be measured close to the scalp surface by use of sensors (Hämäläinen et al., 1993). The magnitude of these measured fields is directly related to neuronal current strength, and hence their measurement will reflect the amplitude of brain activity. The major challenge, however, is to localize active regions inside the head, given the measured magnetic fields outside the head (i.e., given MEG data). This is an ill-posed problem of source localization since the magnetic fields could be caused by an infinite number of neuronal regions. Mathematically, the problem can be stated as follows: one observes a vector of time-series $\mathbf{Y}(t) = (Y_1(t), ..., Y_n(t))^T \in \mathbb{R}^n$, $t = t_j, 1 \leq j \leq J$ from n sensors,

which are linked to candidate sources located at $r_k, 1 \leq k \leq p$ in the brain via the model

$$\mathbf{Y}(t) = \sum_{k=1}^{p} H_k \mathbf{m}_k(t) + \varepsilon(t), \tag{1.1}$$

where H_k is an $n \times 3$ lead field matrix at r_k (i.e., the unit output of the candidate source at location r_k, which is derived from Maxwell's equations), $\mathbf{m}_k(t)$ with covariance matrix Σ_k is a 3×1 moment (time-course) at time t and location r_k, $\varepsilon(t)$ with covariance matrix $\sigma_0^2 I_n$ represents white noises at the MEG channels, and I_n is the $n \times n$ identity matrix. See Mosher et al. (1999) for more details. In practice, when candidate source locations (i.e., voxels) are created by discretizing the source space in the brain, the number of these sources can be substantially larger than the number of available sensors. Moreover, unlike the traditional functional data, not only source time courses but also sensor readings are spatially correlated. Therefore, searching for a small set of latent sources of non-null powers from a large number of candidates poses a challenge to standard i.i.d. sample-based methods in functional data analysis (Ramsay, 2006). Here, the source power at location r_k is referred as the trace of the covariance matrix Σ_k.

Two types of approaches have been proposed for handling the above problem in the literature: global approach and local approach (e.g., Henson et al., 2011; Bolstad et al., 2009; Van Veen et al., 1997; Robinson, 1999; Huang et al., 2004; Quraan et al., 2011). In the global approach, one puts all candidate sources into the model and solves a sparse estimation problem. In the local approach, on other hand, one invokes a list of local models, each is tailored to a particular candidate region. The global approach often requires to specify parametric models, while the local approach is model-free. When the number of candidate sources p is small or moderate compared to the number of available sensors n, one may use a Bayesian method to infer latent sources, with helps of computationally intensive algorithms (e.g., Henson et al., 2011). To make an accurate inference, a large p should be chosen. However, when p is large, the global approach may be computationally intractable and the local approach is preferred. Here, we focus on the so-called linearly constrained minimum variance (LCMV) beamforming (also called vector-beamforming), a local method for solving the above large-p-small-n problem. It involves two steps as follows:

- **Projection step.** For location r_k in the source space, one searches for the optimal $n \times 3$ weighting-matrix W by minimizing the trace of the sample covariance of the projected data $W^T Y(t_j)$, $1 \leq j \leq J$, subject to $W^T H_k = I_3$, where I_3 is a 3×3 identity matrix. This gives the optimal trace

$$\hat{S}_k = \mathrm{tr}([H_k^T \hat{C}^{-1} H_k]^{-1}), \tag{1.2}$$

  where $\hat{C}$ is a sensor covariance estimator and for any invertible matrix A, A^{-1} denotes its inverse, and $\mathrm{tr}(\cdot)$ stands for the matrix trace operator. See Van Veen et al. (1997) for the details.

- **Mapping step.** For each location r_k, calculate the neuronal activity index $\hat{S}_k/(\sigma_0^2 \mathrm{tr}([H_k^T H_k]^{-1}))$, where σ_0^2 is estimated by certain baseline noise data such as the pre-stimulus data. Plot the index against the grid points, creating a neuronal activity map over a given temporal window.

In the **projection step**, the procedure aims at estimating the desired signal from each chosen location while minimizing the contributions of other unknown locations in the presence of noises by optimizing the variation of the projected data. This can be easily seen from the following decomposition of the projected covariance under the constrain $W^T H_k = I_3$:

$$\begin{aligned}\mathrm{tr}\left(\mathrm{cov}(W^T\mathbf{Y}(t))\right) &= \mathrm{tr}(\Sigma_k) + \mathrm{tr}(W^T\mathrm{cov}(\sum_{j\neq k} H_j m_j(t) + \varepsilon(t))W) \\ &\quad +2\mathrm{tr}(\mathrm{cov}(m_k(t), W^T(\sum_{j\neq k} H_j m_j(t) + \varepsilon(t)))),\end{aligned}$$

where the first term is the underlying signal strength at r_k and the last two terms are the contributions of other locations and background noises to the estimated strength of the signal at r_k. Therefore, minimizing the trace of the projected covariance of the data with respect to W is equivalent to minimizing the the contributions of other locations and background noises to estimating the true signal strength at r_k. The further mathematical details can be found in Sekihara and Nagarajan (2008). As pointed out before, in practice, we often have the baseline noise data. Performing the above projection procedure on the noise data under the assumption that the noise covariance matrix is approximately $\sigma_0^2 I_n$, we obtain the optimal trace of the covariance matrix of the projected noise at r_k, $\sigma_0^2\mathrm{tr}([H_k^T H_k]^{-1})$. This implies that the above neuronal activity index is a signal-to-noise ratio (SNR) at location r_k. Therefore, the map generated in the **mapping step** is a SNR map. A similar formula can be derived under a general model of the noise covariance. However, to avoid high-dimensional effects on estimating sensor covariance matrices, we often employed a diagonal noise covariance model even when the true one is not diagonal.

Both theoretical and empirical studies have suggested that the vector-beamforming can provide excellent performance given a sufficient number of observations (e.g., Sekihara et al., 2004; Brookes et al., 2008; Quraan et al., 2011). However, the existing theoretical analyses have been limited to simple cases, where no more than two sources are allowed in the model and the theoretical sensor covariance is assumed known. In limited data scenarios the estimated sensor covariance may possess considerable variation and thus deteriorate the performance of localization. Empirical studies have also demonstrated that the sampling window and rate are generally required to increase as the number of spatial sensors increases. For example, when using the sample covariance matrix to estimate the sensor covariance matrix, the number of statistically independent data records should be three or more times the number of sensors in order to obtain statistically stable source location estimates (e.g., Rodríguez-Rivera et al., 2006). Consequently, the potential advantages of having a large number of sensors are offset by the requirement for increased sampling window and rate. Therefore, it is important to develop a general framework for users to examine the extent of effects to which the spatial dimension (i.e., the lead field matrix) and the temporal dimension (i.e., the temporal correlations of sensor measurements) of MEG on the accuracy of source localization. Furthermore, most brain activities are conducted by neural networks which consist of multiple sources. For example, in the so-called evoked median-nerve MEG response study, scientists have found the relatively large number of neuronal sources activated in a relatively short period of time by the median-nerve stimulation with typical repetition rates, which challenges covariance-based analysis techniques such as beamformer due to source cancellations (Huang et al., 2004). We need to understand how the accuracy

of localization is affected by source cancellations both theoretically and empirically. In particular, we need to address the fundamental questions of whether the neuronal activity map can reveal the true sources when the number of sensors and the width of the sampling window are large enough and of how much multiple source cancellation effects are reduced by increasing spatial and temporal dimensions of MEG.

The goal of the present study is to demonstrate at both theoretical and empirical levels the behavior of a class of vector-beamforming techniques which includes the standard vector-beamformer as a special example. These beamformers are based on thresholding the sample sensor covariance matrix. By thresholding, we aim at reducing the noise level in the sample sensor covariance. We provide an asymptotic theory on these beamformers when the sensor covariance matrix is consistently estimated and when multiple sources exist. We show that the estimated source power is consistent when multiple sources are asymptotically separable in terms of a lead field distance. We further assess the performance of the proposed procedure by both simulations and real data analyses.

The paper is organized as follows. The details of the proposed procedures are given in Section 2. The asymptotic analysis is provided in Section 3. Other covariance estimator-based beamformers are introduced in Section 4. The simulation studies on these beamformers and an application to face-perception data are conducted in Section 5. The discussion and conclusion are made in Section 6. The proofs of the theorems and corollaries are deferred to Section 7. Throughout the paper, let $||A||$ denote the operator norm of matrix A. For a sequence of matrix A_n, we mean by $A_n = O(1)$ that $||A_n||$ is bounded and by $A_n = o(1)$ that $||A_n|| = o(1)$. Similarly, we define the notations O_p and o_p for a sequence of random matrices A_n. For non-negative matrices A and B, we say $A < B$ if $a^T Aa < a^T Ba$ for any a with $||a|| = 1$. We say that random matrix A_n is asymptotically larger than random matrix B_n in probability if $\min_{||a||=1} a^T(A_n - B_n)a$ is asymptotically bounded below from zero in probability.

2. Methodology

Suppose that the sensor measurements $(\mathbf{Y}(t_j) : 1 \le j \le J)$ are weakly stationary time-courses observed from n sensors. We want to identify a small set of non-null sources that underpin these observations. To this end, we introduce a family of vector-beamformers based on thresholding sensor covariance as follows.

2.1 Thresholding the sensor covariance matrix

The sensor covariance matrix of $\mathbf{Y}(t)$, C can be estimated by the sample covariance matrix

$$\hat{C} = (\hat{c}_{ij}) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{Y}(t_j)\mathbf{Y}(t_j)^T - \bar{\mathbf{Y}}\bar{\mathbf{Y}}^T,$$

where $\bar{\mathbf{Y}}$ is the sample mean of $(\mathbf{Y}(t_j) : 1 \le j \le J)$. It is well-known that the sample covariance estimator can breakdown when the dimension n is large (Bickel and Levina, 2008). In the statistical literature (Bickel and Levina, 2008), various sparse estimation procedures have been proposed to fix the sample covariance, including the following thresholded

estimator:

$$\hat{C}(\tau_{nJ}) = (\hat{c}_{ij}(\tau_{nJ}))$$

with $\hat{c}_{ij}(\tau_{nJ}) = \hat{c}_{ij} I(|\hat{c}_{ij}| \geq \tau_{nJ})$, where τ_{nJ} is a varying constant in n and J.

As with the i.i.d. case (Bickel and Levina, 2008), the above thresholded estimator will be shown to converges to positive definite limit with probability tending to 1 in the Lemma 7.2 in Section 7 below. Although the thresholded estimator has good theoretical properties, it may not be always positive definite when the sample size is finite or when sensors are spatially too close to each other. To tackle the issue, we assume that $\hat{C}(\tau_{nJ})$ has the eigen-decomposition $\hat{C}(\tau_{nJ}) = \sum_{k=1}^{n} \hat{\lambda}_k v_k^T v_k$ and then a positive semidefinite estimator can be obtained by setting these negative eigenvalues to zeros. We further shrinkage the covariance matrix estimator by artificially adding $\epsilon_0 I_n$ to it in our implementation, where we choose ϵ_0 to be a tuning constant which is equal to or slightly larger than the maximum eigenvalue of the noise covariance matrix. We will show in the following sections that adding $\epsilon_0 I_n$ to the thresholded covariance matrix does not affect the consistency of the neuronal activity index.

2.2 Beamforming

As before, let Σ_k denote the covariance matrix of the moment $\mathbf{m}_k(t)$ at the location r_k. Based on the thresholded sensor covariance estimator $\hat{C}(\tau_{nJ})$, we estimate $\Sigma_k, 1 \leq k \leq p$ and create a neuronal activity map in the following two steps.

In the projection step, for $1 \leq k \leq p$, we search for an $n \times 3$ weight matrix $\hat{W}_k$ which attains the minimum trace of $W^T \hat{C}(\tau_{nJ}) W$ subject to $W^T H_k = I_3$. When $\hat{C}(\tau_{nJ})$ is invertible, it follows from Van Veen et al. (1997) that

$$\hat{W}_k = \hat{C}(\tau_{nJ})^{-1} H_k \left[H_k^T \hat{C}(\tau_{nJ})^{-1} H_k \right]^{-1}$$

with the resulting moment covariance matrix and trace estimators

$$\hat{\Sigma}_k = \left[H_k^T \hat{C}^{-1}(\tau_{nJ}) H_k \right]^{-1}, \quad \hat{S}_k = \mathrm{tr} \left\{ \left[H_k^T \hat{C}(\tau_{nJ})^{-1} H_k \right]^{-1} \right\}$$

respectively. In the mapping step, we calculate the so-called neuronal activity index

$$\mathrm{NAI}(r_k) = \hat{S}_k / \Big(\sigma_0^2 \mathrm{tr} \left(\left[H_k^T H_k \right]^{-1} \right) \Big),$$

creating a brain activity map, where σ_0^2 is estimated from baseline data (i.e., called pre-stimulus data in the next subsection). One of the underlying sources can be then estimated by the global peak on the map with the associated latent time-course estimated by projecting the data along the optimal weighting vector. The multiple sources can also be identified by grouping the local peaks on the transverse slices of the brain.

2.3 Choosing the thresholding level

In practice, the MEG imaging is often run on a subject first without stimulus and then with stimulus. This allows us to calculate the sample covariance $\hat{C}$ for the stimulus data

as well as the sample covariance $\hat{C}_0$ for the pre-stimulus data. The latter can provide an estimator of the background noise level. In the next section, we will show that the convergence rate of the thresholded sample covariance is $O(\sqrt{\log(n)/J})$. In light of this, we set $\tau_{nJ} = c_0\hat{\sigma}_0^2\sqrt{\log(n)/J}$ with a tuning constant c_0 and threshold $\hat{C}$ by τ_{nJ}, where $\hat{\sigma}_0^2$ is the minimum diagonal element in $\hat{C}_0$ and c_0 is a tuning constant. Note that, when $c_0 = 0$, the proposed procedure reduces to the standard vector-beamformer implemented in the software FieldTrip (Oostenveld et al., 2010). For each value of c_0, we apply the proposed procedure to the data and calculate the maximum neuronal activity index

$$\text{NAI}_{c_0} = \max\{\text{NAI}(r) : r \text{ is running over the grid}\}. \tag{2.3}$$

In simulations, we will show that $c_0 \in D_0 = \{0, 0.5, 1, 1.5, 2\}$ has covered its useful range. Our simulations also suggests that there is an optimal value of c_0, which depends on several factors including the strengths of signals and source interferences. To exploit these two factors, we choose c_0 in which NAI_{c_0} attains maximum or minimum, resulting in two procedures called **ma** and **mi** respectively. By choosing c_0, the procedure **ma** intends to increase the maximum SNR value, while the procedure **mi** tries to reduce source interferences. The simulation studies in Section 5 suggest that **mi** can perform better than **ma** when sources are correlated.

2.4 Two sets of stimuli

Suppose now that MEG measurements $(\mathbf{Y}^{(1)}(t))$ and $(\mathbf{Y}^{(2)}(t))$ are made under two different sets of stimuli and pre-stimuli with the associated neuronal activity indices denoted by $\text{NAI}^{(1)}(r_k)$ and $\text{NAI}^{(2)}(r_k)$ respectively. The previous strategy for selecting the tuning constant c_0 can be adopted here when we calculate these indices. To identify source locations that respond to the change of stimulus set, we calculate a log-contrast $\log(\text{NAI}^{(1)}(r_k)/\text{NAI}^{(2)}(r_k))$ between the two sets of stimuli at location r_k, $1 \le k \le p$, creating a log-contrast map. The resulting log-contrast map is equivalent to the map based on index ratio $\text{NAI}^{(1)}(r_k)/\text{NAI}^{(2)}(r_k)$, which was often seen in the literature (e.g., Hillebrand et al., 2005). We further take the global peak of the log-contrast as the maximum location estimator for a source location that contributes to the difference between the two sets of MEG measurements.

3. Theory

In this section, we develop a theory on the consistency as well as the convergence rate of the hard thresholding-based beamformer estimator under regularity conditions. In particular, we show that the consistency holds true under regularity conditions if we let the hard threshold $\tau_{nJ} = A\sqrt{\log(n)/J}$ with constant A. This provides a theoretical basis for using the proposed procedures **ma** and **mi**.

Without loss of generality, we assume that the first $q \le p$ moment vectors are of non-zero covariance matrices $\Sigma_k, 1 \le k \le q$, where q is unknown and often much smaller than p in practice. For the simplicity of mathematical derivations, we also assume that Σ_k does

not grow with the number of sensors n. Our task is to identify the unknown true model

$$\mathbf{Y}(t) = \sum_{k=1}^{q} H_k \mathbf{m}_k(t) + \varepsilon(t), \tag{3.4}$$

from the working model (1.1) by using the proposed procedure, where the unknown moments $\mathbf{m}_k(t), 1 \le k \le q$ are of non-zero powers $\mathrm{tr}(\Sigma_k), 1 \le k \le q$. To establish a theory for the proposed procedures, we assume that

(A1): Both the moment vectors $(\mathbf{m}_k(t) : 1 \le k \le q)$ and the white noise process $(\varepsilon(t))$ are stationary with zero means and temporally uncorrelated with each other. Also, $\mathbf{m}_k(t)$ is temporally uncorrelated with $\mathbf{m}_j(t)$ for $k \ne j$.

Under Condition (A1), the sensor covariance matrix of $\mathbf{Y}(t)$, C can be expressed in the form

$$C = \sum_{k=1}^{q} H_k \Sigma_k H_k^T + \sigma_0^2 I_n.$$

As pointed out by Sekihara and Nagarajan (2008, Chapter 9), Condition (A1) is one of fundamental assumptions in the vector-beamforming. However, source activities in the brain are inevitably correlated to some degree, and in strict sense, (A1) cannot be satisfied. The theoretical influence of temporally correlated sources has been investigated by Sekihara and Nagarajan (2008, Chapter 9). The equation (9.3) in Sekihara and Nagarajan (2008, Chapter 9) implies that the influence can be ignored if the partial correlations between sources are close to zeros in order of $o(1/n)$ when the number of sensors n is sufficiently large. Note that although in practice the number of sensors is limited to a few hundreds, we still ideally let n tend to infinity to identify potential spatial factors that affect the performance of a vector-beamformer. In the next section, by using simulations, we will demonstrate that the source correlations can mask some true sources.

To show the consistency of the estimators $\hat{\Sigma}_k$ and S_k, we need more notations and condition as follows. Let H_k denote the lead field matrix at the location r_k. For the simplicity of the technical derivations later, we further assume that the lead field matrices satisfy the condition that for any location r_k, $H_k^T H_k/n \to G$ in terms of the operator norm as n tends infinity, where G is a 3×3 positive definite matrix.

Under the above condition, we can find a positive definite matrix Q_k satisfying that $H_k^T H_k = n Q_k Q_k^T$ and $Q_k^{-1} H_k^T H_k Q_k^{-T} = n I_3$ when n is large enough, where I_3 is an identity matrix. Letting $H_k^* = H_k Q_k^{-T}$, $\mathbf{m}_k^*(t) = Q_k^T \mathbf{m}_k$ and $\Sigma_k^* = Q_k^T \Sigma_k Q_k$, we reparametrize the model (1.1) as follows:

$$\mathbf{Y}(t) = \sum_{k=1}^{p} H_k^* \mathbf{m}_k^* + \varepsilon(t)$$

with the covariance matrix $C = \sum_{k=1}^{p} H_k^* \Sigma_k^* H_k^{*T} + \sigma_0^2 I_n$. Then, under the reparametrized model, the estimators

$$\begin{aligned} \hat{\Sigma}_k^* &= \left[H_k^{*T} \hat{C}(\tau_{nJ})^{-1} H_k^* \right]^{-1} = \left[Q_k^{-1} H_k^T \hat{C}(\tau_{nJ})^{-1} H_k Q_k^{-T} \right]^{-1} = Q_k^T \hat{\Sigma}_k Q_k. \\ \hat{S}_k &= \mathrm{tr}(Q_k^{-T} \hat{\Sigma}_k^* Q_k^{-1}). \end{aligned}$$

Consequently, $\hat{\Sigma}_k^*$ is consistent with Σ_k^* if and only if $\hat{\Sigma}_k$ is consistent with Σ_k. Therefore, without loss of generality, hereinafter we assume that

(A2): $H_k^T H_k = nI_3$ for any location r_k.

We process the remaining analysis in two stages: In the first stage, we develop an asymptotic theory for the proposed vector-beamformers when the sensor covariance matrix C is known. The sensor covariance matrix can be assumed known if the width of the sampling window can be arbitrarily large. In the second stage, we extend the theory to the case where C is estimated by $\hat{C}(\tau_{nJ})$.

3.1 Beamformer analysis when C is known

We begin with introducing some more notations. For any locations r_x and r_y, let H_x and H_y denote their lead field matrices. Define the lead field coherent matrix by $\rho_{xy} = \rho(r_x, r_y) = H_x^T H_y/n$. Note that $\rho_{xy} + \rho_{yx} = I_3 - (H_x - H_y)^T(H_x - H_y)/(2n)$. Therefore, $I_3 - (\rho_{xy} + \rho_{yx})$ indicates how close r_x is to r_y. In general, the partial coherence factor matrices (or called partial correlation matrices) $a_{yx|k}$, $1 \le k \le q$ are defined iteratively by the so-called sweep operation (Goodnight, 1979) as follows:

$$\begin{aligned} a_{yx|1} &= \sigma_0^{-2}\rho(r_y, r_1, r_x) = \sigma_0^{-2}\left(\rho(r_y, r_x) - \rho(r_y, r_1)\rho(r_1, r_x)\right), \\ a_{yx|(k+1)} &= a_{yx|k} - a_{y(k+1)|k}\left[a_{(k+1)(k+1)|k}\right]^{-1} a_{(k+1)x|k}, \quad 1 \le k \le q-1. \end{aligned}$$

For example, we have

$$\begin{aligned} \sigma_0^2 a_{yx|1} &= \rho_{yx} - \rho_{y1}\rho_{1x}, \quad \sigma_0^2 a_{22|1} = I_3 - \rho_{12}^T\rho_{12}, \\ \sigma_0^2 a_{33|2} &= I_3 - \rho_{13}^T\rho_{13} - \left(\rho_{23} - \rho_{12}^T\rho_{13}\right)^T\left[I_3 - \rho_{12}^T\rho_{12}\right]^{-1}\left(\rho_{23} - \rho_{12}^T\rho_{13}\right). \end{aligned}$$

Note that $\sigma_0^2 a_{(k+1)(k+1)|k}$ gauges the partial variability of r_{k+1} given the previous $r_k's$ while $\sigma_0^2 a_{yx|(k+1)}$ shows the partial coherence between r_x and r_y given $\{r_1, ..., r_{k+1}\}$. We expect that $a_{yx|(k+1)}$ will be small if r_y and r_x are spatially far away from each other. We define $b_{yx|k}$, $1 \le k \le q$, by letting $b_{yx|1} = \rho_{y1}\Sigma_1^{-1}\rho_{1x}$ and

$$\begin{aligned} b_{yx|k} &= b_{yx|(k-1)} - b_{yk|(k-1)}\left[a_{kk|(k-1)}\right]^{-1} a_{kx|(k-1)} - a_{yk|(k-1)}\left[a_{kk|(k-1)}\right]^{-1} b_{kx|(k-1)} \\ &\quad + a_{yk|(k-1)}\left[a_{kk|(k-1)}\right]^{-1}\left\{\Sigma_k^{-1} + b_{kk|(k-1)}\right\}\left[a_{kk|(k-1)}\right]^{-1} a_{kx|(k-1)}. \end{aligned}$$

We also define $c_{jj|k}$, $1 \le j \le k \le q$ by

$$c_{jj|k} = \begin{cases} -\Sigma_k^{-1}\left[a_{kk|(k-1)}\right]^{-1}\Sigma_k^{-1}, & j = k \\ c_{jj|(k-1)} - b_{jk|(k-1)}\left[a_{kk|(k-1)}\right]^{-1} b_{jk|(k-1)}^T, & 1 \le j \le k-1. \end{cases}$$

Let $a_{nq} = n\min_{1\le k\le q-1}||a_{(k+1)(k+1)|k}||$, and let $k_m = 0$ if $a_{nq} \to \infty$ and $k_m = \min\{1 \le k \le q-1 : n||a_{(k+1)(k+1)|k}|| = O(1)\}$ if $a_{nq} = O(1)$. Let $d_{x|q} = \max_{2\le k\le q}||a_{kx|(k-1)}a_{kk|(k-1)}^{-1}||$, which measures the maximum absolute partial correlation among q sources by using their lead field matrix. As the lead field matrix measures the unit outputs of sources recorded by sensors, the maximum absolute partial correlation may increase when the number of sensors n increases. In the following theorem, for any location r_x of interest, the condition that

$d_{x|q} = O(1)$ (i.e., the maximum absolute partial correlation will be bounded) is imposed on the lead field matrix. The condition is used to ensure the coherence stability for the grid approximation to the lead field. Our numerical experience suggests that the condition roughly holds when the underlying sources are asymptotically not close to each other. See the discussion in Section 7. The following theorem shows when the source covariance estimator is consistent and when it is not.

Theorem 1 *Under Conditions (A1)∼(A2) and C is known, we have:*

(1) If $a_{nq} = O(1)$ and $\max_{1\le k\le q} d_{k|q} = O(1)$, then the estimated source covariance at r_{k_m+1} $\left[{H_{k_m+1}}^T C^{-1} H_{k_m+1}\right]^{-1}$ is asymptotically larger than Σ_{k_m+1}.

(2) If $a_{nq} \to \infty$, then for $1 \le j \le q$, the estimated source covariance at r_j admits

$$\left[H_j^T C^{-1} H_j\right]^{-1} = \Sigma_j + \frac{1}{n}\Sigma_j c_{jj|q}\Sigma_j + O(a_{nq}^{-2}),$$

provided $\max_{1\le k\le q} d_{k|q} = O(1)$, where $||\Sigma_j c_{jj|q}\Sigma_j/n|| = O(a_{nq}^{-1})$ as $n \to \infty$.

(3) If $a_{nq} \to \infty$, then for $r_x \notin \{r_1, ..., r_q\}$, the estimated source covariance at r_x admits

$$\left[H_x^T C^{-1} H_x\right]^{-1} = \frac{1}{n}a_{xx|q}^{-1} - \frac{1}{n^2}a_{xx|q}^{-1} b_{xx|q} a_{xx|q}^{-1} + O(a_{nq}^{-3}),$$

provided $\max_{1\le j\le q} d_{j|q} = O(1)$, $||n a_{xx|q}|| \to \infty$, and $d_{x|q} = O(1)$ as n tends to infinity, where $b_{xx|q} = O(1)$ as $n \to \infty$.

The following lemma shows when the source power estimator is consistent.

Corollary 2 *Under Condition (A1)∼(A2), we have:*

(1) If $a_{nq} = O(1)$ and $\max_{1\le k\le q} d_{k|q} = O(1)$, then the estimated source power at r_{k_m+1} $tr\left(\left[{H_{k_m+1}}^T C^{-1} H_{k_m+1}\right]^{-1}\right)$ is asymptotically larger than $tr\,(\Sigma_{k_m+1})$.

(2) If $a_{nq} \to \infty$, then for $1 \le j \le q$, the estimated source power at r_j admits

$$tr\left(\left[H_j^T C^{-1} H_j\right]^{-1}\right) = tr(\Sigma_j) + \frac{1}{n}tr(\Sigma_j c_{jj|q}\Sigma_j) + O(a_{nq}^{-2}),$$

provided $\max_{1\le k\le q} d_{k|q} = O(1)$, where $||\Sigma_j c_{jj|q}\Sigma_j/n|| = O(a_{nq}^{-1})$ as $n \to \infty$.

(3) If $a_{nq} \to \infty$, then for $r_x \notin \{r_1, ..., r_q\}$, the estimated source power at r_x admits

$$tr\left(\left[H_x^T C^{-1} H_x\right]^{-1}\right) = \frac{1}{n}tr(a_{xx|q}^{-1}) - \frac{1}{n^2}tr(a_{xx|q}^{-1} b_{xx|q} a_{xx|q}^{-1}) + O(a_{nq}^{-3}),$$

provided $\max_{1\le j\le q} d_{j|q} = O(1)$, $||n a_{xx|q}|| \to \infty$, and $d_{x|q} = O(1)$ as n tends to infinity, where $b_{xx|q} = O(1)$ as $n \to \infty$.

Remark 3 *It follows from the definition that $c_{jj|q}$ is proportional to σ_0^2, which implies the convergence rate of the neuronal activity index is of order $O(\sigma_0^2/(\sigma_0^2 a_{nq}))$, where $\sigma_0^2 a_{nq}$ is independent of σ_0^2. Therefore, the effect of adding $\epsilon_0 I_n$ to C on the above convergence rate is increasing or decreasing the rate by the amount of $O(\epsilon_0/((\sigma_0^2+\epsilon_0)a_{nq}))$. In particular, adding $\epsilon_0 I_n$ to C does not affect the consistency of the neuronal activity index if a_{nq} tends infinity.*

Remark 4 *From the proof in Section 7, we can see that if we relax the coherence stability condition $\max_{1\le k\le q} d_{k|q} = O(1)$ to $\max_{1\le k\le q} d_{k|q} = O(\log(n))$, then the convergence rates in the theorem will be reduced by a factor of $\log(n)$.*

Remark 5 *If there are MEG measurements made under two different sets of stimuli and pre-stimuli, we let $C^{(1)} = \sum_{k=1}^p H_k^T \Sigma_k^{(1)} H_k + \sigma_{01}^2 I_n$ and $C^{(2)} = \sum_{k=1}^p H_k^T \Sigma_k^{(2)} H_k + \sigma_{02}^2 I_n$ be the corresponding sensor covariance matrices. We perform the proposed beamformers on $C^{(1)}$ and $C^{(2)}$ respectively. Then, under certain conditions, Theorem 1 can be extended to this setting. When r_k is a source location for both sets of stimuli, the log-contrast tends to the true one as $n \to \infty$; when r_k is a source for stimulus set 1 but not for stimulus set 2, the log-contrast tends to infinite; when r_k is a source location for stimulus set 2 but not for stimulus set 1, the log-contrast tends to $-\infty$; when r_j is neither a source for stimulus set 1 nor a source for stimulus 2, the log-contrast tends to a finite value depending on the associated values of $a_{xx|q}$. The details are omitted here.*

3.2 Beamformer analysis when C is estimated

We now estimate the sensor covariance matrix by using the sensor observations over J time instants. Following Bickel and Levina (2008) and Fan et al. (2011), we establish the asymptotic theory for the resulting beamformer estimators when both n and J are tending to infinity.

In addition to Conditions (A1) and (A2), we need the following two conditions for conducting the asymptotic analysis above. The first one is imposed to regularize the tail behavior of the sensor processes.

(A3): There exist positive constants κ_1 and τ_1 such that for any $u > 0$ and all t,

$$\max_{1\le i\le n} P(||Y_i(t)|| > u) \le \exp(1 - \tau_1 u^{\kappa_1})$$

and $\max_{1\le i\le n} E||Y_i(t)||^2 < +\infty$, where the noise covariance matrix is $\sigma_0^2 I_n$ and $||\cdot||$ is the L_2 norm.

Note that Condition (A3) holds if $\mathbf{Y}(t)$ is a multivariate normal.

In the second additional condition, we assume that the sensor processes are strong mixing. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_k^\infty$ denote the σ-algebras generated by $\{\mathbf{Y}(t) : -\infty \le t \le 0\}$ and $\{\mathbf{Y}(t) : t \ge k\}$ respectively. Define the mixing coefficient

$$\alpha(k) = \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_k^\infty} |P(A)P(B) - P(AB)|.$$

The mixing coefficient $\alpha(k)$ quantifies the degree of the temporal dependence of the process $\{\mathbf{Y}(t)\}$ at lag k. We assume that $\alpha(k)$ is decreasing exponentially fast as lag k is increasing.

(A4): There exist positive constants κ_2 and τ_2 such that

$$\alpha(k) \leq \exp(-\tau_2 k^{\kappa_2}).$$

Condition (A4) is a commonly used assumption for studying asymptotic behavior of time series.

For a constant A, let $\tau_{nJ} = A\sqrt{\log(n)/J}$. As before, let $\bar{Y}_i$ be the sample mean of the i-th sensor and

$$\hat{c}_{ik} = \frac{1}{J}\sum_{j=1}^{J}(Y_i(t_j) - \bar{Y}_i)(Y_k(t_j) - \bar{Y}_k), \quad \hat{C}(\tau_{nJ}) = (\hat{c}_{ik} I(\hat{c}_{ik} \geq \tau_{nJ})),$$

where $I(\cdot)$ is the indicator.

We are now in position to generalize Theorem 1 to the case where the sensor covariance is estimated by the thresholded covariance estimator.

Theorem 6 *Under Conditions (A1)∼(A4) and assuming that* $n^2\sqrt{\log(n)/J} = o(1)$ *as* n *and* J *tend to infinity, we have:*

(1) If $a_{nq} = O(1)$ *and* $\max_{1\leq k\leq q} d_{k|q} = O(1)$*, then as* n *and* J *tend to infinity, the estimated source covariance at* r_{k_m+1} $\hat{\Sigma}_{k_m+1}$ *is asymptotically larger than* Σ_{k_m+1} *in probability.*

(2) If $a_{nq} \to \infty$*, then as* n *and* J *tend to infinity, for* $1 \leq j \leq q$*, the estimated source covariance at* r_j *admits*

$$\hat{\Sigma}_j = \Sigma_j + \frac{1}{n}\Sigma_j c_{jj|q}\Sigma_j + O_p(a_{nq}^{-2} + n^2\sqrt{\log(n)/J}),$$

provided $\max_{1\leq k\leq q} d_{k|q} = O(1)$*, where* $||\Sigma_j c_{jj|q}\Sigma_j/n|| = O(a_{nq}^{-1})$ *as* $n \to \infty$*.*

(3) If $a_{nq} \to \infty$*, then as* n *and* J *tend to infinity, for* $r_x \notin \{r_1, ..., r_q\}$*, the estimated source covariance at* r_x *admits*

$$\hat{\Sigma}_x \quad = \quad \frac{1}{n}a_{xx|q}^{-1} - \frac{1}{n^2}a_{xx|q}^{-1}b_{xx|q}a_{xx|q}^{-1} + O(a_{nq}^{-3} + n^2\sqrt{\log(n)/J}),$$

provided $\max_{1\leq j\leq q} d_{j|q} = O(1)$*,* $||na_{xx|q}|| \to \infty$*, and* $d_{x|q} = O(1)$ *as* n *tends to infinity, where* $b_{xx|q} = O(1)$ *as* $n \to \infty$*.*

Corollary 7 *Under Conditions (A1)∼(A4) and assuming that* $n^2\sqrt{\log(n)/J} = o(1)$ *as* n *and* J *tend to infinity, we have:*

(1) If $a_{nq} = O(1)$*,* $\max_{1\leq k\leq q} d_{k|q} = O(1)$*, as* n *and* J *tend to infinity, the estimated source power at* r_{k_m+1}*,* $\hat{S}_{k_m+1}$ *is asymptotically larger than* $tr(\Sigma_{k_m+1})$*.*

(2) If $a_{nq} \to \infty$*, then as* n *and* J *tend to infinity, for* $1 \leq j \leq q$*, the estimated source power at* r_j *admits*

$$\hat{S}_j = tr(\Sigma_j) + \frac{1}{n}tr(\Sigma_j c_{jj|q}\Sigma_j) + O(a_{nq}^{-2} + n^2\sqrt{\log(n)/J}),$$

provided $\max_{1\leq k\leq q} d_{k|q} = O(1)$*, where* $||\Sigma_j c_{jj|q}\Sigma_j/n|| = O(a_{nq}^{-1})$ *as* $n \to \infty$*.*

(3) If $a_{nq} \to \infty$, then as n and J tend to infinity, for $r_x \notin \{r_1, ..., r_q\}$, the estimated source power at r_x admits

$$\hat{S}_x = \frac{1}{n} tr(a^{-1}_{xx|q}) - \frac{1}{n^2} tr(a^{-1}_{xx|q} b_{xx|q} a^{-1}_{xx|q}) + O(a^{-3}_{nq} + n^2\sqrt{\log(n)/J}),$$

provided $\max_{1\le j\le q} d_{j|q} = O(1)$, $||na_{xx|q}|| \to \infty$, *and* $d_{x|q} = O(1)$ *as n tends to infinity, where* $b_{xx|q} = O(1)$ *as* $n \to \infty$.

Remark 8 *Theorem 6 indicates the convergence rate of the vector-beamformer estimation is much slower than the empirical rate suggested by Rodríguez-Rivera et al. (2006). However, the result is in agreement with an empirical result of Brookes et al. (2008). In fact, using their heuristic arguments, we can show that the error of the power estimation at location r_x is determined by the factor $H_x(\hat{C}(\tau_{nJ})^{-1} - C^{-1})H_x$, which has a rate of $n^2\sqrt{\log(n)/J}$.*

Theorem 6 can be also extended to the scenarios where MEG data are obtained under two different sets of stimuli.

Remark 9 *From the proof of Theorem 6, we can see that the thresholded covariance is still consistent with the true C even when the underlying sources are correlated.*

4. Other covariance estimator-based beamformers

There are various ways to estimate the sensor covariance matrix. Each can be used to construct a beamformer. These covariance estimators can be roughly divided into two categories, namely global shrinkage-based methods and elementwise thresholding-based methods. In shrinkage-based settings, the sample covariance is shrinking toward a target structure (for example, a diagonal matrix). The so-called optimal shrinkage estimator belongs to this category (Ledoit and Wolf, 2004). In thresholding-based settings, an elementwise thresholding is applied to the sample covariance estimator. Examples of these approaches include hard thresholding, generalized thresholding and adaptive thresholding (Bickel and Levina, 2008; Rothman et al., 2009; Cai and Liu, 2011). Here, we focus on the following three methods recommended by the above authors.

The optimal shrinkage covariance matrix is defined by

$$\hat{C}_{opt} = \frac{b_n^2}{d_n^2}\mu_n I_n + \frac{d_n^2 - b_n^2}{d_n^2}\hat{C},$$

where

$$\mu_n = \left\langle \hat{C}, I_n \right\rangle, \quad d_n^2 = \left\langle \hat{C} - \mu_n I_n, \hat{C} - \mu_n I_n \right\rangle,$$

$$\bar{b}_n^2 = \frac{1}{J^2}\sum_{j=1}^{J} \left\langle \mathbf{Y}_j\mathbf{Y}_j^T - \hat{C}, \mathbf{Y}_j\mathbf{Y}_j^T - \hat{C} \right\rangle, \quad b_n^2 = \min(\bar{b}_n^2, d_n^2),$$

and the operator $< A, B >= \mathrm{tr}(AB^T)/n$ for any $n \times n$ matrices A and B. The idea behind the above estimator is to find the optimal weighted average of the sample covariance matrix $\hat{C}$ and the identity matrix via minimizing the expected squared loss. Under certain

conditions $\hat{C}_{opt}$ converges to the true covariance C as n tends infinity, implying that $\hat{C}_{opt}$ can be degenerate if C is degenerate (Ledoit and Wolf, 2004). As before, we tackle the issue by adding $\epsilon_0 I_n$ to $\hat{C}_{opt}$, where ϵ_0 is determined by the maximum eigenvalue of the pre-stimulus sample covariance matrix. The beamformer based on the above covariance estimator is denoted as **sh**.

A family of generalized thresholding-based covariance estimators indexed by tuning constants $c_0 \geq 0$ and $\delta_0 > 0$ can be defined by replacing the hard thresholding in Subsection 2.1 with the generalized thresholding, i.e.,

$$\hat{C}_g = (g(\hat{c}_{ij}))$$

with $g(\hat{c}_{ij}) = \hat{c}_{ij}(1 - (\tau_{nJ}/|\hat{c}_{ij}|)^{\delta_0})$, where $\tau_{nJ} = c_0\hat{\sigma}_0^2\sqrt{\log(n)/J}$ and $\hat{\sigma}_0^2$ is estimated from a baseline sample. Following the suggestion of Rothman et al. (2009), we choose $\delta_0 = 4$. The same maximum/minimum strategy as in Subsection 2.3 can be adapted to choose the tuning constant c_0 when we use the above estimator to construct a beamformer. The corresponding beamformers are denoted by **gmax** and **gmin** respectively.

Similarly, an adaptive thresholding estimator can be introduced by replacing the above τ_{nJ} in the g function by $\lambda_{ij} = 2\sqrt{\hat{\theta}_{ij}\log(n)/J}$, where $\hat{\theta}_{ij}$ is the estimated variance of the (i,j)-th entry $\hat{c}_{ij}$ and is defined by

$$\hat{\theta}_{ij} = \frac{1}{J}\sum_{k=1}^{J}[(Y_{ik} - \bar{Y}_i)(Y_{jk} - \bar{Y}_j) - \hat{c}_{ij}]^2$$

and $\bar{Y}_i$ and $\bar{Y}_j$ are the sample means of the i-th and the j-th sensors. See Cai and Liu (2011). The corresponding beamformer is denoted by **adp**.

5. Numerical results

In this section, we compare the proposed procedures to the standard vector-beamformer (with the tuning $c_0 = 0$) and to the other covariance estimator-based beamformers in terms of localization bias by simulation studies and real data analyses. Here, for any estimator $\hat{r}$ of a source location r, the localization bias $|\hat{r} - r|$ is the L_1 distance between $\hat{r}$ and r. The spatial correlation $\rho_{\max}$ between locations r_1 and r_2 is measured by the maximum correlation between the projected lead field vectors at r_1 and r_2:

$$\rho_{\max}(r_1, r_2) = \left\{ \max_{||\eta_1||=1, ||\eta_2||=1} \frac{(l(r_1)\eta_1)^T l(r_1)\eta_1}{||l(r_1)\eta_1|| \cdot |l(r_2)\eta_2||} \right\}.$$

By simulations, we attempted to answer the following questions:

- Has the vector-beamformer been improved by using the thresholded covariance estimator?
- To what extent will the performance of the proposed beamformer procedure deteriorate by source interferences (or source cancellations) and source spatial correlations?
- Can the proposed beamformers **ma** and **mi** be superior to the other covariance estimator-based beamformers?

5.1 Simulated data

We started with specifying the following two head models (Sarvas, 1987). The simple head model that uses a homogeneous sphere in simulating the magnetic fields emanating from current electric dipole neuronal activity possesses the advantage that the lead field matrix can be calculated analytically. However, with more realistic head models, the numerical approximations such as a finite element method have to be used when we calculate the lead field matrix. Here, we considered both of them: the simple one is a spherical volume conductor with 10cm radius from the origin and with 91 sensors, created by using the software FieldTrip (Oostenveld et al., 2010), and the realistic one is a single shell head model by using the magnetic resonance imaging (MRI) scan of a human brain provided by Henson et al. (2011). We then discretized the inside brain space into a 3D-grid of resolution 1 cm. This yielded a grid with 2222 points for the simple model and 1487 points for the realistic model. The grids was further sliced into 10 and 14 transverse layers along the z-axis of the brain respectively. We put two non-null sources at r_1 and r_2 or three sources at r_1, r_2 and r_3 respectively, where two sources $\{r_1, r_2\}$ are equal to $\{(3,-1,4)^T, (-5,2,6)^T\}$ cm or $\{(-5,5,6)^T, (-6,-2,5)^T\}$ cm, and three sources $\{r_1, r_2, r_3\}$ are equal to $\{(3,-1,4)^T, (-5,2,6)^T, (5,5,6)^T\}$ in the Subject Coordinate System (SCS/CTF). Note that the second set of source locations was obtained in our real data analyses which will be presented later. These sources were located in the region of the parietal and occipital lobes, where visual, auditory and touch information is processed. We considered two types of sources in the brain: evoked responses that are phase-locked to the stimulus and induced responses that are not. The induced responses often have oscillatory patterns. Combining these sources with the two head models, we had the following four scenarios:

- **Scenario 1**: For the simple head model, we put two oscillatory sources at locations $r_1 = (3,-1,4)^T$ and $r_2 = (-5,2,6)^T$ with time-courses

$$\mathbf{m}_k(t) = \eta_k \cos(20t\pi), \quad k = 1, 2,$$

  respectively, where $\eta_1 = (10,1,1)^T$ and $\eta_2 = (8,0,0)^T$. We considered two values of the signal-to-noise-ratio (SNR): 0.04 and $1/0.64 = 1.5625$.

- **Scenario 2**: For the simple head model, we put the above oscillatory sources at locations $r_1 = (-5,5,6)^T$ and $r_2 = (-6,-2,5)^T$. We also considered two values of the SNR: 0.04 and $1/0.64 = 1.5625$.

- **Scenario 3**: For the realistic head model, we put the following evoked response sources at locations $r_1 = (3,-1,4)^T$ and $r_2 = (-5,2,6)^T$ with moments (i.e., time-courses)

$$\mathbf{m}_k(t) = \alpha_k \exp(-(t-\tau_{k1})^2/\omega_k^2) \sin(f_k 2\pi(t-\tau_{k2})), k = 1, 2,$$

  respectively, where $\alpha_1 = (5,0,0)^T$, $\alpha_2 = (20,0,0)^T$, $\tau_{11} = 0.239$, $\tau_{12} = 0.139$, $\tau_{21} = 0.199$, $\tau_{22} = 0.139$, $f_1 = 4.75$, $f_2 = 6.25$, and $\omega_1 = \omega_2 = 0.067$. We considered three values of the SNR: $1/0.35^2 = 8.16, 1/0.4^2 = 6.25, 1/0.5^2 = 4$.

- **Scenario 4**: For the realistic head model, we put the following evoked response sources at locations $r_1 = (-5, 5, 6)^T$ and $r_2 = (-6, -2, 5)^T$ with moments (i.e., time-courses)

$$\mathbf{m}_k(t) = \alpha_k \exp(-(t - \tau_{k1})^2/\omega_k^2) \sin(f_k 2\pi(t - \tau_{k2})), k = 1, 2,$$

  respectively, where $\alpha_1 = (2, 0, 0)^T$, $\alpha_2 = (18, 0, 0)^T$, $\tau_{11} = 0.439$, $\tau_{12} = 0.139$, $\tau_{21} = 0.399$, $\tau_{22} = 0.139$, $f_1 = 6$, $f_2 = 9$, and $\omega_1 = \omega_2 = 2$. We considered three values of the SNR: $1/0.7^2 = 2.04, 1/0.76^2 = 1.73, 1/0.78^2 = 1.64$.

- **Scenario 5**: We added another evoked response source at location $r_3 = (5, 5, 6)^T$ to the model in Scenario 3 with moment

$$\mathbf{m}_3(t) = \alpha_3 \exp(-(t - \tau_{31})^2/\omega_3^2) \sin(f_3 2\pi(t - \tau_{32})),$$

  where $\alpha_3 = (2.5, 0.25, 0.25)$, $\tau_{31} = 0.1$, $\tau_{32} = 0.139$, $f_3 = 1.25$, and $w_3 = 0.067$. The three source locations are highly spatially correlated with the pairwise spatial correlations $\rho(r_1, r_2) = 0.7289$, $\rho(r_1, r_3) = 0.7935$, and $\rho(r_2, r_3) = 0.5924$. We considered the same SNR values as in Scenario 3.

The pair sources $\mathbf{m}_k(t), k = 1, 2$ for the first four scenarios and the treble sources $\mathbf{m}_k(t), k = 1, 2, 3$ for Scenario 5 are plotted respectively in Figure 1. By Scenarios 1 and 2, we compared the proposed procedure to the standard vector-beamformer (with $c_0 = 0$) and to the other estimator-based beamformer, when there existed two highly correlated oscillatory sources (they have the same frequency and phase, but with slightly different amplitudes). By Scenarios 3 and 4, we tested these beamformers when there existed two unbalanced evoked response (or slightly dumped-oscillatory) sources. By Scenario 5, we assessed these beamformers when there were three spatially correlated source locations. In each scenario, with time-window width 1 and sample rate J, we sampled 30 data sets of $\mathbf{Y}(t)$ from the model

$$\mathbf{Y}(t) = \sum_{k=1}^{p} H_k \mathbf{m}_k(t) + \varepsilon(t), \tag{5.5}$$

where in Scenarios 1∼4, $\mathbf{m}_k(t)$, $k = 1, 2$ are non-null time-courses at the two locations and $\mathbf{m}_k(t), 3 \le k \le p$ are null time-courses at other grid points, while in Scenario 5, $\mathbf{m}_k(t)$, $k = 1, 2, 3$ are non-null time-courses at the three locations and $\mathbf{m}_k(t), 4 \le k \le p$ are null time-courses at other grid points. As before, $\{\varepsilon(t)\}$ is a white noise process with noise level σ_0^2. We considered various combinations of $(n, p) = (91, 2222)$ and $(102, 1487)$, and $J = 500, 1000, 2000$, and 3000. Note that p is substantially larger than n and that the sources are sparse in the sense that there are only two or three non-null sources among p candidates.

We first applied the proposed procedures **ma**, **mi** and **sh** to each data set. We calculated the maximum indices over the grids and the L_1-biases of the maximum location estimates to two sources respectively. For each combination of (n, p, J) and the SNR, we then summarized these values in the form of a box-whisker plot as in Figures 2, 3, 4, and 5 corresponding to Scenarios 1, 2, 3, and 4 respectively. The results demonstrate that the proposed hard thresholding-based procedure **mi** can outperform both the conventional vector-beamformer

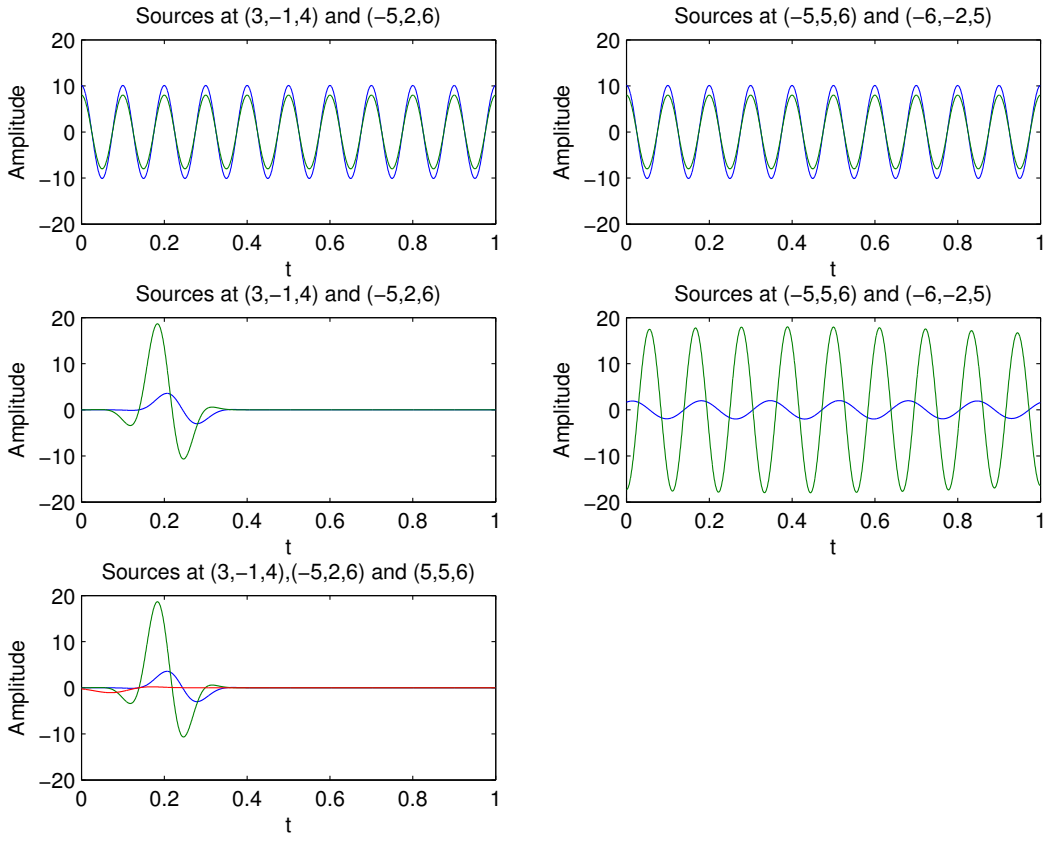

Figure 1: The amplitude plots of $\mathbf{m}_k(t), k = 1, 2$ for Scenarios 1 to 4 and the amplitude plots of $\mathbf{m}_k(t), k = 1, 2, 3$ for Scenario 5. In these plots, the blue, green and red colored curves are corresponding to the amplitudes of $\mathbf{m}_k(t)$, $k = 1, 2, 3$ respectively.

and the procedures **ma** and **sh** in all four scenarios, in particular when the SNR is low. We note that in several cases, the localization bias and the maximum index were degenerate to a single value with some outliers, indicating that random variations have not changed the global peak location although they have effects on local peaks on the map. The simulations also suggest that the proposed procedure may be unable to detect evoked response sources of low SNR values. The local peak box-whisker plots in these figures reveal that all the local peaks on the transverse slices are not close to the source location r_1, implying that the source at r_1 has been masked on the neuronal activity index-based map even when two sources have a similar power level. This may be due to source cancellations as the lead field vectors at these two locations were correlated and the sensor positions might favor the detection of r_2. Finally, we note that the results are robust to the choice of J in the sense that increasing sampling frequency has only slightly reduced both the mean and standard error of localization bias.

To compare the procedures **ma**, **mi** and **sh** with the procedures **gma**, **gmi** and **adp** based on the generalized and adaptive thresholding, we again generated 30 data sets from model (5.5) for each of the above four scenarios and for each combination of $(n, p) = (91, 2222)$ and $(102, 1487)$, and $J = 500, 1000, 2000$, and 3000. We applied these procedures to each data set and calculated their localization biases respectively. As before, we displayed these biases by multiple box-whisker plots in Figures 6, 7 and 8. From these figures, we can see a dramatic improvement in localization performance of the hard thresholding-based procedure **mi** over the other procedures in Scenarios 1 and 2 and a slightly better or similar performance to **ma**, **gma**, **gmi**, **adp** and **sh** in Scenarios 3 and 4. This is striking because the existing studies have already shown that the soft (or generalized) and adaptive thresholding-based covariance estimators can improve the hard thresholding-based covariance estimator in terms of estimation loss. The potential explanations for this phenomena are as follows: (1) The procedure **adp** may lose efficiency by not using the pre-stimulus data. (2) The existing covariance estimators were aimed to improve the estimation accuracy by reducing the estimation loss (the distance between the estimator and the true covariance matrix) or by increasing the sensitivity and specificity in recovering sparse entries in the true covariance matrix (Rothman et al., 2009; Cai and Liu, 2011). Unfortunately, the sparsity in MEG means a sparse signal distribution, which is quite different from the entry sparsity of the sensor covariance matrix. Therefore, these estimators may be not efficient for improving the accuracy of the beamformer estimation which is related to the signal sparsity. In fact, our simulation experience suggests that besides the covariance estimation, there are other factors that can affect the performance of a beamformer such as the lead field matrix and the spatial distribution of signals in the brain. Therefore, the covariance estimator with a smaller estimation loss may not give rise to a beamformer with a lower localization bias.

To assess the performances of the six procedures **ma**, **mi**, **gma**, **gmi**, **adp** and **sh** when there are more than two spatially correlated sources, we applied these procedures to the 30 data sets generated for Scenario 5. We calculated the average localization bias for each procedure and presented them in Figure 9. It can be seen from these plots that like in two-source scenarios, **mi** can have superior performance over the other procedures. However, compared the above result to those in Scenario 3, we can see that the source cancellation from r_3 has increased the average localization bias from zero to the value of three.

Note that although Theorem 2 suggests that in general the localization bias will be reduced as the sampling rate increases, it does not implies the localization bias is a monotone function of the sampling rate (or the number of time instances). In fact, from row 4 in Figure 2 and row one in Figure 9, it can be seen that the localization bias when $J = 500$ is smaller than when $J = 1000, 2000$ and 3000. A potential explanation is that in finite cases a higher sampling rate may cause a higher amount of leakage of background noises (in a neighborhood of the target location) into the neuronal activity index calculation.

Finally, we notice that we also carried out simulations with the soft thresholding ($\delta_0 = 1$). The result is very similar to the case with $\delta_0 = 4$. For reasons of space, we do not report it here.

5.2 Face-perception data

We applied the proposed methodology to human MEG data acquired in five sessions by Wakeman and Henson (Henson et al., 2011). In each session, 96 face trials and 50 scrambled face trials were performed on a healthy young adult subject. Each trial started with a central fixation cross (presented for a random duration of 400 to 600 ms), followed by a face or scrambled face (presented for a random duration of 800 to 1000 ms), and followed by a central circle for 1700 ms. The subject used either his/her left or right index finger to report whether he/she thought the stimulus was symmetrical or asymmetrical vertically through its center. The data were collected with a Neuromag VectorView system, containing a magnetometer and two orthogonal, planar gradiometers located at each of 102 positions within a hemispherical array situated in a light, magnetically shielded room. The sampling rate was 1100Hz. We focused our analysis on localizing non-null source positions, where neuronal activity increases for the face stimuli relative to the scrambled face stimuli.

For this purpose, we normalized the subject's MRI scan to a MRI template by using the FieldTrip, on which a grid CTF system of 1 cm resolution was created with 1487 points. For each session, we applied the neuroimaging software SPM8 to read and preprocess the recorded data, and to epoch and average the data generated from the face stimulus trials and the scrambled face stimulus trials respectively. This gives rise to five 306×771 data matrices: the first 220 columns for 200ms pre-stimuli and the later 551 columns for the stimuli. For each session, we calculated the sample covariance $\hat{C}$ and noise covariance $\hat{C}_0$ by using the stimulus data and the pre-stimulus data respectively. We estimated the baseline noise level by $\hat{\sigma}_0^2$, the minimum diagonal element in $\hat{C}_0$. We applied the beamforming procedures **ma**, **mi**, **gma**, **gmi**, **adp**, and **sh** to the face data set and the scrambled face data set respectively, obtaining the log-contrasts at each grid point. Here, if there exist the negative eigenvalues of the covariance estimators (used in **ma**, **mi**, **gma**, **gmi**, **adp** and **sh**), we set them to zeros and added ϵ_0 to them to make the resulting covariance estimators positive definite, where ϵ_0 was determined by the maximum eigenvalue of the noise matrix $\hat{C}_0$. For each procedure, we interpolated and overlaid its log-contrasts on the structural MRI of the subject, obtaining its index map. There were no visible differences among the maps derived from **ma**, **mi**, **gma**, **gmi** and **sh**. The map derived from the **adp** slightly differed from the rest. So, we reported only the **mi**-based and **adp**-based maps below.

For each session, we first identified the global peak location from each map, followed by slicing the maps through their global peak locations as shown in Figure 10. For sessions $1 \sim 4$, the global peaks derived from the mi and adp were the same, which were located at $(-4, 3, 8)$cm, $(-1, -6, 8)$cm, $(-6, -2, 5)$cm, and $(-4, -4, 6)$cm respectively. However, for session 5, the global peaks derived from the **mi** and the **adp** were located at two slightly different positions, $(-4, -4, 6)$cm and $(-7, -3, 6)$cm. We then projected the data along the associated optimal weight directions, obtaining estimated time-courses at these global peaks. For reasons of space, we presented only these time-courses derived from the procedure **mi**. See Figure 12. Finally, we made 20 transverse slices along the z-axis to identify the local peaks. There were some subtle differences between the **mi**-based and the **adp**-based local peaks. For example, in session 1, the **mi**-based local peaks were located at $(1, 5, 2)$ cm, $(0, -1, 11)$ cm, $(3, 2, 10)$ cm, $(3, 4, 9)$ cm, $(-5, -3, 3)$ cm, $(-4, -3, 4)$ cm, $(-2, 1, 1)$ cm,$(-4, -3, -1)$ cm,$(-2, 1, 0)$ cm,$(-4, -5, 5)$ cm,$(-4, 2, 6)$ cm,$(-5, 3, 7)$ cm and

$(-4, 3, 8)$ cm, whereas the **adp**-based local peaks were located at $(3, 2, 2)$cm, $(0, -1, 11)$cm, $(-4, 3, 9)$cm, $(-6, -2, 1)$cm, $(-4, -3, 4)$cm, $(2, 3, 10)$cm, $(-4, -3, -1)$ cm, $(-1, 1, 0)$ cm, $(-3, 6, 3)$ cm, $(-4, -4, 5)$ cm, $(-4, 2, 6)$ cm, $(-5, 3, 7)$ cm, and $(-4, 3, 8)$cm. They are not the same as shown in Figure 11. Note that the previous simulations demonstrated that the procedure **mi** was expected to give a more accurate localization result than did the procedure **adp**.

Although the areas highlighted in Figures 10 and 11 were varying over sessions, they did reveal the following known regions of face perception: the occipital face area (OFA), the inferior occipital gyrus (IOG), and the superior temporal sulcus (STS), and the precuneus (PCu). Interestingly, in each session, we identified a pair of nearly symmetric sources, of which one was strongly powered while the other was weakly powered. This phenomenon occurred due to source cancellations that prevented the second source from identification as we have demonstrated in our simulation studies. The time-courses plots in Figure 12 showed the response differences under face stimuli and scrambled face stimuli during the time period 100ms$\sim$300ms. The results are consistent with recent findings in face-perception studies by using an MEG-based multiple sparse prior approach (Friston et al., 2006; Henson et al., 2011) and by other empirical approaches (e.g., Pitcher et al., 2011; Kanwisher et al., 1997). However, in the first two papers, the authors made a parametric model assumption on source temporal correlation structures and imposed a limit on the number of candidate sources in the model, whereas in our approach, the model is non-parametric and allows for arbitrary number of candidate sources.

6. Discussion and Conclusion

In the present study, we have proposed a class of vector-beamformers by thresholding the sensor sample covariance matrix. The consistency and the convergence rate of the proposed vector-beamformer estimation have been proved in the presence of multiple sources. The theory has provided a basis for choosing the threshold $\tau_{nJ} = c_0\sigma_0^2\sqrt{\log(n)/J}$ in the beamformer construction. However, it requires a number of conditions. As pointed out in Section 3, conditions (A1)$\sim$(A4) are commonly used assumptions in literature for studying multiple time series (Sekihara and Nagarajan, 2008; Fan et al., 2011). We only need to validate the coherence stability condition which is new. Intuitively, the strength of correlations between sensors (therefore the absolute partial correlation) will increase when the number of sensors increases in general. Taking the face-perception data (session 1) as an example, we show how to validate it empirically by random sub-samples of the 306 sensors below. We take the first two peaks in Figure 8 as two true sources. They are located at CTF (-4,3,8) cm and (-4,-5,5) cm respectively. First, we reparametrize the lead field matrix as in Section 3. Then, for $k = 1, 2, ..., 306$, we randomly choose k sensors, obtaining a $k \times 4461$ sub lead field matrix for the 1487 voxels in the brain. We calculate the maximum absolute partial correlation $d_{12}(k) = \max\{d_{1|2}, d_{2|2}\}$ between the two sources and the maximum absolute correlation $d_{\max}(k) = \max d_{x|2}$ for all voxels, where x is running over these voxels. Finally, we plot $d_{12}(k)$, $d_{\max}(k)$, and $\log(\log(k))$ against $k = 1, 2, ..., 306$ respectively as displayed in Figure 13. As expected, the result shows that both $d_{12}(k)$ and $d_{\max}(k)$ change very slowly when the number of sensors k changes, with a rate much slower than $\log(\log(k))$. This implies that the coherence stability condition nearly holds.

In real world situations, the underlying number of true sources, q needs to be estimated. The influence of q on the beamformer estimators can be measured by the lead field partial correlation coefficient a_{nq} defined in Section 3. In this paper, local peaks on transverse slices have been used to reduce the search space of sources. We can cluster the local peak values into two groups, one of which is taken as a group of potential sources. The size of the selected group gives an estimate of q. In the face-perception data, we have only presented the first two sources which are ranked higher than the remaining local peaks, because these two are of clear neurological implications. Our approach is non-parametric in the sense that we have not made any parametric assumptions on the model (1.1). However, if we are willing to assume a family of parametric models for background noises, then we can determine q via model selection criteria such as Bayesian information criterion.

By theoretical and empirical studies, we have shown that due to source cancellations, the beamformer power estimator can be inconsistent if the underlying multiple sources are not well separated in terms of a lead field distance. Unlike the existing theories in the literature, the new theory is applicable to more general scenarios, where multiple sources exist and the sensor covariance matrix are estimated from the data. In the new theory, we do assume that the powers of the unknown no-null sources as well as the underlying number q are not growing with the number of sensors n. This assumption is natural to neurologists and has simplified mathematical derivations of the theory very much. However, the theory can be extended to the case where these quantities are growing with n. In the theory, we have not impose any constraint on p as we only consider local behavior of beamformers. If we want to investigate global properties of the neuronal activity map, then some constraints need to be imposed on the growth rate of p with respect to n.

The performances of the proposed beamformers have further been assessed by simulations and real data analyses. We have demonstrated that thresholding the sensor covariance matrix can help reduce the source localization bias when the data have a low SNR value. We have applied the vector-beamformer to an MEG data set for identifying the active regions related to human face perception. Some excellent agreements have been found between the current results and the existing neurological facts on human face perception. Finally, we note that there are other ways to measure the contrast between two source covariances such as the information-divergence. The theory can be easily extended to this case. The details will be presented elsewhere.

7. Proofs

In this section we prove the theorems and corollaries in Section 3.

To prove Theorem 1, we need the following lemma.

Lemma 10 *If $a_{nq} \to \infty$ as $n \to \infty$, then we have*

$$\begin{aligned}
H_j^T C_k^{-1} H_j &= b_{jj|k} + \frac{c_{jj|k}}{n} + O(a_{nk}^{-2}), \quad b_{jj|k} = \Sigma_j^{-1}, \quad \text{for } 1 \le j \le k \\
H_{j_1}^T C_k^{-1} H_{j_2} &= \frac{c_{j_1 j_2|k}}{n} + O(a_{nk}^{-2}), \quad \text{for } 1 \le j_1 \ne j_2 \le k \\
H_j^T C_k^{-1} H_x &= b_{jx|k} + \frac{c_{jx|k}}{n} + O(a_{nk}^{-2}), \quad \text{for } 1 \le j \le k,\ x \notin R_k \\
H_y^T C_k^{-1} H_x &= n a_{yx|k} + b_{yx|k} + O(a_{nk}^{-1}), \quad \text{for } x, y \notin R_k
\end{aligned}$$

where $a_{nk} = n \min_{1\le j\le k-1} tr\left(a_{(j+1)(j+1)|j}\right)$, $R_k = \{r_1, \ldots, r_k\}$, $C_k = \sum_{j=1}^{k} H_j^T \Sigma_j H_j + \sigma_0^2 I_n$, *and* $a_{yx|k}$, $b_{yx|k}$ *and* $c_{jj|k}$ *are defined before and the other* c'*s are defined iteratively as follows:*

$$c_{j_1 j_2|k} = \begin{cases} b_{j_1 k|(k-1)} \Sigma_k^{-1} a_{kk|(k-1)}^{-1}, & 1 \le j_1 \le k-1, j_2 = k \\ \Sigma_k^{-1} a_{kk|(k-1)}^{-1} b_{kj_2|(k-1)}, & 1 \le j_2 \le k-1, j_1 = k \\ c_{j_1 j_2|(k-1)} - b_{j_1 k|(k-1)} a_{kk|(k-1)}^{-1} b_{kj_2|(k-1)}, & 1 \le j_1 \ne j_2 \le k-1. \end{cases}$$

$$c_{jx|k} = \begin{cases} \left(a_{kk|(k-1)}\Sigma_k\right)^{-1} \left\{ b_{kx|(k-1)} \right. \\ \left. - \left(I_3 + b_{kk|(k-1)}\right) \left(a_{kk|(k-1)}\Sigma_k\right)^{-1} a_{kx|(k-1)} \right\}, & j = k \\ c_{jx|(k-1)} - c_{jk|(k-1)} a_{kk|(k-1)}^{-1} a_{kx|(k-1)} \\ -b_{jk|(k-1)} a_{kk|(k-1)}^{-1} b_{kx|(k-1)} & 1 \le j \le k-1 \\ +b_{jk|(k-1)} a_{kk|(k-1)}^{-1} \left[\Sigma_k^{-1} + b_{kk|(k-1)}\right] a_{kk|(k-1)}^{-1} a_{kx|(k-1)}. \end{cases}$$

Proof Note that under the stability condition and the assumption that $a_{nq} \to \infty$, we have $b_{yx|k} = O(1),\ 1 \le k \le q$. And for any r_x in the source space,

$$\frac{c_{1x|1}}{n} = O(n^{-1}), \quad \frac{c_{yx|k}}{n} = O(a_{n(k-1)}^{-1}), 1 \le y \le k, 2 \le k \le q.$$

We prove the lemma by induction. For $k = 1$, we have

$$\begin{aligned} C_1^{-1} &= \sigma_0^{-2} I_n - \sigma_0^{-4} H_1 (\Sigma_1^{-1} + n\sigma_0^{-2} I_3)^{-1} H_1^T, \\ H_1^T C_1^{-1} H_1 &= n\sigma_0^{-2} I_3 - n^2 \sigma_0^{-4} (\Sigma_1^{-1} + n\sigma_0^{-2} I_3)^{-1} \\ &= n\sigma_0^{-2} \left(I_3 - \left(I_3 + \Sigma_1^{-1} \frac{\sigma_0^2}{n} \right)^{-1} \right) \\ &= n\sigma_0^{-2} \left(I_3 + n\Sigma_1 \sigma_0^{-2} \right)^{-1} \\ &= \Sigma_1^{-1} \left(I_3 - \sigma_0^2 \Sigma_1^{-1} / n \right) + O(n^{-2}) \\ &= \Sigma_1^{-1} - \Sigma_1^{-1} \sigma_0^2 \Sigma_1^{-1} / n + O(n^{-2}) \\ &= b_{11|1} + \frac{c_{11|1}}{n} + O(n^{-2}), \end{aligned}$$

where

$$b_{11|1} = \Sigma_1^{-1},\ c_{11|1} = -\sigma_0^2 \Sigma_1^{-2}.$$

Analogously,

$$\begin{aligned} H_1^T C_1^{-1} H_x &= \sigma^{-2} H_1^T H_x - \sigma_0^{-4} n (\Sigma_1^{-1} + n\sigma_0^{-2} I_3)^{-1} H_1^T H_x \\ &= \left(I_3 - \left(I_3 + \Sigma_1^{-1} \sigma_0^2 / n \right)^{-1} \right) H_1^T H_x \\ &= \left(I_3 + \frac{n}{\sigma_0^2} \Sigma_1 \right)^{-1} H_1^T H_x \\ &= \Sigma_1^{-1} \left(I_3 + \frac{\sigma_0^2}{n} \Sigma_1^{-1} \right)^{-1} \rho_{1x} \\ &= \Sigma_1^{-1} \rho_{1x} - \Sigma_1^{-1} \sigma_0^2 \Sigma_1^{-1} \rho_{1x} / n + O(n^{-2}) \\ &= b_{1x|1} + \frac{c_{1x|1}}{n} + O(n^{-2}), \end{aligned}$$

where

$$b_{1x|1} = \Sigma_1^{-1}\rho_{1x},\ c_{11|1} = -\Sigma_1^{-2}\sigma_0^2\rho_{1x}.$$

And

$$\begin{aligned} H_y^T C_1^{-1} H_x &= \sigma_0^{-2} H_y^T H_x - \sigma_0^{-4} H_y^T H_1 (\Sigma_1^{-1} + n\sigma_0^{-2} I_3)^{-1} H_1^T H_x \\ &= n\sigma_0^{-2}\rho_{yx} - \sigma_0^{-4} H_y^T H_1 \frac{\sigma_0^2}{n} \left(I_3 + \sigma_0^2\Sigma_1^{-1}/n\right)^{-1} H_1^T H_x \\ &= n\sigma_0^{-2}\rho_{yx} - n\sigma_0^{-2}\rho_{y1}\left(I_3 - \sigma_0^2\Sigma_1^{-1}/n\right)\rho_{1x} + O(n^{-1}) \\ &= n\sigma_0^{-2}\rho_{y1x} + \rho_{y1}\Sigma_1^{-1}\rho_{1x} + O(n^{-1}) \\ &= n a_{yx|1} + b_{yx|1} + O(n^{-1}), \end{aligned}$$

where

$$\rho_{y1x} = \rho_{yx} - \rho_{y1}\rho_{1x},\ a_{yx|1} = \sigma_0^{-2}\rho_{y1x},\ b_{yx|1} = \rho_{y1}\Sigma_1^{-1}\rho_{1x}.$$

This implies the lemma holds for $k = 1$.

Assuming the lemma holds for the cases with less or equal to k sources, we show that it is also true for the case with $k+1$ sources by invoking the matrix inversion formulas

$$\begin{aligned} C_{k+1}^{-1} &= C_k^{-1} - C_k^{-1} H_{k+1}\left(\Sigma_{k+1}^{-1} + H_{k+1}^T C_k^{-1} H_{k+1}\right)^{-1} H_{k+1}^T C_k^{-1}, \\ C_k^{-1} &= C_{k+1}^{-1} + C_{k+1}^{-1} H_{k+1}\Sigma_{k+1} H_{k+1}^T C_k^{-1}. \end{aligned} \tag{7.6}$$

The details are as follows.

For $1 \le j \le k$,

$$\begin{aligned} H_j^T C_{k+1}^{-1} H_j &= H_j^T C_k^{-1} H_j - \left(H_j^T C_k^{-1} H_{k+1}\right) \times \left(\Sigma_{k+1}^{-1} + H_{k+1}^T C_k^{-1} H_{k+1}\right)^{-1} \left(H_{k+1}^T C_k^{-1} H_j\right) \\ &= b_{jj|k} + \frac{c_{jj|k}}{n} + O(a_{nk}^{-2}) - \left(b_{j(k+1)|k} + \frac{c_{j(k+1)|k}}{n} + O(a_{nk}^{-2})\right) \\ &\quad \times \left(\Sigma_{k+1}^{-1} + n a_{(k+1)(k+1)|k} + b_{(k+1)(k+1)|k} + O(a_{nk}^{-1})\right)^{-1} \\ &\quad \times \left(b_{j(k+1)|k} + \frac{c_{j(k+1)|k}}{n} + O(a_{nk}^{-2})\right)^T \\ &= b_{jj|k} + \frac{c_{jj|k}}{n} + O(a_{nk}^{-2}) \\ &\quad - \left(b_{j(k+1)|k} + O(a_{nk}^{-1})\right)\left((n a_{(k+1)(k+1)|k})^{-1} - O(a_{n(k+1)}^{-2})\right) \\ &\quad \times \left(b_{j(k+1)|k} + O(a_{nk}^{-1})\right)^T \\ &= b_{jj|k} + \frac{c_{jj|k}}{n} - \frac{1}{n} b_{j(k+1)|k} a_{(k+1)(k+1)|k}^{-1} b_{j(k+1)|k}^T + O(a_{n(k+1)}^{-2}) \\ &= b_{jj|(k+1)} + \frac{c_{jj|(k+1)}}{n} + O(a_{n(k+1)}^{-2}). \end{aligned}$$

For $j = k+1$, we have

$$\begin{aligned} H_j^T C_{k+1}^{-1} H_j &= H_{k+1}^T C_{k+1}^{-1} H_{k+1} = H_{k+1}^T C_k^{-1} H_{k+1}\left(I_3 + \Sigma_{k+1} H_{k+1}^T C_k^{-1} H_{k+1}\right)^{-1} \\ &= \left(n a_{(k+1)(k+1)|k} + b_{(k+1)(k+1)|k} + O(a_{nk}^{-1})\right) \\ &\quad \times \left(I_3 + \Sigma_{k+1}\left(n a_{(k+1)(k+1)|k} + b_{(k+1)(k+1)|k} + O(a_{nk}^{-1})\right)\right)^{-1} \end{aligned}$$

$$\begin{aligned}
&= \left(na_{(k+1)(k+1)|k} + b_{(k+1)(k+1)|k} + O(a_{nk}^{-1})\right)\left(na_{(k+1)(k+1)|k}\right)^{-1} \\
&\quad \times \left(I_3 + (na_{(k+1)(k+1)|k})^{-1}\Sigma_{k+1}^{-1} + O(a_{n(k+1)}^{-2})\right)^{-1}\Sigma_{k+1}^{-1} \\
&= \Sigma_{k+1}^{-1} - \frac{1}{n}\Sigma_{k+1}^{-1}a_{(k+1)(k+1)|k}^{-1}\Sigma_{k+1}^{-1} + O(a_{n(k+1)}^{-2}) \\
&= b_{(k+1)(k+1)|(k+1)} + \frac{c_{(k+1)(k+1)|(k+1)}}{n} + O(a_{n(k+1)}^{-2}).
\end{aligned}$$

This completes the proof of the first equation in the lemma.

To prove the second equation in the lemma, we let

$$\begin{aligned}
B &= \left[\Sigma_{k+1}na_{(k+1)(k+1)|k}\right]^{-1} + \left[\Sigma_{k+1}na_{(k+1)(k+1)|k}\right]^{-\frac{1}{2}} \\
&\quad \times\Sigma_{k+1}b_{(k+1)(k+1)|k}\left[\Sigma_{k+1}na_{(k+1)(k+1)|k}\right]^{-\frac{1}{2}}.
\end{aligned}$$

Then, when $1 \le j_1 \le k$, $j_2 = k+1$, we have

$$\begin{aligned}
H_{j_1}^T C_{k+1}^{-1}H_{j_2} &= H_{j_1}^T C_{k+1}^{-1}H_{k+1} = H_{j_1}^T C_k^{-1}H_{k+1}\left\{I_3 + \Sigma_{k+1}H_{k+1}^T C_k^{-1}H_{k+1}\right\}^{-1} \\
&= \left(b_{j_1(k+1)|k} + \frac{1}{n}c_{j_1(k+1)|k} + O(a_{nk}^{-2})\right) \\
&\quad \times\left(I_3 + \Sigma_{k+1}\left(na_{(k+1)(k+1)|k} + b_{(k+1)(k+1)|k} + O(a_{nk}^{-1})\right)\right)^{-1} \\
&= \left(b_{j_1(k+1)|k} + \frac{1}{n}c_{j_1(k+1)|k} + O(a_{nk}^{-2})\right)\left[\Sigma_{k+1}na_{(k+1)(k+1)|k}\right]^{-\frac{1}{2}} \\
&\quad \times(I_3 + B + O(a_{n(k+1)}^{-2}))^{-1}\left[\Sigma_{k+1}na_{(k+1)(k+1)|k}\right]^{-\frac{1}{2}} \\
&= b_{j_1(k+1)|k}\left[\Sigma_{k+1}na_{(k+1)(k+1)|k}\right]^{-\frac{1}{2}} \\
&\quad \times\left(I_3 + O\left((na_{(k+1)(k+1)|k})^{-1}\right)\right)\left[\Sigma_{k+1}na_{(k+1)(k+1)|k}\right]^{-\frac{1}{2}} + O(a_{n(k+1)}^{-2}) \\
&= \frac{1}{n}b_{j_1(k+1)|k}\Sigma_{k+1}^{-1}a_{(k+1)(k+1)|k}^{-1} + O(a_{n(k+1)}^{-2}) \\
&= \frac{c_{j_1(k+1)|(k+1)}}{n} + O(a_{n(k+1)}^{-2}).
\end{aligned}$$

Similarly, when $1 \le j_1 \ne j_2 \le k$, we have

$$\begin{aligned}
H_{j_1}^T C_{k+1}^{-1}H_{j_2} &= H_{j_1}^T C_k^{-1}H_{j_2} - H_{j_1}^T C_k^{-1}H_{k+1}\left(\Sigma_{k+1}^{-1} + H_{k+1}^T C_k^{-1}H_{k+1}\right)^{-1}H_{k+1}^T C_k^{-1}H_{j_2} \\
&= \frac{1}{n}c_{j_1j_2|k} + O(a_{nk}^{-2}) - \frac{1}{n}b_{j_1(k+1|k)}a_{(k+1)(k+1)|k}^{-1}b_{(k+1)j_2|k} + O(a_{n(k+1)}^{-2}) \\
&= \frac{1}{n}c_{j_1j_2|(k+1)} + O(a_{n(k+1)}^{-2}).
\end{aligned}$$

We complete the proof of the second equation in the lemma.

To prove the third equation in the lemma, we let

$$\begin{aligned}
D &= \left(na_{(k+1)(k+1)|k}\Sigma_{k+1}\right)^{-1} + \left(na_{(k+1)(k+1)|k}\Sigma_{k+1}\right)^{-\frac{1}{2}} \\
&\quad \times b_{(k+1)(k+1k|k)}\Sigma_{k+1}\left(na_{(k+1)(k+1)|k}\Sigma_{k+1}\right)^{-\frac{1}{2}}, \\
F &= (na_{(k+1)(k+1)|k})^{-\frac{1}{2}}\left(\Sigma_{k+1}^{-1} + b_{(k+1)(k+1)|k}\right)(na_{(k+1)(k+1)|k})^{-\frac{1}{2}}.
\end{aligned}$$

Then, for $j = k+1$,

$$\begin{aligned}
H_j^T C_{k+1}^{-1} H_x &= \left[I_3 + H_{k+1}^T C_k^{-1} H_{k+1} \Sigma_{k+1}\right]^{-1} H_{k+1}^T C_k^{-1} H_x \\
&= \left[I_3 + n a_{(k+1)(k+1)|k} \Sigma_{k+1} + b_{(k+1)(k+1)|k} \Sigma_{k+1} + O(a_{nk}^{-1})\right]^{-1} \\
&\quad \times \left[n a_{(k+1)x|k} + b_{(k+1)x|k} + O(a_{nk}^{-1})\right] \\
&= \left(n a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-\frac{1}{2}} \left(I_3 + D + O(a_{n(k+1)}^{-2})\right)^{-1} \left(n a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-\frac{1}{2}} \\
&\quad \times \left[n a_{(k+1)x|k} + b_{(k+1)x|k} + O(a_{nk}^{-1})\right] \\
&= \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-\frac{1}{2}} \left\{I_3 - D + O(a_{n(k+1)}^{-2})\right\} \\
&\quad \times \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-\frac{1}{2}} \left(a_{(k+1)x|k} + b_{(k+1)x|k}/n + O(a_{nk}^{-1})/n\right) \\
&= \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-1} a_{(k+1)x|k} \\
&\quad - \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-1/2} D \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-1/2} a_{(k+1)x|k} \\
&\quad + O(a_{n(k+1)}^{-3}) + \frac{1}{n} \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-1} b_{(k+1)x|k} + O(a_{n(k+1)}^{-2}) \\
&= \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-1} a_{(k+1)x|k} + \frac{1}{n} \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-1} \\
&\quad \times \left\{b_{(k+1)x|k} - \left(I_3 + b_{(k+1)(k+1)|k} \Sigma_{k+1}\right) \left(a_{(k+1)(k+1)|k} \Sigma_{k+1}\right)^{-1} a_{(k+1)x|k}\right\} \\
&\quad + O(a_{n(k+1)}^{-2}) \\
&= b_{(k+1)x|(k+1)} + \frac{1}{n} c_{(k+1)x|(k+1)} + O(a_{n(k+1)}^{-2}).
\end{aligned}$$

For $1 \le j \le k$, we have

$$\begin{aligned}
H_j^T C_{k+1}^{-1} H_x &= H_j^T C_k^{-1} H_x - H_j^T C_k^{-1} H_{k+1} \left(\Sigma_{k+1}^{-1} + H_{k+1}^T C_k^{-1} H_{k+1}\right)^{-1} H_{k+1}^T C_k^{-1} H_x \\
&= b_{jx|k} + \frac{1}{n} c_{jx|k} + O(a_{nk}^{-2}) - \left(b_{j(k+1)|k} + \frac{1}{n} c_{j(k+1)|k} + O(a_{nk}^{-2})\right) \\
&\quad \times \left(\Sigma_{k+1}^{-1} + n a_{(k+1)(k+1)|k} + b_{(k+1)(k+1)|k} + O(a_{nk}^{-2})\right)^{-1} \\
&\quad \times \left(n a_{(k+1)x|k} + b_{(k+1)x|k} + O(a_{nk}^{-1})\right) \\
&= b_{jx|k} + \frac{1}{n} c_{jx|k} + O(a_{nk}^{-2}) - \left(b_{j(k+1)|k} + \frac{1}{n} c_{j(k+1)|k} + O(a_{nk}^{-2})\right) \\
&\quad \times (n a_{(k+1)(k+1)|k})^{-1/2} \left[I_3 + F + O(a_{nk}^{-2})\right]^{-1} (n a_{(k+1)(k+1)|k})^{-1/2} \\
&\quad \times \left(n a_{(k+1)x|k} + b_{(k+1)x|k} + O(a_{nk}^{-1})\right) \\
&= b_{jx|k} + \frac{1}{n} c_{jx|k} + O(a_{nk}^{-2}) - \left(b_{j(k+1)|k} + \frac{1}{n} c_{j(k+1)|k} + O(a_{nk}^{-2})\right) \\
&\quad \times \Big(a_{(k+1)(k+1)|k}^{-1} a_{(k+1)x|k} - a_{(k+1)(k+1)|k}^{-1/2} F a_{(k+1)(k+1)|k}^{-1/2} a_{(k+1)x|k} + O(a_{nk}^{-2}) \\
&\quad (n a_{(k+1)(k+1)|k})^{-1} b_{(k+1)x|k} - a_{(k+1)(k+1)|k}^{-1/2} F a_{(k+1)(k+1)|k}^{-1/2} b_{(k+1)x|k}/n \\
&\quad + O(a_{n(k+1)}^{-2})\Big)
\end{aligned}$$

$$
\begin{aligned}
&= b_{jx|k} + \frac{1}{n}c_{jx|k} + O(a_{nk}^{-2}) - \left(b_{j(k+1)|k} + \frac{1}{n}c_{j(k+1)|k} + O(a_{nk}^{-2})\right) \\
&\quad \times \left(a^{-1}_{(k+1)(k+1)|k}a_{(k+1)x|k} - \frac{1}{n}a^{-1}_{(k+1)(k+1)|k}(\Sigma_{k+1}^{-1} + b_{(k+1)(k+1)|k}) \right. \\
&\quad \left. \times a^{-1}_{(k+1)(k+1)|k}a_{(k+1)x|k}\frac{1}{n}a^{-1}_{(k+1)(k+1)|k}b_{(k+1)x|k} + O(a^{-2}_{n(k+1)})\right) \\
&= b_{jx|k} + \frac{1}{n}c_{jx|k} - b_{j(k+1)|k}a^{-1}_{(k+1)(k+1)|k}a_{(k+1)x|k} \\
&\quad -\frac{1}{n}\left\{c_{j(k+1)|k}a^{-1}_{(k+1)(k+1)|k}a_{(k+1)x|k} + b_{j(k+1)|k}a^{-1}_{(k+1)(k+1)|k}b_{(k+1)x|k}\right. \\
&\quad \left. -b_{j(k+1)|k}a^{-1}_{(k+1)(k+1)|k}\left[\Sigma_{k+1}^{-1} + b_{(k+1)(k+1)|k}\right]a^{-1}_{(k+1)(k+1)|k}a_{(k+1)x|k}\right\} \\
&\quad +O(a^{-2}_{n(k+1)}) \\
&= b_{jx|k} - b_{j(k+1)|k}a^{-1}_{(k+1)(k+1)|k}a_{(k+1)x|k} + \frac{1}{n}c_{jx|(k+1)} + O(a^{-2}_{n(k+1)}). \qquad (7.7)
\end{aligned}
$$

Note that for $k = j$,

$$a_{jx|j} = a_{jx|(j-1)} - a_{jj|(j-1)}a^{-1}_{jj|(j-1)}a_{jx|(j-1)} = 0.$$

Assuming that for $k = j + m, m > 0$, the statement is true, i.e., $a_{jx|(k+m)} = 0$ for all x. Then,

$$
\begin{aligned}
a_{jx|(j+m+1)} &= a_{jx|(j+m)} - a_{j(j+m+1)|(j+m)}a^{-1}_{(j+m+1)(j+m+1)|(j+m)}a_{(j+m+1)x|(j+m)} \\
&= 0.
\end{aligned}
$$

By induction, we have that $a_{jx|k} = 0$ for all x, $j \le k$. This implies that and

$$b_{jx|(k+1)} = b_{jx|k} - b_{j(k+1)|k}a^{-1}_{(k+1)(k+1)|k}a_{(k+1)x|k}$$

by the definition of $b_{jx|(k+1)}$. Combining this with (7.7), we complete the proof of the third equation in the lemma.

Finally, we turn to the last equation in the lemma. Assume that the equation holds for the case k. We show that it also holds for $k + 1$ below. For $x, y \notin R_{k+1}$ (thus $x, y \notin R_k$), by the assumption, we have

$$H_y^T C_k^{-1} H_x = na_{yx|k} + b_{yx|k} + O(a_{nk}^{-1}).$$

This together with (7.6) yields

$$
\begin{aligned}
H_y^T C_{k+1}^{-1} H_x &= H_y^T C_k^{-1} H_x - H_y^T C_k^{-1} H_{k+1}\left(\Sigma_{k+1}^{-1} + H_{k+1}^T C_k^{-1} H_{k+1}\right)^{-1} H_{k+1}^T C_k^{-1} H_x \\
&= na_{yx|k} + b_{yx|k} + O(a_{nk}^{-1}) - \left(na_{y(k+1)|k} + b_{y(k+1)|k} + O(a_{nk}^{-1})\right) \\
&\quad \times \left(\Sigma_{k+1}^{-1} + na_{(k+1)(k+1)|k} + b_{(k+1)(k+1)|k} + O(a_{nk}^{-1})\right)^{-1} \\
&\quad \times \left(na_{(k+1)x|k} + b_{(k+1)x|k} + O(a_{nk}^{-1})\right) \\
&= na_{yx|k} + b_{yx|k} + O(a_{nk}^{-1}) - \left(na_{y(k+1)|k} + b_{y(k+1)|k} + O(a_{nk}^{-1})\right)
\end{aligned}
$$

$$
\begin{aligned}
&\times \left((na_{(k+1)(k+1)|k})^{-1} - \frac{1}{n^2} a^{-1}_{(k+1)(k+1)|k} \left(\Sigma^{-1}_{k+1} + b_{(k+1)(k+1)|k} \right) \right. \\
&\left. \times a^{-1}_{(k+1)(k+1)|k} + O(a^{-3}_{n(k+1)}) \right) \left(na_{y(k+1)|k} + b_{y(k+1)|k} + O(a^{-1}_{nk}) \right) \\
= \;& na_{yx|k} + b_{yx|k} + O(a^{-1}_{nk}) - \left\{ a_{y(k+1)|k} a^{-1}_{(k+1)(k+1)|k} \right. \\
&- a_{y(k+1)|k} a^{-1}_{(k+1)(k+1)|k} \left(\Sigma^{-1}_{k+1} + b_{(k+1)(k+1)|k} \right) a^{-1}_{(k+1)(k+1)|k} / n \\
&\left. + b_{y(k+1)|k} a^{-1}_{(k+1)(k+1)|k} / n + O(a^{-2}_{n(k+1)}) \right\} \\
&\times \left(na_{(k+1)x|k} + b_{(k+1)x|k} + O(a^{-1}_{nk}) \right) \\
= \;& n \left[a_{yx|k} - a_{y(k+1)|k} a^{-1}_{(k+1)(k+1)|k} a_{(k+1)x|k} \right] \\
&+ \left[b_{yx|k} - b_{y(k+1)|k} a^{-1}_{(k+1)(k+1)|k} a_{(k+1)x|k} - a_{y(k+1)|k} a^{-1}_{(k+1)(k+1)|k} b_{(k+1)x|k} \right. \\
&\left. + a_{y(k+1)|k} a^{-1}_{(k+1)(k+1)|k} \left(\Sigma^{-1}_{k+1} + b_{(k+1)(k+1)|k} \right) a^{-1}_{(k+1)(k+1)|k} a_{(k+1)x|k} \right] \\
&+ O(a^{-1}_{n(k+1)}) \\
= \;& na_{yx|(k+1)} + b_{yx|(k+1)} + O(a^{-1}_{n(k+1)}).
\end{aligned}
$$

The proof is completed. ■

Proof of Theorem 1. Note that $b_{yx|1} = \rho(r_y, r_1)\Sigma_1^{-1}\rho(x_1, x)$, $a_{yx|1} = \sigma_0^{-2}(\rho_{yx} - \rho_{y1}\rho_{1x})$, and both are bounded. By induction and the stability condition, it can be shown that $a_{yx|k}$ and $b_{yx|k}$ are bounded for $2 \le k \le q$. If a_{nq} is bounded, then there exists k_m such that $na_{(k_m+1)(k_m+1)|k_m} = O(1)$ and $a_{nk_m} = \min_{1 \le j \le k_m - 1} na_{(j+1)(j+1)|j} \to \infty$ as n tends to infinity. By Lemma 10, we have

$$
H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \;=\; na_{(k_m+1)(k_m+1)|k_m} + b_{(k_m+1)(k_m+1)|k_m} + O(a^{-1}_{nk_m}),
$$

which is bounded and non-negative definite. Furthermore, there exists an orthogonal matrix Q and a diagonal matrix $D = \mathrm{diag}(d_1, d_2, d_3)$ such that

$$
\Sigma^{1/2}_{k_m+1} H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \Sigma^{1/2}_{k_m+1} = QDQ^T.
$$

Therefore,

$$
\begin{aligned}
& H^T_{k_m+1} C^{-1}_{k_m+1} H_{k_m+1} \\
&= H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \left(I_3 - \left(\Sigma^{-1}_{k_m+1} + H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \right)^{-1} H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \right) \\
&= H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \left(\Sigma^{-1}_{k_m+1} + H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \right)^{-1} \Sigma^{-1}_{k_m+1} \\
&= H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \Sigma^{1/2}_{k_m+1} \left(I_3 + \Sigma^{1/2}_{k_m+1} H^T_{k_m+1} C^{-1}_{k_m} H_{k_m+1} \Sigma^{1/2}_{k_m+1} \right)^{-1} \Sigma^{-1/2}_{k_m+1} \\
&= \Sigma^{-1/2}_{k_m+1} QDQ^T \left(I_3 + QDQ^T \right)^{-1} \Sigma^{-1/2}_{k_m+1} \\
&= \Sigma^{-1/2}_{k_m+1} \left(I_3 + QD^{-1}Q^T \right)^{-1} \Sigma^{-1/2}_{k_m+1} \\
&= \Sigma^{-1/2}_{k_m+1} \left(Q(I_3 + D^{-1})Q^T \right)^{-1} \Sigma^{-1/2}_{k_m+1} \\
&= \Sigma^{-1/2}_{k_m+1} Q(I_3 + D^{-1})^{-1} Q^T \Sigma^{-1/2}_{k_m+1}. \qquad (7.8)
\end{aligned}
$$

Note that $\Sigma_{k_m+1}^{1/2} H_{k_m+1}^T C_{k_m}^{-1} H_{k_m+1} \Sigma_{k_m+1}^{1/2} = O(1)$, which implies that $d_k \geq 0, 1 \leq k \leq 3$ are bounded. We can find a positive constant ϵ_0 such that $\max_{1\leq k\leq 3}(1+d_k^{-1})^{-1} < (1+\epsilon_0)^{-1}$ when n is large enough. Consequently, for any vector $a \in \mathbb{R}^3$ with $||a|| = 1$, we have

$$a^T \Sigma_{k_m+1}^{1/2} Q(I_3 + D^{-1}) Q^T \Sigma_{k_m+1}^{1/2} a > (1+\epsilon_0) a^T \Sigma_{k_m+1}^{1/2} Q Q^T \Sigma_{k_m+1}^{1/2} a,$$

which shows that $\Sigma_{k_m+1}^{1/2} Q(I_3 + D^{-1}) Q^T \Sigma_{k_m+1}^{1/2}$ (thus $\left[H_{k_m+1}^T C_{k_m+1}^{-1} H_{k_m+1}\right]^{-1}$ due to (7.8)) is asymptotically larger than $\Sigma_{k_m+1}(1+\epsilon_0)$.

We now consider the case where $a_{nq} \to \infty$. For $j = q$, by Lemma 10, we have

$$\begin{aligned}
\frac{c_{qq|q}}{n} &= -\Sigma_q^{-1} \left[n a_{qq|(q-1)}\right]^{-1} \Sigma_l^{-1} = O(a_{nq}^{-1}), \\
\left[H_q^T C_q^{-1} H_q\right]^{-1} &= \left[\Sigma_q^{-1} + \frac{c_{qq|q}}{n} + O(a_{nq}^{-2})\right]^{-1}, \\
&= \Sigma_q^{1/2} \left[I_3 + \Sigma_q^{1/2} \frac{c_{qq|q}}{n} \Sigma_q^{1/2} + O(a_{nq}^{-2})\right]^{-1} \Sigma_q^{1/2} \\
&= \Sigma_q^{1/2} \left[I_3 - \Sigma_q^{1/2} \frac{c_{qq|q}}{n} \Sigma_q^{1/2} + O(a_{nq}^{-2})\right] \Sigma_q^{1/2} \\
&= \Sigma_q - \left[n a_{qq|(q-1)}\right]^{-1} + O(a_{nq}^{-2})
\end{aligned}$$

as $n \to \infty$. For $1 \leq j \leq q-1$, by Lemma 10, we have

$$H_j^T C_q^{-1} H_j = \Sigma_j^{-1} + \frac{c_{jj|q}}{n} + O(a_{nq}^{-2}),$$

where $\frac{c_{jj|q}}{n} = O(a_{nq}^{-1})$. This entails

$$\begin{aligned}
\left[H_j^T C_q^{-1} H_j\right]^{-1} &= \Sigma_j^{1/2} \left(I_3 + \frac{1}{n} \Sigma_j^{1/2} c_{jj|q} \Sigma_j^{1/2} + O(a_{nq}^{-2})\right)^{-1} \Sigma_j^{1/2} \\
&= \Sigma_j^{1/2} \left(I_3 - \frac{1}{n} \Sigma_j^{1/2} c_{jj|q} \Sigma_j^{1/2} + O(a_{nq}^{-2})\right) \Sigma_j^{1/2} \\
&= \Sigma_j - \frac{1}{n} \Sigma_j c_{jj|q} \Sigma_j + O(a_{nq}^{-2}).
\end{aligned}$$

For any location r_x, by Lemma 10, we have

$$\begin{aligned}
\left[H_x^T C_q^{-1} H_x\right]^{-1} &= \frac{1}{n} \left[I_3 + \frac{1}{n} a_{xx|q}^{-1} b_{xx|q} + O(a_{nq}^{-2})\right]^{-1} a_{xx|q}^{-1} \\
&= \frac{1}{n} \left[I_3 - \frac{1}{n} a_{xx|q}^{-1} b_{xx|q} + O(a_{nq}^{-2})\right] a_{xx|q}^{-1} \\
&= \frac{1}{n} a_{xx|q}^{-1} - \frac{1}{n^2} a_{xx|q}^{-1} b_{xx|q} a_{xx|q}^{-1} + O(a_{nq}^{-3}).
\end{aligned}$$

The proof is completed.

Proof of Corollary 2. First, let $A_n = \left[{H_{k_m+1}}^T C_l^{-1} H_{k_m+1}\right]^{-1}$. If $a_{nq} = O(1)$ and $\max_{1\leq k\leq q} d_{k|q} = O(1)$, then by Theorem (1), there exists a positive constant ϵ_0 such that

$\min_{||a||=1} a^T(A_n - \Sigma_{k_m+1})a > \epsilon_0$ for large n. Let $a_1 = (1,0,0)^T, a_2 = (0,1,0)^T$ and $a_3 = (0,0,1)^T$. Then, we have

$$\begin{aligned}
\mathrm{tr}(A_n) &= \mathrm{tr}(A_n \sum_{k=1}^{3} a_k a_k^T) = \sum_{k=1}^{3} \mathrm{tr}(A_n a_k a_k^T) \\
&= \sum_{k=1}^{3} a_k^T A_n a_k > 3\epsilon_0 + \sum_{k=1}^{3} a_k \Sigma_{k_m+1} a_k^T \\
&= 3\epsilon_0 + \sum_{k=1}^{3} \mathrm{tr}(\Sigma_{k_m+1} a_k a_k^T) = 3\epsilon_0 + \mathrm{tr}(\Sigma_{k_m+1} \sum_{k=1}^{3} a_k a_k^T) \\
&= 3\epsilon_0 + \mathrm{tr}(\Sigma_{k_m+1}),
\end{aligned}$$

which implies $\mathrm{tr}(A_n)$ is asymptotically larger than Σ_{k_m+1}.

To prove Theorem 2, we need two more lemmas as follows and the following condition

$$(A1') : \{\mathbf{Y}(t_j) : 1 \le j \le J\} \text{ is stationary and has a finite covariance matrix.}$$

Lemma 11 *Under Conditions (A1') and (A3)∼(A4), if* $\tau_{nJ} = O(\sqrt{\log(n)/J})$ *and* $n\tau_{nJ} = o(1)$ *as* $n \to \infty$ *and* $J \to \infty$, *then*

(i) $\max_{1\le i,j\le n} |\hat{c}_{ij} - c_{ij}| = O_p(\sqrt{\log(n)/J})$,

(ii) $||\hat{C}(\tau_{nJ}) - C|| = O_p(m_n\sqrt{\log(n)/J})$,

(iii) $||\hat{C}(0) - C|| \le (m_n + n)\tau_{nJ}$,

where $m_n = \max_{1\le i\le n} \sum_{j=1}^{n} I(c_{ij} \ne 0) \le n$.

Proof. Let $\kappa_3 = \max\{2(2/\kappa_1 + 1/\kappa_2) - 1, (4/3)(1/\kappa_1 + 1/\kappa_2) - 1/3, 1\}$. Then $n\sqrt{\log(n)/J} = o(1)$ yields $(\log(n))^{\kappa_3}/J = o(1)$. We adopted the techniques of Bickel and Levina (2008); Fan et al. (2011); Zhang et al. (2014) to prove it. To prove (i), we set up more notations. Let $\tau(t)$ be the so-called Dedecker-Prieur τ-mixing coefficients (Merlevède et al., 2011, see). Let

$$\Theta(u,t) = \infty\{v > 0 : P(|y_1(t)y_2(t)| > v) \le u\}, \quad \psi_y(M,t) = \max\{\min\{y_i(t)y_j(t), M\}, -M\}.$$

It follows from Lemma 7 in Dedecker and Prieur (2004) that

$$\sup_t \Theta(u,t) \le b_1(1 - \log(u))^{2/\kappa_1},$$

which, under Condition (A4), gives $\tau(t) \le b_2 \exp(-b_3 t^{\kappa_2})$. Similarly, it is derived from Remark 3 in Merlevède et al. (2011) that

$$\begin{aligned}
&\sup_{M>0}[\sup_t \mathrm{var}(\psi_y(M,t)) + 2\sum_{t_1>t_2} |\mathrm{cov}(\psi_y(M,t_1), \psi_y(M,t_2))] \\
&\quad \le \sup_{M>0}\sup_t \mathrm{var}(\psi_y(M,t)) \\
&+2\left(\sup_{M>0}\sup_t \mathrm{var}(\psi_y(M,t)) + 4\sum_{t>0}\int_0^{2\alpha(t)} (\sup_t \Theta(u))^2 du\right) < \infty.
\end{aligned}$$

Let $1/\kappa = 2/\kappa_1 + 1/\kappa_2$. By Theorem 1 in Merlevède et al. (2011), we can find positive constants $d_k, 1 \le k \le 5$ that only depend on $\tau_1, \kappa_2, b_2, b_3$ such that

$$\begin{aligned} P\left(|\frac{1}{J}\sum_{t=1} y_i(t)y_j(t) - c_{ij}| \ge u\right) &\le J\exp\left(-\frac{(Ju)^\kappa}{d_1}\right) + \exp\left(-\frac{(Ju)^2}{d_2(1+Jd_3)})\right) \\ &\quad + \exp\left(-\frac{(Ju)^2}{d_4 J}\exp\left(\frac{(Ju)^{\kappa(1-\kappa)}}{d_5(\log(Ju))^\kappa}\right)\right). \end{aligned}$$

Consequently,

$$\begin{aligned} &P\left(\max_{1\le i,j\le n}|\frac{1}{J}\sum_{t=1}^{J} y_i(t)y_j(t) - c_{ij}| > u\right) \\ &\le n^2 \max_{1\le i,j\le n} P\left(|\frac{1}{J}\sum_{t=1}^{J} y_i(t)y_j(t) - c_{ij}| > u\right) \\ &\le n^2 J\exp\left(-\frac{(Ju)^\kappa}{d_1}\right) + n^2\exp\left(-\frac{(Ju)^2}{d_2(1+Jd_3)}\right) \\ &+n^2\exp\left(-\frac{(Ju)^2}{d_4 J}\exp\left(\frac{(Ju)^{\kappa(1-\kappa)}}{d_5(\log(Ju))^\kappa}\right)\right). \end{aligned}$$

Let $u = A\sqrt{\log(n)/J}$. Then $Ju = \sqrt{J\log(n)}$. When both n and J tend to infinity, we have

$$\begin{aligned} n^2 J\exp\left(-\frac{(Ju)^\kappa}{d_1}\right) &= \exp\left(2\log(n) + \log(J) - \frac{(A\sqrt{J\log(n)})^\kappa}{d_1}\right) \\ &= \exp\left((2\frac{(\log(n))^{1-\kappa/2}}{J^{\kappa/2}} - \frac{A}{d_1})(J\log(n))^{\kappa/2} + \log(J)\right) \\ &= o(1), \end{aligned}$$

since $(\log(n))^{1-\kappa/2}/J^{\kappa/2} = o(1)$. Similarly, if we choose $A > \sqrt{2d_2(d_3+1)}$, we have

$$\begin{aligned} n^2\exp\left(-\frac{(Ju)^2}{d_2(1+Jd_3)}\right) &= n^2\exp\left(-\frac{A^2 J\log(n)}{d_2(1+Jd_3)}\right) \\ &= \exp\left(\left(2 - \frac{A^2}{d_2(d_3+1/J)}\right)\log(n)\right) = o(1). \end{aligned}$$

And

$$\begin{aligned} &n^2\exp\left(-\frac{(Ju)^2}{d_4 J}\exp\left(\frac{(Ju)^{\kappa(1-\kappa)}}{d_5(\log(Ju))^\kappa}\right)\right) \\ &= \exp\left(\log(n)\left(2 - \frac{A^2}{d_4}\exp\left(\frac{A^{\kappa(1-\kappa)}(J\log(n))^{\kappa(1-\kappa)/2}}{d_5(\log(A\sqrt{J\log(n)}))^\kappa}\right)\right)\right) \\ &= o(1). \end{aligned}$$

Therefore,

$$P\left(\max_{1\le i,j\le n}|\frac{1}{J}\sum_{t=1}^{J}y_i(t)y_j(t)-c_{ij}|>u\right)=o(1). \tag{7.9}$$

Note that for $u=A\sqrt{\log(n)/J}$, there exist positive constants $d_k, 1\le k\le 5$ so that

$$\begin{aligned}
P(\max_{1\le i,j\le n}|\bar{y}_i||\bar{y}_j|>u) &= P(\max_{1\le i\le n}|\bar{y}_i|>\sqrt{u})\\
&\le n\max_{1\le i\le n}P(|\bar{y}_i|>\sqrt{u})\\
&= nJ\exp\left(-\frac{(J\sqrt{u})^{\kappa_1}}{d_1}\right)+n\exp\left(-\frac{(J\sqrt{u})^2}{d_2(1+Jd_3)}\right)\\
&\quad +n\exp\left(-\frac{(J\sqrt{u})^2}{d_4J}\exp\left(\frac{(J\sqrt{u})^{\kappa_1(1-\kappa_1)}}{d_5(\log(Ju))^{\kappa_1}}\right)\right)\\
&= o(1),
\end{aligned}$$

since $(\log(n))^{4/(3\kappa_1)-1/3}/J=o(1)$ and $\log(n)/J=o(1)$. This together with (7.9) yields that for $u=O(\sqrt{\log(n)/J})$,

$$\begin{aligned}
P\left(\max_{1\le i,j\le n}|\hat{c}_{ij}-c_{ij}|>u\right) &\le P\left(\max_{1\le i,j\le n}|\frac{1}{J}\sum_{t=1}^{J}y_i(t)y_j(t)-c_{ij}|>u\right)\\
&\quad +P\left(\max_{1\le i,j\le n}|\bar{y}_i||\bar{y}_j|>u\right)=o(1),
\end{aligned}$$

which implies

$$\max_{1\le i,j\le n}|\hat{c}_{ij}-c_{ij}|=O_p\left(\sqrt{\log(n)/J}\right).$$

We turn to $\hat{C}(\tau_{nJ})$ in (ii). Let $T_1=||(\hat{c}_{ij}I(|\hat{c}_{ij}|>\tau_{nJ})-c_{ij}I(|c_{ij}|>\tau_{nJ}))||$. We have

$$\begin{aligned}
||\hat{C}(\tau_{nJ})-C|| &\le ||(\hat{c}_{ij}I(|\hat{c}_{ij}|>\tau_{nJ})-c_{ij}I(|c_{ij}|>\tau_{nJ}))||+||(c_{ij}I(|c_{ij}|\le\tau_{nJ}))||\\
&\le T_1+\max_i\sum_{j=1}^{n}|c_{ij}|I(|c_{ij}|\le\tau_{nJ})\\
&\le T_1+\tau_{nJ}m_n. \qquad (7.10)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
||\hat{C}(0)-C|| &\le T_1+\tau_{nJ}m_n+\max_i\sum_{j=1}^{n}|\hat{c}_{ij}|I(|\hat{c}_{ij}|\le\tau_{nJ})\\
&\le T_1+(m_n+n)\tau_{nJ}.
\end{aligned}$$

Note that

$$T_1 \le \max_i\sum_{j=1}^{n}|\hat{c}_{ij}I(|\hat{c}_{ij}|>\tau_{nJ})-c_{ij}I(|c_{ij}|>\tau_{nJ})|$$

$$
\begin{aligned}
&= \max_i \sum_{j=1}^n |\hat{c}_{ij}\,(I(|\hat{c}_{ij}| > \tau_{nJ}, |c_{ij}| \le \tau_{nJ}) + I(|\hat{c}_{ij}| > \tau_{nJ}, |c_{ij}| > \tau_{nJ})) \\
&\quad -c_{ij}\,(I(|c_{ij}| > \tau_{nJ}, |\hat{c}_{ij}| > \tau_{nJ}) + I(|c_{ij}| > \tau_{nJ}, |\hat{c}_{ij}| \le \tau_{nJ}))\,| \\
&\le \text{I} + \text{II} + \text{III},
\end{aligned}
$$

where

$$
\begin{aligned}
\text{I} &= \max_i \sum_{j=1}^n |\hat{c}_{ij} - c_{ij}| I(|\hat{c}_{ij}| > \tau_{nJ}, |c_{ij}| > \tau_{nJ}), \\
\text{II} &= \max_i \sum_{j=1}^n |\hat{c}_{ij}| I(|\hat{c}_{ij}| > \tau_{nJ}, |c_{ij}| \le \tau_{nJ}), \\
\text{III} &= \max_i \sum_{J} |c_{ij}| I(|c_{ij}| > \tau_{nJ}, |\hat{c}_{ij}| \le \tau_{nJ}).
\end{aligned}
$$

We bound the above three terms as follows.

$$
\begin{aligned}
\text{I} &\le \max_{i,j} |\hat{c}_{ij} - c_{ij}| \max_i \sum_{j=1}^n I(|c_{ij}| > 0) \\
&= O_p\left(m_n \sqrt{\log(n)/J}\right). \qquad (7.11)
\end{aligned}
$$

For $\delta > 0$, using the equality in (i), we have

$$
\begin{aligned}
\text{II} &\le \max_i \sum_{j=1}^n |\hat{c}_{ij} - c_{ij}| I(|\hat{c}_{ij}| > \tau_{nJ}, |c_{ij}| \le \tau_{nJ}) \\
&\quad + \max_i \sum_{j=1}^n |c_{ij}| I(|c_{ij}| \le \tau_{nJ}) \\
&\le \max_i \sum_{j=1}^n |\hat{c}_{ij} - c_{ij}| I(|\hat{c}_{ij}| > \tau_{nJ}, |c_{ij}| \le \delta\tau_{nJ}) \\
&\quad + \max_i \sum_{j=1}^n |\hat{c}_{ij} - c_{ij}| I(|\hat{c}_{ij}| > \tau_{nJ}, \delta\tau_{nJ} < |c_{ij}| < \tau_{nJ}) + \tau_{nJ} m_n \\
&\le \max_{i,j} |\hat{c}_{ij} - c_{ij}| \left(\max_i \sum_{j=1}^n I(|\hat{c}_{ij}| > \tau_{nJ}, |c_{ij}| \le \delta\tau_{nJ}) + m_n \right) + \tau_{nJ} m_n \\
&\le O_p(\sqrt{\log(n)/J}) \left(\max_i \sum_{j=1}^n I(|\hat{c}_{ij} - c_{ij}| \ge (1-\delta)\tau_{nJ}) + m_n \right) + \tau_{nJ} m_n \\
&= O_p(\sqrt{\log(n)/J})(o_p(1) + m_n) + \tau_{nJ} m_n = O_p(\tau_{nJ} m_n), \qquad (7.12)
\end{aligned}
$$

since

$$
P\left(\max_i \sum_{j=1}^n I(|\hat{c}_{ij} - c_{ij}| \ge (1-\delta)\tau_{nJ}) > \epsilon \right) \le P\left(\max_{i,j} |\hat{c}_{ij} - c_{i,j}| \ge (1-\delta)\tau_{nJ} \right)
$$

$$= \quad o(1).$$

Similarly,

$$\begin{aligned}
\text{III} &\leq \max_i \sum_{j=1}^{n} \left(|\hat{c}_{ij} - c_{ij}| + |\hat{c}_{ij}||\right) I(|c_{ij}| > \tau_{nJ}, |\hat{c}_{ij}| \leq \tau_{nJ}) \\
&\leq \max_{i,j} |\hat{c}_{ij} - c_{ij}| \sum_{j=1}^{n} I(|c_{ij}| > \tau_{nJ}) + \tau_{nJ} \max_i \sum_{j=1}^{n} I(|c_{ij}| > \tau_{nJ}) \\
&\leq O_p(\tau_{nJ}) m_n + \tau_{nJ} m_n = O_p(\tau_{nJ} m_n).
\end{aligned}$$

Combining this with (7.11), (7.12) and (7.10), we obtain the desired result in (ii). The proof is completed.

Lemma 12 *Under Conditions (A1') and (A3)∼(A4), if* $\tau_{nJ} = O(\sqrt{\log(n)/J})$ *and* $n\tau_{nJ} = o(1)$ *as* $n \to \infty$ *and* $J \to \infty$, *then*

(i) $||\hat{C}(\tau_{nJ})^{-1} - C^{-1}|| = O_p(m_n \tau_{nJ})$ *and* $||\hat{C}(\tau_{nJ})^{-2} - C^{-2}|| = O_p(m_n \tau_{nJ})$,

(ii) $||\hat{C}(0)^{-1} - C^{-1}|| \leq O_p(\tau_{nJ}(m_n + n))$; $||\hat{C}(0)^{-2} - C^{-2}|| \leq O_p(\tau_{nJ}(m_n + n))$,

where $m_n = \max_{1 \leq i \leq n} \sum_{j=1}^{n} I(c_{ij} \neq 0) \leq n$.

Proof. Let $\kappa_3 = \max\{2(2/\kappa_1 + 1/\kappa_2) - 1, (4/3)(1/\kappa_1 + 1/\kappa_2) - 1/3, 1\}$. Then $n\sqrt{\log(n)/J}$ $= o(1)$ yields $(\log(n))^{\kappa_3}/J = o(1)$. If let $\lambda_{\min}(C)$ denote the minimum eigenvalue of C, then we have that $\lambda_{\min}(C) \geq \sigma_0^2$. If let $\lambda_{\min}(\hat{C}(\tau_{nJ}))$ denote the minimum eigenvalue of $\hat{C}(\tau_{nJ})$, then it follows from Lemma 11 that

$$\begin{aligned}
\lambda_{\min}(\hat{C}(\tau_{nJ})) &= \lambda_{\min}(C) + O_p(m_n \tau_{nJ}) \\
&\geq \sigma_0^2 + O_p(m_n \tau_{nJ}),
\end{aligned}$$

which is bounded below by $\sigma_0^2/2$ if $\tau_{nJ} m_n$ is small enough. Therefore, we have

$$\begin{aligned}
||\hat{C}(\tau_{nJ})^{-1} - C^{-1}|| &= ||\hat{C}(\tau_{nJ})^{-1}(C - \hat{C}(\tau_{nJ}))C^{-1}|| \\
&\leq ||\hat{C}(\tau_{nJ})^{-1}||(||C - \hat{C}(\tau_{nJ})||)||C^{-1}|| \\
&\leq \lambda_{\min}(\hat{C}(\tau_{nJ}))^{-1} \lambda_{\min}(C)^{-1} ||C - \hat{C}(\tau_{nJ})|| = O_p(\tau_{nJ} m_n). \\
||\hat{C}(\tau_{nJ})^{-2} - C^{-2}|| &\leq ||\hat{C}(\tau_{nJ})^{-1}(\hat{C}(\tau_{nJ})^{-1} - C^{-1})|| + ||(\hat{C}(\tau_{nJ})^{-1} - C^{-1})C^{-1}|| \\
&\leq ||\hat{C}(\tau_{nJ})^{-1}||||\hat{C}(\tau_{nJ})^{-1} - C^{-1}|| + ||\hat{C}(\tau_{nJ})^{-1} - C^{-1}||||C^{-1}|| \\
&= (\lambda_{\min}(\hat{C}(\tau_{nJ}))^{-1} + \lambda_{\min}(C)^{-1})||\hat{C}(\tau_{nJ})^{-1} - C^{-1}|| \\
&= O_p(\tau_{nJ} m_n). \\
||\hat{C}(0)^{-1} - C^{-1}|| &\leq O_p(\tau_{nJ}(m_n + n)), \\
||\hat{C}(0)^{-2} - C^{-2}|| &\leq O_p(\tau_{nJ}(m_n + n)).
\end{aligned}$$

Proof of Theorem 6. Note that for any x, $H_x^T H_x = n$. We have

$$\begin{aligned}
&|| \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j\right]^{-1} - \left[H_j^T C^{-1} H_j\right]^{-1} || \\
&= || \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j\right]^{-1} \left(H_j^T C^{-1} H_j - H_j^T \hat{C}(\tau_{nJ})^{-1} H_j\right) \left[H_j^T C^{-1} H_j\right]^{-1} || \\
&\leq \frac{1}{n} || \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j/n\right]^{-1} ||||H_j^T C^{-1} H_j/n - H_j^T \hat{C}(\tau_{nJ})^{-1} H_j/n|| \\
&\quad \times || \left[H_j^T C^{-1} H_j/n\right]^{-1} || \\
&\leq \frac{1}{n} || \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j/n\right]^{-1} ||||C^{-1} - \hat{C}(\tau_{nJ})^{-1}|||| \left[H_j^T C^{-1} H_j/n\right]^{-1} ||,
\end{aligned}$$

which combining with Lemma 11 yields

$$\begin{aligned}
&|| \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j\right]^{-1} - \left[H_j^T C^{-1} H_j\right]^{-1} || \\
&\leq O_p(n^2 \sqrt{\log(n)/J}) || \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j\right]^{-1} |||| \left[H_j^T C^{-1} H_j\right]^{-1} ||. \qquad (7.13)
\end{aligned}$$

Let λ_m and $\hat{\lambda}_m$ denote the smallest eigenvalues of $H_j^T C^{-1} H_j$ and $H_j^T \hat{C}(\tau_{nJ})^{-1} H_j$ respectively. Invoking Theorem 1, $(H_j^T C^{-1} H_j)^{-1} = \Sigma_j + o(1)$. There exists a positive constant ϵ_0 such that for large n, $\lambda_m \geq \epsilon_0$. By the definition, there exists $a_m \in \mathbb{R}^3$ with $||a_m|| = 1$, such that $\hat{\lambda}_m = a_m^T H_j^T \hat{C}(\tau_{nJ})^{-1} H_j a_m$. So

$$\begin{aligned}
|\hat{\lambda}_m - a_m^T H_j^T C^{-1} H_j a_m| &= |(H_i a_m)^T (\hat{C}^{-1} - C^{-1}) H_i a_m| \leq n||\hat{C}(\tau_{nJ})^{-1} - C^{-1}|| \\
&\leq O_p(n^2 \sqrt{\log(n)/J}),
\end{aligned}$$

which implies

$$\begin{aligned}
\hat{\lambda}_m &\geq a_m^T H_j^T C^{-1} H_j a_m - O_p(n^2 \sqrt{\log(n)/J}) \\
&\geq \lambda_m - O_p(n^2 \sqrt{\log(n)/J}) \geq \epsilon_0 - O_p(n^2 \sqrt{\log(n)/J}).
\end{aligned}$$

This shows that for large n, $\hat{\lambda}_m$ is bounded below from zero. Consequently, we have

$$|| \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j\right]^{-1} || = O(1), \quad || \left[H_j^T C^{-1} H_j\right]^{-1} || = O(1).$$

This together with (7.13) proves that

$$|| \left[H_j^T \hat{C}(\tau_{nJ})^{-1} H_j\right]^{-1} - \left[H_j^T C^{-1} H_j\right]^{-1} || = O_p(n^2 \sqrt{\log(n)/J}).$$

The proof is completed.

Proof of Corollary 7. It follows from Theorem 6 directly. The details are omitted.

Acknowledgments

We thank Professor Richard Henson from MRC Cognition and Brain Sciences Unit, Cambridge for sharing his MEG data with us. The software SPM8 is available at `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`. We also thank the Editor and three anonymous reviewers for their helpful comments.

References

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

A. Bolstad, B. Van Veen, and R. Nowak. Space–time event sparse penalization for magneto-/electroencephalography. *NeuroImage*, 46(4):1066–1081, 2009.

M. J. Brookes, J. Vrba, S. E. Robinson, C. M. Stevenson, A. M. Peters, G. R. Barnes, A. Hillebrand, and P. G. Morris. Optimising experimental design for meg beamformer imaging. *Neuroimage*, 39(4):1788–1802, 2008.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

J. Dedecker and C. Prieur. Coupling for τ-dependent sequences and applications. *Journal of Theoretical Probability*, 17(4):861–885, 2004.

J. Fan, Y. Liao, and M. Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320, 2011.

K. Friston, R. Henson, C. Phillips, and J. Mattout. Bayesian estimation of evoked and induced responses. *Human brain mapping*, 27(9):722–735, 2006.

J. H. Goodnight. A tutorial on the sweep operator. *The American Statistician*, 33(3): 149–158, 1979.

M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

R. N. Henson, D. G. Wakeman, V. Litvak, and K. J. Friston. A parametric empirical bayesian framework for the eeg/meg inverse problem: generative models for multi-subject and multi-modal integration. *Frontiers in human neuroscience*, 5, 2011.

A. Hillebrand, K. D. Singh, I. E. Holliday, P. L. Furlong, and G. R. Barnes. A new approach to neuroimaging with magnetoencephalography. *Human brain mapping*, 25(2):199–211, 2005.

M. X. Huang, J. J. Shih, R. R. Lee, D. L. Harrington, R. J. Thoma, M. P. Weisend, F. Hanlon, K. M. Paulson, T. Li, K. Martin, et al. Commonalities and differences among vectorized beamformers in electromagnetic source imaging. *Brain topography*, 16(3):139–158, 2004.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11): 4302–4311, 1997.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

F. Merlevède, M. Peligrad, and E. Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474, 2011.

J. C. Mosher, R. M. Leahy, and P. S. Lewis. Eeg and meg: forward solutions for inverse methods. *Biomedical Engineering, IEEE Transactions on*, 46(3):245–259, 1999.

R. Oostenveld, P. Fries, E. Maris, and J. Schoffelen. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011, 2010.

D. Pitcher, D. D. Dilks, R. R. Saxe, C. Triantafyllou, and N. Kanwisher. Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage*, 56(4):2356–2363, 2011.

M. A. Quraan, S. N. Moses, Y. Hung, T. Mills, and M. J. Taylor. Detection and localization of hippocampal activity using beamformers with meg: a detailed investigation using simulations and empirical data. *Human brain mapping*, 32(5):812–827, 2011.

J. O. Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

S. E. Robinson. Functional neuroimaging by synthetic aperture magnetometry (sam). *Recent advances in biomagnetism*, pages 302–305, 1999.

A. Rodríguez-Rivera, B. V. Baryshnikov, B. D. Van Veen, and R. T. Wakai. Meg and eeg source localization in beamspace. *Biomedical Engineering, IEEE Transactions on*, 53(3): 430–441, 2006.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

J. Sarvas. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Physics in medicine and biology*, 32(1):11, 1987.

K. Sekihara and S. S. Nagarajan. *Adaptive spatial filters for electromagnetic brain imaging*. Springer Science & Business Media, 2008.

K. Sekihara, S. S. Nagarajan, D. Poeppel, and A. Marantz. Asymptotic snr of scalar and vector minimum-variance beamformers for neuromagnetic source reconstruction. *Biomedical Engineering, IEEE Transactions on*, 51(10):1726–1734, 2004.

B. D. Van Veen, W. Van Drongelen, M. Yuchtman, and A. Suzuki. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *Biomedical Engineering, IEEE Transactions on*, 44(9):867–880, 1997.

J. Zhang, C. Liu, and G. Green. Source localization with meg data: A beamforming approach based on covariance thresholding. *Biometrics*, 70(1):121–131, 2014.

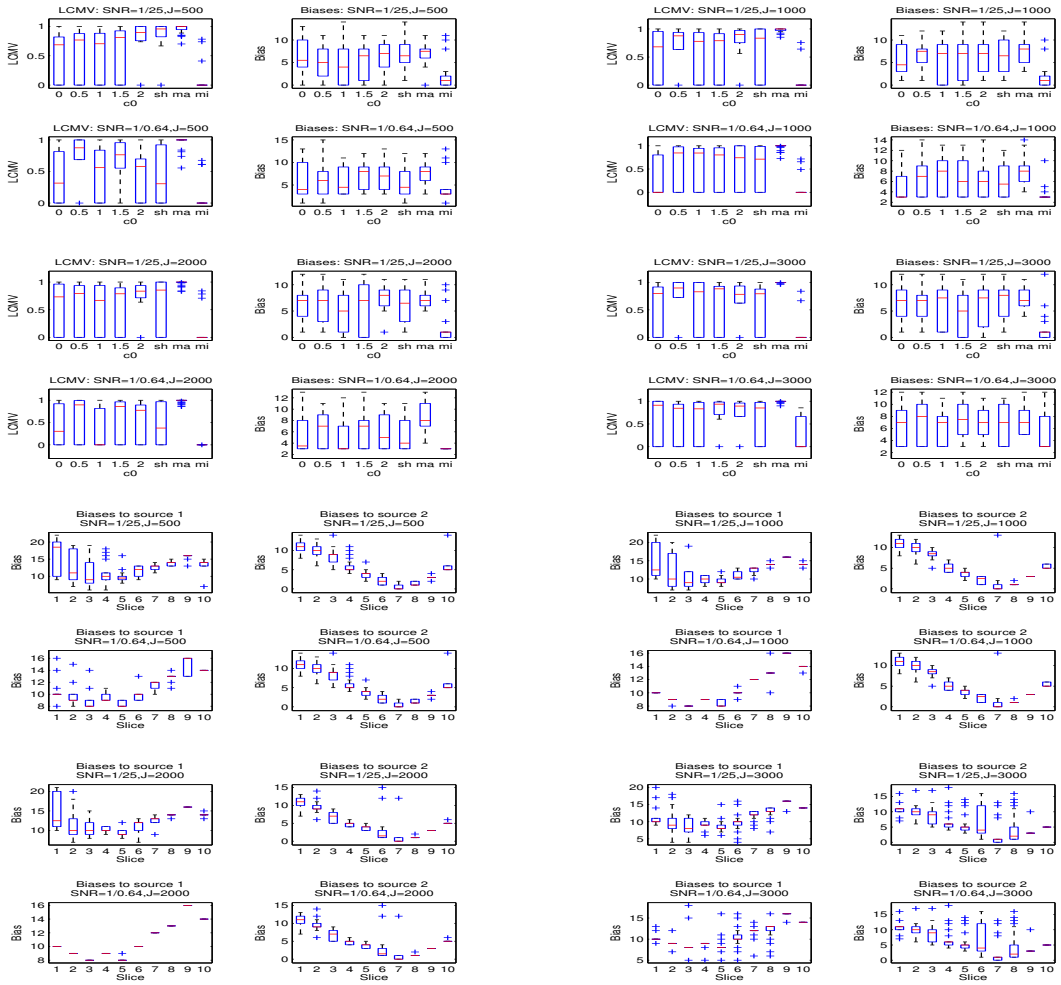

Figure 2: Scenario 1: Two sources located at CTF coordinates $(3, -1, 4)^T$ cm and $(-5, 2, 6)^T$ cm respectively. The first four rows display the box-and-whisker plots of the index values and the localization biases against the tuning constant $c_0 = 0, 0.5, 1, 1.5, 2$, **ma**, **mi** and **sh** for the combinations of $n = 91$, SNR$= 1/25, 1/0.64$, and $J = 500, 1000, 2000, 3000$ respectively. Here, **ma** and **mi** stand for the proposed hard-thresholded covariance based methods. **sh** stands for the optimal shrinkage-based method. With a slightly abuse of notation, $c_0 =$ **ma**, **mi**, **sh** refer to that **ma**, **mi**, and **sh** are used. The remaining rows present the box-and-whisker plots of the local localization bias to the sources r_1 and r_2 against the transverse slice indices from 0 to 10 when c_0 was selected by the minimum strategy for the above combinations respectively. The red colored lines in the boxes are the medians. Note that when the distribution of the localization biases are degenerate, the upper and lower quartiles and medians of localization biases will be equal. Consequently, the box in the plot will reduce to a red colored line. The plots in the last four rows show that all the local peaks on the transverse slices are not close to the source location r_1, implying that the source 1 has been masked on the neuronal activity map by source cancellations.

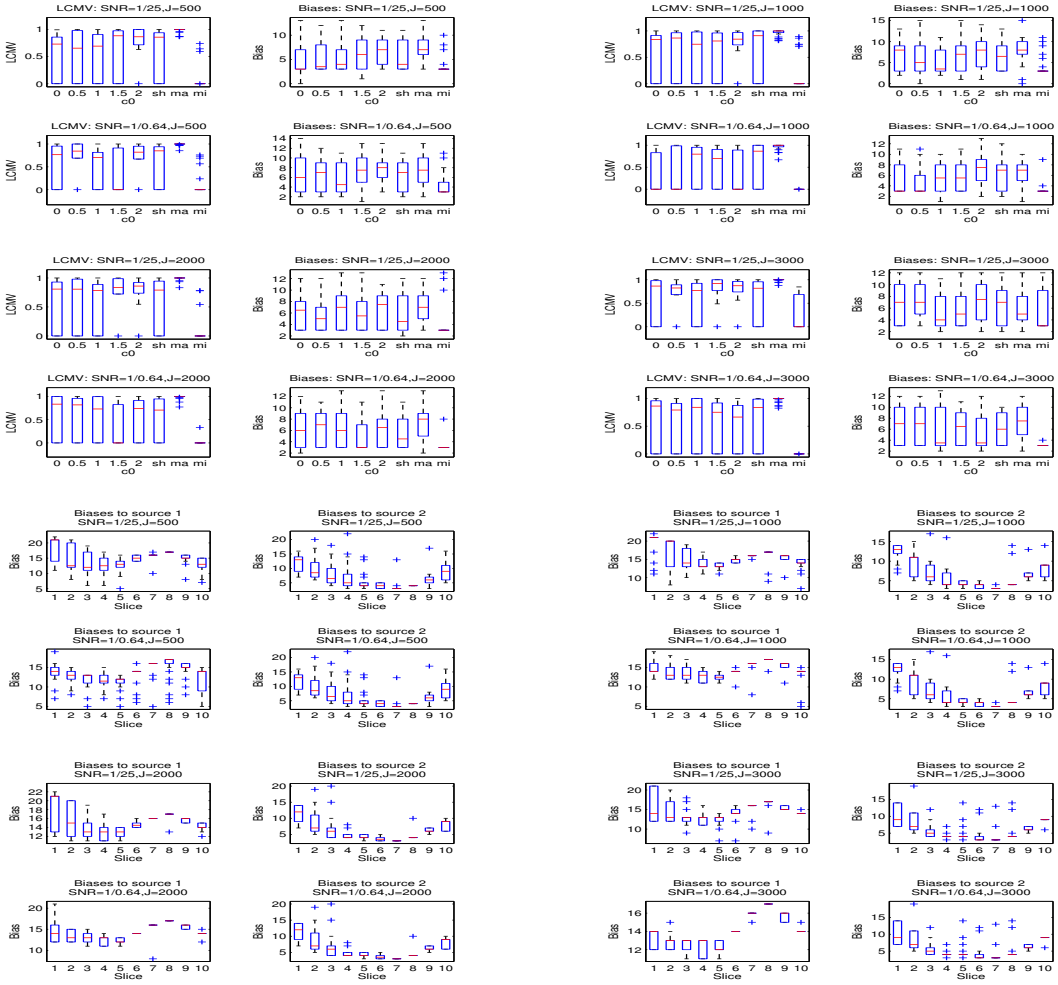

Figure 3: Scenario 2: Two sources located at CTF coordinates $(-5, 5, 6)^T$ cm and $(-6, -2, 5)^T$ cm respectively. The first four rows display the box-and-whisker plots of the index values and the localization biases against the tuning constant $c_0 = 0, 0.5, 1, 1.5, 2$, **ma**, **mi** and **sh** for the combinations of $n = 91$, SNR$= 1/25, 1/0.64$, and $J = 500, 1000, 2000, 3000$ respectively. The remaining rows present the box-and-whisker plots of the minimum local localization bias to the sources r_1 and r_2 against the transverse slice indices from 0 to 10 when c_0 is selected by the minimum strategy for the above combinations respectively. The red colored lines in the boxes are the medians. When the upper and lower quartiles and medians of localization biases have the same value, the box in the plot will reduce to a red colored line. The plots in the last four rows show that all the local peaks on the transverse slices are not close to the source location r_1, implying the source 1 has been masked on the neuronal activity map by source cancellations.

Figure 4: Scenario 3: Two sources located at CTF coordinates $(3, -1, 4)^T$ cm and $(-5, 2, 6)^T$ cm respectively. The first six rows show the box-and-whisker plots of the index values and the localization biases against the tuning constant $c_0 = 0, 0.5, 1, 1.5, 2$, **ma**, **mi** and **sh** for the combinations of $n = 102$ sensors, SNR$= 1/0.35^2, 1/0.4^2, 1/0.5^2$, and the sample rates $J = 500, 1000, 2000, 3000$ respectively. The last six rows give the box-and-whisker plots of the minimum local localization bias to the sources r_1 and r_2 against the transverse slice indices from 0 to 10 when c_0 is selected by the minimum strategy for these combinations respectively. The red colored lines in the boxes are the medians. When the upper and lower quartiles and medians of localization biases are equal, the box in the plot will reduce to a red colored line. The last six rows of the plots show all the local peaks on the transverse slices are not close to the source location r_1, implying the source 1 has been masked on the neuronal activity map by source cancellations.

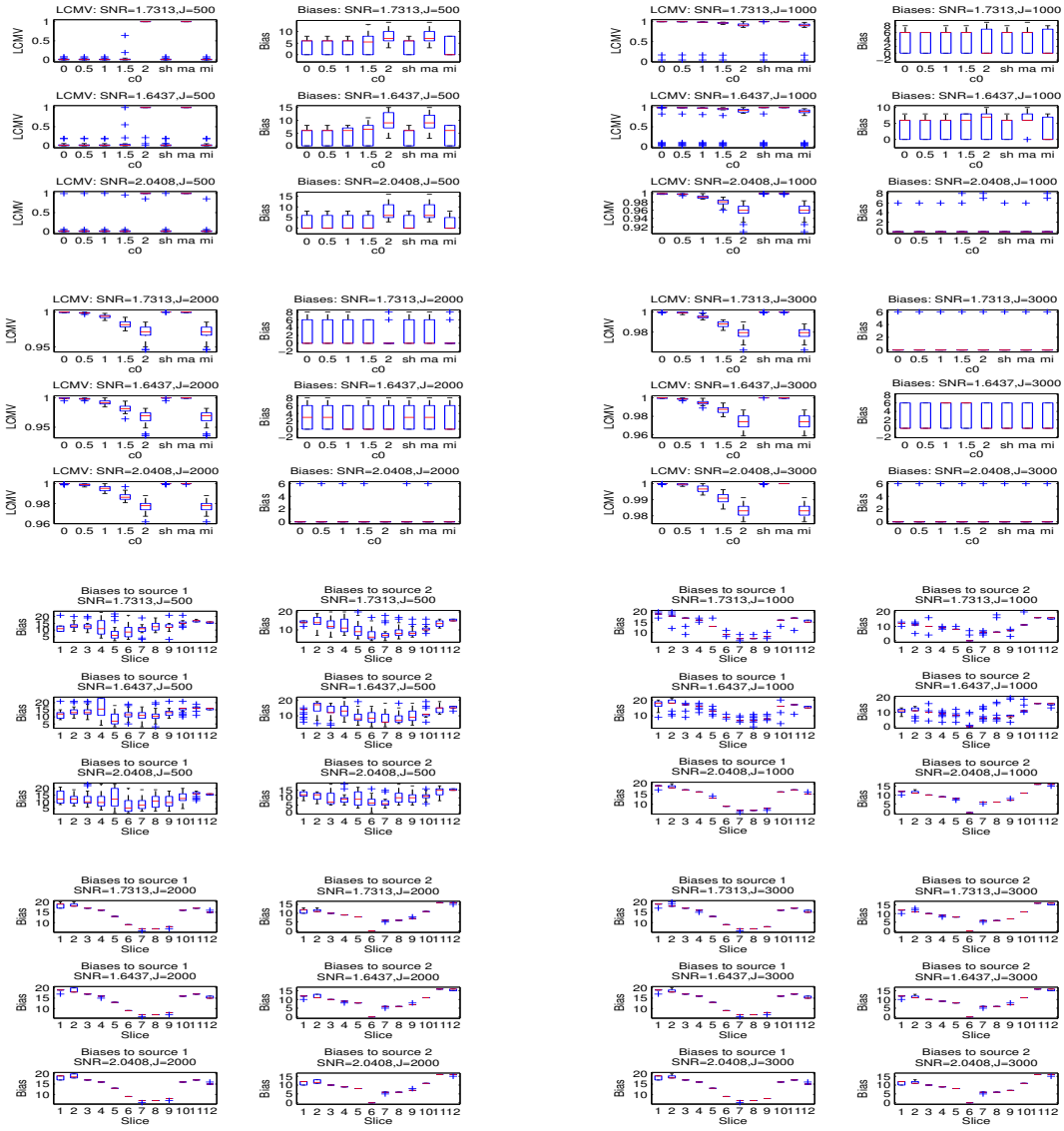

Figure 5: Scenario 4: Two sources located at CTF coordinates $(-5, 5, 6)^T$ cm and $(-6, -2, 5)^T$ cm respectively. The first six rows show the box-and-whisker plots of the index values and the localization biases against the tuning constant $c_0 = 0, 0.5, 1, 1.5, 2$, **ma**, **mi** and **sh** for the combinations of $n = 102$ sensors, SNR$= 1/0.35^2, 1/0.4^2, 1/0.5^2$, and the sample rates $J = 500, 1000, 2000, 3000$ respectively. The last six rows give the box-and-whisker plots of the minimum local localization bias to the sources r_1 and r_2 against the transverse slice indices from 0 to 10 when c_0 was selected by the minimum strategy for these combinations respectively. The red colored lines in the boxes are the medians. When the upper and lower quartiles and medians of localization biases are equal, the box in the plot will reduce to a red colored line. The last six rows of the plots show all the local peaks on the transverse slices are not close to the source location r_1, implying the source 1 has been masked on the neuronal activity map by source cancellations.

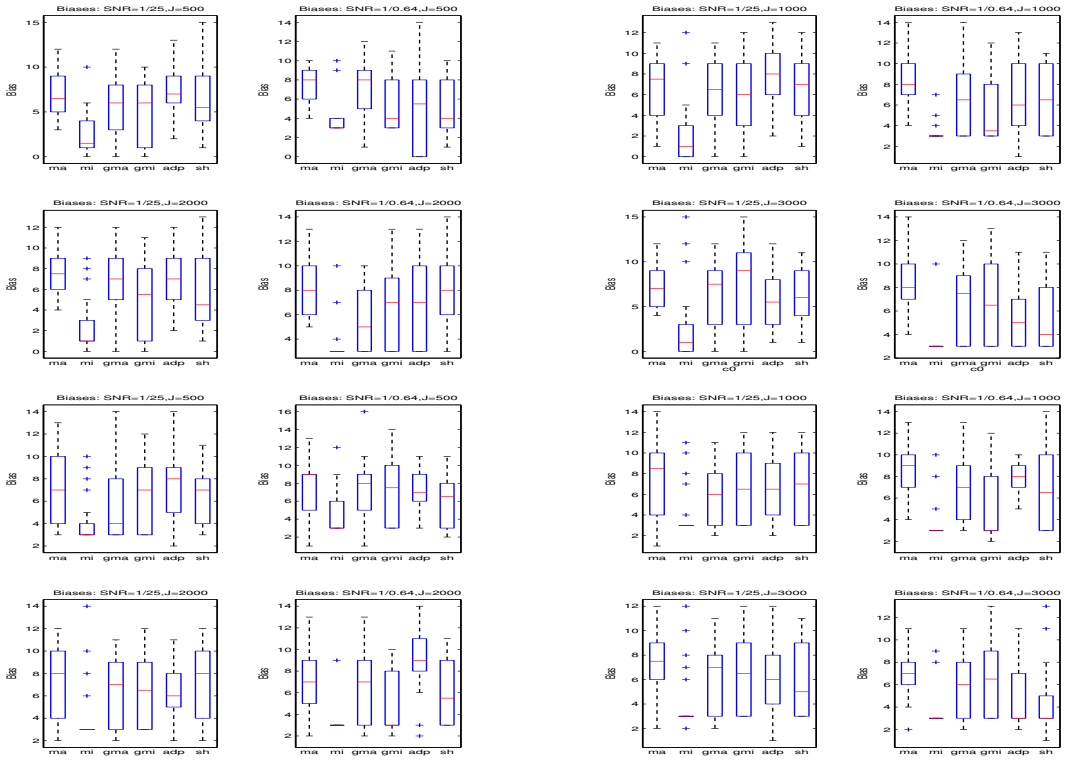

Figure 6: Performance comparison of the six different beamformers, namely **ma**, **mi**, **gma**, **gmi**, **adp** and **sh** in Scenarios 1 and 2. Here, **ma** and **mi** stand for the hard-thresholded covariance based methods when the tuning constant c_0 is chosen by use of the maximum strategy and the minimum strategy respectively; **gma** and **gmi** stand for the generalized thresholded covariance based methods when the tuning constant c_0 is chosen by use of the maximum strategy and the minimum strategy respectively; **adp** and **sh** stand for the adaptive thresholding-based method and the optimal shrinkage-based method. The upper two rows of multiple box-whisker plots are for the combinations of $n = 91$, SNR$= 1/25, 1/0.64$, and $J = 500, 1000, 2000, 3000$ in Scenario 1, while the lower two rows are for the combinations of $n = 91$, SNR$= 1/25, 1/0.64$, and $J = 500, 1000, 2000, 3000$ in Scenario 2. Each panel shows the localization biases against the above six different beamformer methods. The red colored lines in the boxes are the medians. When the upper and lower quartiles and medians of localization biases are equal, the box in the plot will reduce to a red colored line.

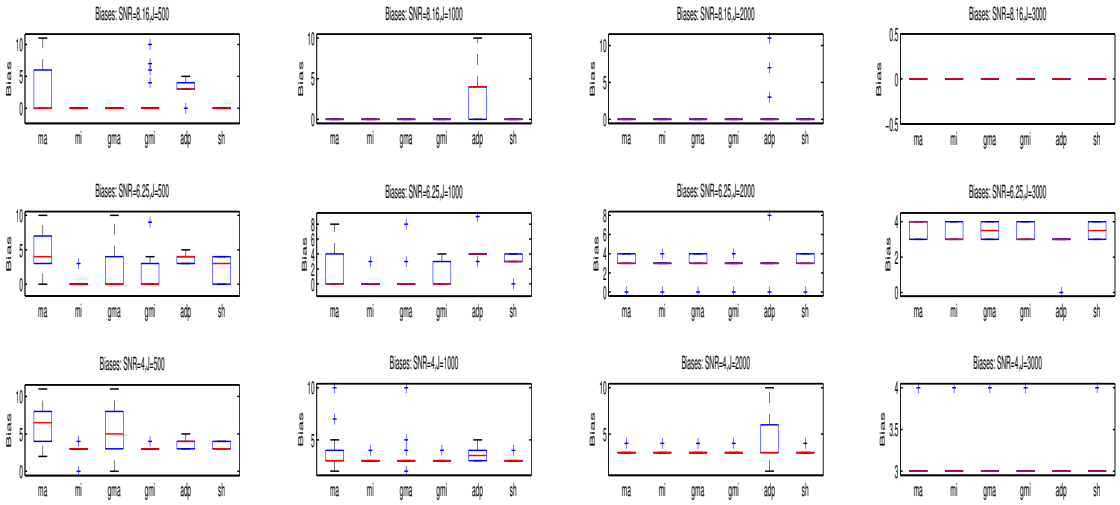

Figure 7: Performance comparison of the six different beamformers, namely **ma**, **mi**, **gma**, **gmi**, **adp** and **sh** in Scenario 3. Multiple box-whisker plots of localization biases are displayed for the combinations of $n = 102$, SNR$= 1/0.35^2$, $1/0.4^2$, $1/0.5^2$, and $J = 500, 1000, 2000, 3000$. Each panel shows the localization biases against the six different beamformer methods, namely **ma**, **mi**, **gma**, **gmi**, **adp** and **sh**. The red colored lines in the boxes are the medians. When the upper and lower quartiles and medians of localization biases are equal, the box in the plot will reduce to a red colored line.

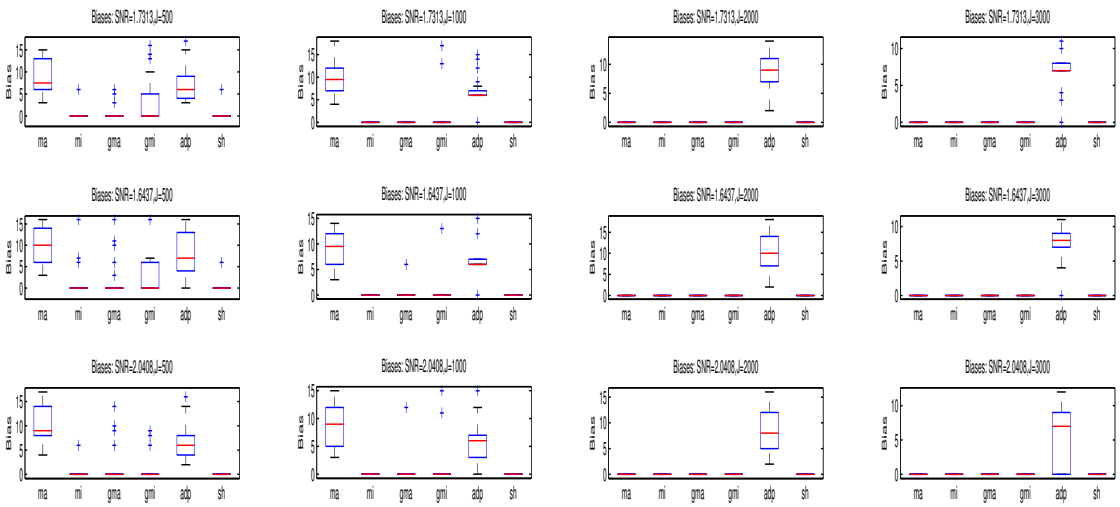

Figure 8: Performance comparison of the six different beamformers, namely **ma**, **mi**, **gma**, **gmi**, **adp** and **sh** in Scenario 4. Multiple box-whisker plots of localization biases are displayed for the combinations of $n = 102$, SNR$= 1/0.35^2$, $1/0.4^2$, $1/0.5^2$, and $J = 500, 1000, 2000, 3000$ in Scenario 4. Each panel shows the localization biases against the six different beamformer methods, namely **ma**, **mi**, **gma**, **gmi**, **adp** and **sh**. The red colored lines in the boxes are the medians. When the upper and lower quartiles and medians of localization biases are equal, the box in the plot will reduce to a red colored line.

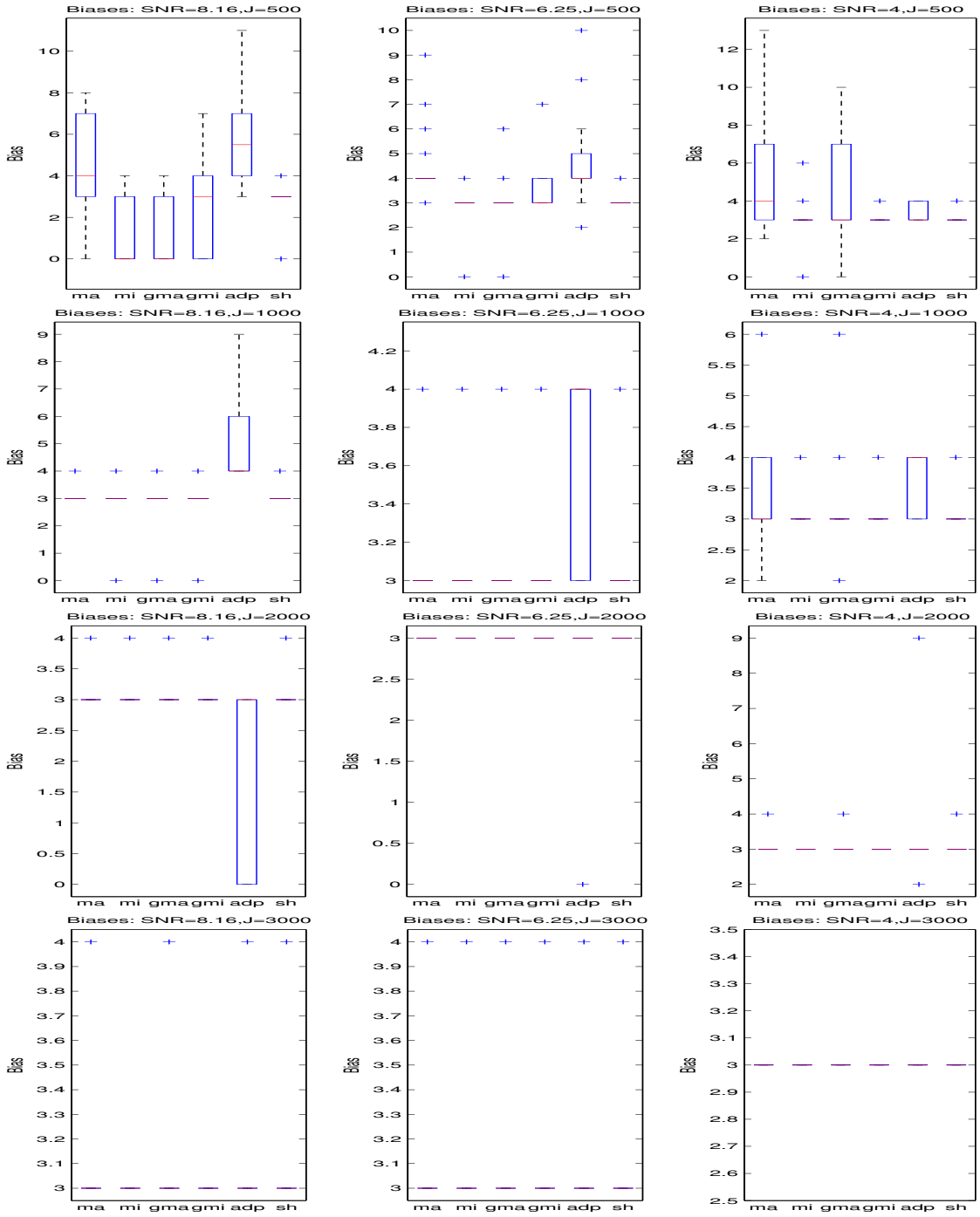

Figure 9: Performance comparison of the six different beamformers, namely **ma**, **mi**, **gma**, **gmi**, **adp** and **sh** in Scenario 5. Multiple box-whisker plots of localization biases are displayed for the combinations of $n = 102$, SNR$= 1/0.35^2$, $1/0.4^2$, $1/0.5^2$, and $J = 500, 1000, 2000, 3000$ respectively. Each panel shows the localization biases against the six different beamformer methods, namely **ma**, **mi**, **gma**, **gmi**, **adp** and **sh**.

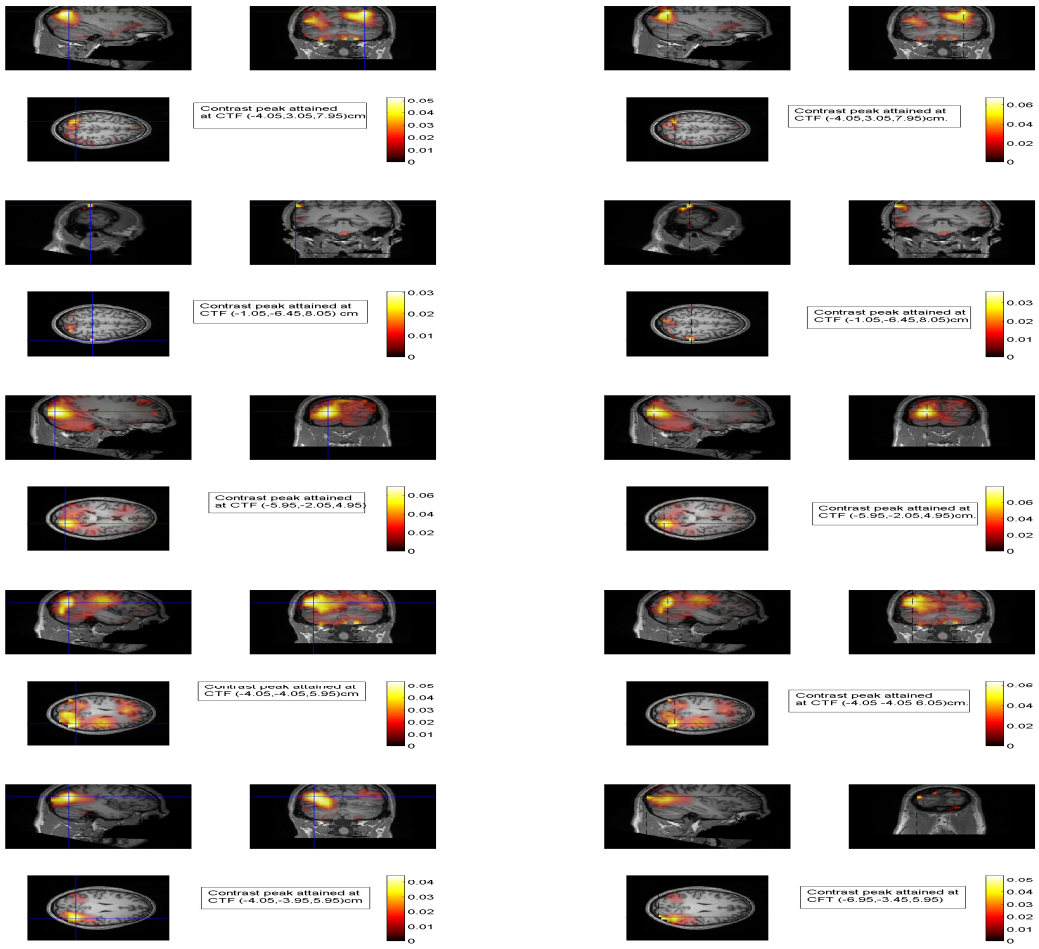

Figure 10: Plots of the log-contrasts between the faces and scrambled faces on three orthogonal slices through the peak locations for each of five sessions, which are overlaid on the subject's MRI scan. The plots in the left-hand two columns and the right-hand two columns are derived from the procedures **mi** and **adp** respectively. Rows 1 and 2, 3 and 4, 5 and 6, and 7 and 8 are for sessions $1 \sim 5$ respectively. The highlighted yellow colored areas revealed neuronal activity increases or decreases for the faces relative to the scrambled faces. The areas shown in the left-hand two columns are in or close to the IOG, STS, and PCu regions which are known to be related to the human face perception.

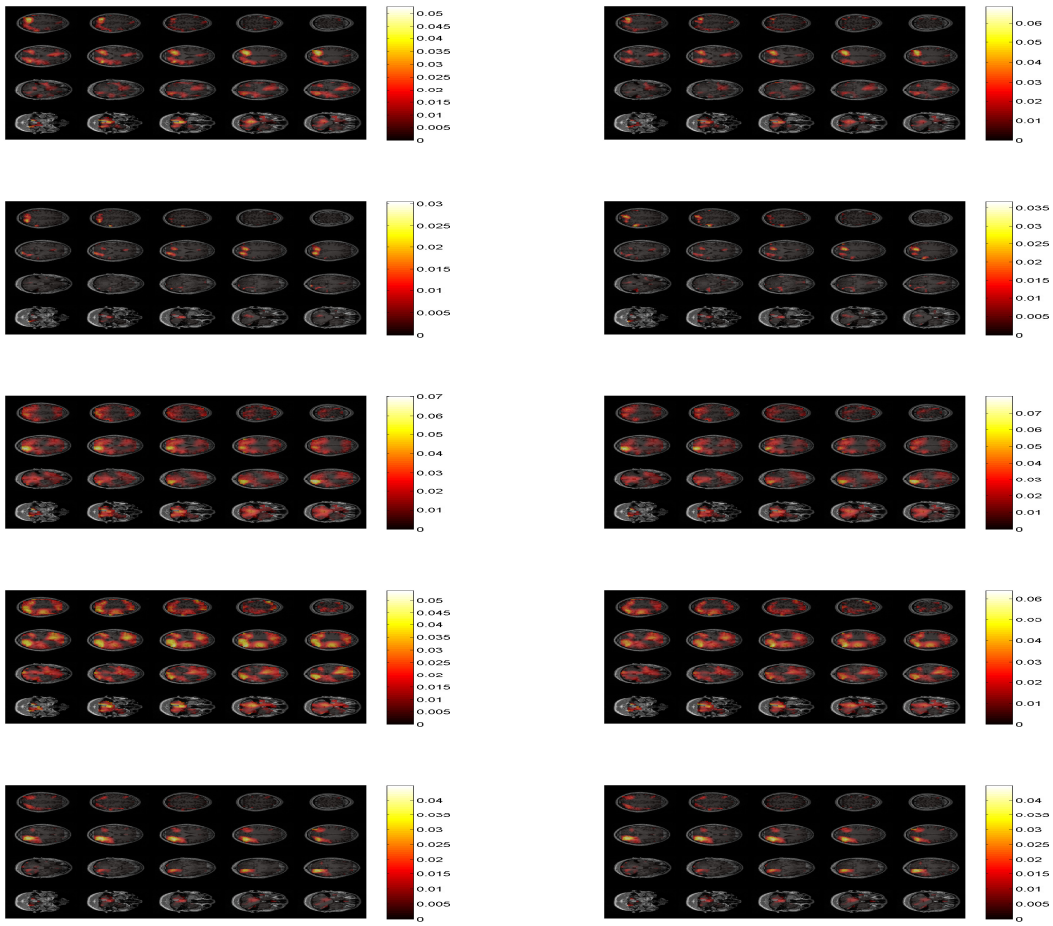

Figure 11: Plots of the log-contrasts between the faces and scrambled faces on 20 transverse slices for each of five sessions, which are overlaid on the subject's MRI scan. The plots in the left-hand column and the right-hand column are derived from the procedures **mi** and **adp** respectively. Rows $1 \sim 5$ are for sessions $1 \sim 5$ respectively. The highlighted yellow colored areas revealed neuronal activity increases for the faces relative to the scrambled faces. The areas highlighted in the first column are in or close to the OFA, IOG, STS, and PCu regions which are known to be related to the human face perception.

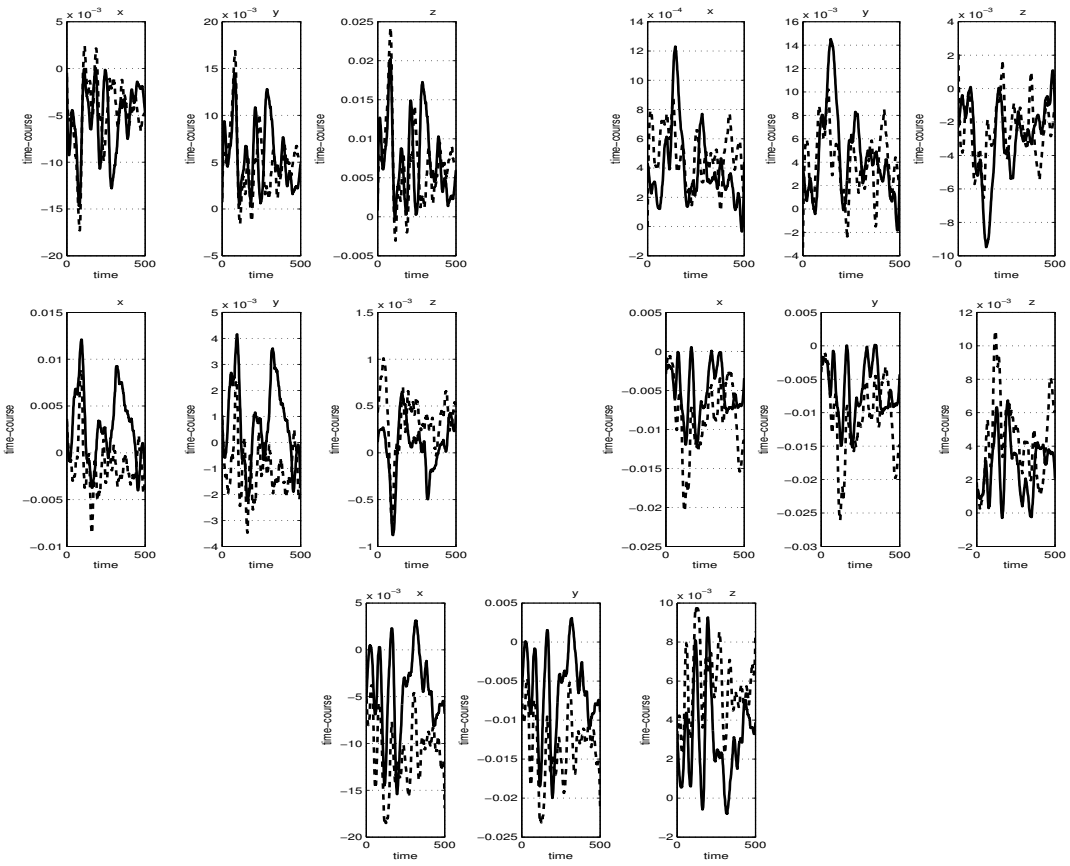

Figure 12: Plots of the estimated time-courses at the global peaks along x, y and z-axes respectively for each of five sessions. The solid curve and the dashed curve in each plot stand for the estimated time-courses under the faces and the scrambled faces respectively. The plots are ordered from the top left panel to the right panel to the bottom panel corresponding to sessions $1 \sim 5$.

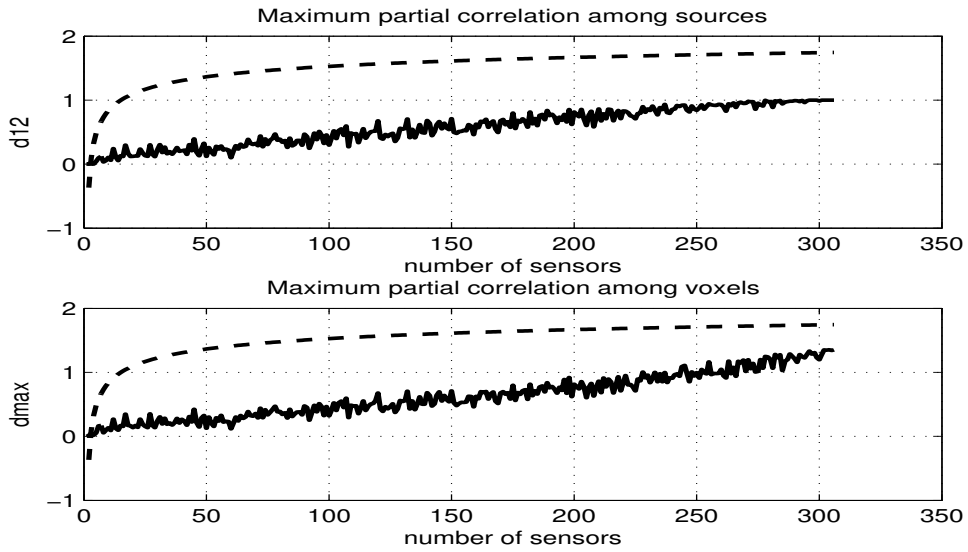

Figure 13: Plots of $d_{12}(k)$ and $d_{\max}(k)$ against $k = 1, 2, ..., 306$ respectively, where k stands for k randomly chosen sensors from the 306 sensors in the face-perception data, two sources are located at CTF (-4, 3,8)cm and (-4,-5,5) cm respectively, and the dashed curve in each plot is for the function $\log(\log(k))$.