Maximum-Likelihood Estimation of Low-Rank Signals for Multiepoch MEG/EEG Analysis

Boris V. Baryshnikov*, Barry D. Van Veen, *Fellow, IEEE*, and Ronald T. Wakai

***Abstract*—A maximum-likelihood-based algorithm is presented for reducing the effects of spatially colored noise in evoked response magneto- and electro-encephalography data. The repeated component of the data, or signal of interest, is modeled as the mean, while the noise is modeled as the Kronecker product of a spatial and a temporal covariance matrix. The temporal covariance matrix is assumed known or estimated prior to the application of the algorithm. The spatial covariance structure is estimated as part of the maximum-likelihood procedure. The mean matrix representing the signal of interest is assumed to be low-rank due to the temporal and spatial structure of the data. The maximum-likelihood estimates of the components of the low-rank signal structure are derived in order to estimate the signal component. The relationship between this approach and principal component analysis (PCA) is explored. In contrast to prestimulus-based whitening followed by PCA, the maximum-likelihood approach does not require signal-free data for noise whitening. Consequently, the maximum-likelihood approach is much more effective with nonstationary noise and produces better quality whitening for a given data record length. The efficacy of this approach is demonstrated using simulated and real MEG data.**

***Index Terms*—Electroencephalography, evoked responses, magnetoencephalography, maximum likelihood parameter estimation, principal component analysis.**

I. Introduction

EVOKED response data in magneto- and electro-encephalography (MEG/EEG) typically has very low signal-to-noise ratio (SNR); the presence and character of the response to a stimulus is often obscured by background noise. Averaging is typically used to improve the SNR. The underlying assumption behind averaging is that the response of interest is the same across trials while the noise is independent. Although the noise is independent from trial to trial, it is not spatially or temporally white [1]. Hence, appropriate space-time processing can exploit the noise coloration to improve the SNR.

We present a maximum-likelihood approach for estimating the repeated component of the evoked response, modeled by a low-rank or structured mean, assuming the spatial component of the noise modeled as Gaussian with zero mean and unknown covariance. Basis vectors for the temporal structure of the mean are determined using knowledge of the frequency band of interest and the assumption that the response of interest is constant across trials. The spatial component of the mean is assumed to be low-rank, but with unknown structure. Thus, the mean of the space (columns) time (rows) signal measured by the detector array takes the form $\mathbf{H}\boldsymbol{\theta}\mathbf{D}^T$, where $\mathbf{D}^T$ is a matrix whose rows contain the known temporal basis vectors, $\mathbf{H}$ is a matrix whose columns contain the unknown spatial basis vectors, and $\boldsymbol{\theta}$ is an unknown matrix of signal amplitude parameters. The signal of interest is estimated from the maximum-likelihood estimates (MLEs) of $\mathbf{H}$ and $\boldsymbol{\theta}$.

The MEG/EEG noise covariance matrix is modeled as the Kronecker product of spatial and temporal covariance matrices as suggested by de Munck, *et al.* [2]. The spatial covariance matrix of the noise is a nuisance parameter that is estimated by the maximum-likelihood procedure. We assume the temporal covariance structure of the noise is either known or may be approximated as independent and identically distributed across time samples. While the independence assumption is effective in many situations, procedures for modeling and estimating the temporal covariance structure from the data [2], [3] may also be employed.

Similar signal and noise models have been used in radar, e.g., [4], and in growth curve models of multivariate statistics, e.g., [5]. Dogandžić and Nehorai [6] adapted these methods to localize dipolar sources in MEG/EEG data. In these applications, the matrix $\mathbf{H}$ is assumed to be known or to depend on a small number of parameters. Our application differs in that we employ the MLE of an unstructured $\mathbf{H}$, analogous to reduced-rank regression as described in [7]. By assuming an unstructured model for $\mathbf{H}$, this approach is not restricted to dipolar source models or even focal activity for describing the signal of interest. Simulated and real adult and fetal data is used to demonstrate the effectiveness of this approach. The relationship between the maximum-likelihood method and principal component analysis (PCA) [8] is also derived in this paper.

One method for dealing with spatially colored noise in MEG/EEG is prewhitening [9], wherein the data is whitened using a noise covariance matrix estimated from a signal-free portion of the recording. This approach not only requires signal-free data, but assumes that the noise statistics are identical in the segments of the data containing the signal and those containing only noise. Our method is not subject to these limitations since we exploit the known temporal structure of the signal to estimate the spatial covariance matrix of the noise using data containing the signal of interest. Another class of approaches for discriminating against spatially colored noise is to use a beamformer [10], [11] to estimate the signal amplitude associated with a specific signal spatial pattern, e.g., a current

Manuscript received December 30, 2003; revised March 31, 2004. This work was supported in part by the National Institutes of Health (NIH) under Grant R01NS037740 and Grant R01HL063174. *Asterisk indicates corresponding author.*

*B. V. Baryshnikov is with the Department of Medical Physics, University of Wisconsin-Madison, 1300 University Ave., MSC 1530, Madison, WI 53706 USA (e-mail: boris@cs.wisc.edu).

B. D. Van Veen is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706 USA. (e-mail: vanveen@engr.wisc.edu).

R. T. Wakai is with the Department of Medical Physics, University of Wisconsin-Madison, Madison, WI 53706 USA (e-mail: rtwakai@wisc.edu).

Digital Object Identifier 10.1109/TBME.2004.834285

dipole pattern, and then reconstruct the signal using the estimated amplitude and known spatial pattern. Such approaches have the disadvantage of assuming specific source models for the evoked responses and do not explicitly exploit the temporal structure of the signal.

Section II of this paper develops the signal model for the MEG/EEG evoked response data. The MLEs of the unknown signal and noise parameters are derived in Section III. The efficacy of the maximum-likelihood approach is demonstrated in Section IV using simulated and real MEG data. Section V presents a discussion of the proposed method. Our notation uses bold lower and uppercase symbols to denote vectors and matrices, respectively.

II. Problem Formulation

In a typical multiepoch MEG or EEG experiment, multiple observations or epochs of a spatio-temporal series are obtained. The signal of interest is normally considered to be the repeatable component of each epoch, while the nonrepeatable component is noise. Assuming N spatial channels and T time samples, the data for the jth epoch is represented by the $N \times T$ matrix $\mathbf{X}_j = \mathbf{S} + \mathbf{N}_j$, where $\mathbf{S}$ is the $N \times T$ matrix of signal samples and $\mathbf{N}_j$ is $N \times T$ matrix of noise samples in the jth epoch. During the experiment, J data epochs are collected.

We assume that the signal time series in each channel lies in the space spanned by the columns of the known $T \times L$ matrix $\mathbf{C}$, where L is the number of temporal basis vectors. This implies that the rows of $\mathbf{S}$ lie in the space spanned by $\mathbf{C}^T$. In this paper, we use the frequency band of interest to identify the set of temporal basis vectors in $\mathbf{C}$. The details of the method used to construct the temporal basis matrix $\mathbf{C}$ are given in the Appendix. Information other than the frequency band may also be used to identify $\mathbf{C}$. For example, if the signal lies within a range of time shifts and scales, then $\mathbf{C}$ can be selected using the appropriate columns of the wavelet transform. Note that if the signal is known to be absent during a portion of the T samples, then the corresponding columns of $\mathbf{C}$ are set to zero and the basis functions designed over the remainder of the interval.

Similarly, let the columns of the unknown $N \times P$ matrix $\mathbf{H}$ be a basis for the P-dimensional space in which the columns of the signal matrix $\mathbf{S}$ lie. If the activity of interest is modeled as an equivalent current dipole in an MEG experiment as in [6], then $\mathbf{H}$ is $N \times 2$ and may be derived from the magnetic field associated with the unit amplitude dipoles in the x, y, and z directions at a given location in the brain.[1] If the activity is dipolar and the moment orientation is fixed over the T time samples, then $\mathbf{H}$ is $N \times 1$. In general, if there are multiple dipolar sources or the activity is not dipolar, then $P > 2$. We assume $\mathbf{H}$ is unknown and do not restrict the structure of $\mathbf{H}$ to correspond to dipolar sources. Thus, the signal matrix can be expressed in the form

$$\mathbf{S} = \mathbf{H}\boldsymbol{\theta}\mathbf{C}^T \tag{1}$$

where $\boldsymbol{\theta}$ is a $P \times L$ matrix of unknown signal amplitude parameters. Note that since $\mathbf{H}$ is unstructured, a unique signal matrix $\mathbf{S}$ does not correspond to unique matrices $\mathbf{H}$ and $\boldsymbol{\theta}$. If $\mathbf{V}$ is a nonsingular $P \times P$ matrix, then $\mathbf{H}\boldsymbol{\theta} = \mathbf{H}'\boldsymbol{\theta}'$, where $\mathbf{H}' = \mathbf{H}\mathbf{V}$ and $\boldsymbol{\theta}' = \mathbf{V}^{-1}\boldsymbol{\theta}$. This ambiguity is inherent to all unstructured low-rank signal models and is generally not of concern since the product $\mathbf{H}\boldsymbol{\theta}$ as well as spaces spanned by the columns of $\mathbf{H}$ and rows of $\boldsymbol{\theta}$ are unique.

Let $\mathbf{X} = [\,\mathbf{X}_1 \;\; \mathbf{X}_2 \;\; \ldots \;\; \mathbf{X}_J\,]$ be the $N \times JT$ data matrix formed from all J epochs, that is

$$\mathbf{X} = [\,\mathbf{S} \;\; \mathbf{S} \;\; \ldots \;\; \mathbf{S}\,] + [\,\mathbf{N}_1 \;\; \mathbf{N}_2 \;\; \ldots \;\; \mathbf{N}_J\,]. \tag{2}$$

We expand the signal component of the full dataset of J epochs in terms of the low-rank model, since the signal of interest $\mathbf{S}$ is assumed to be identical in each epoch

$$[\,\mathbf{S} \;\; \mathbf{S} \;\; \ldots \;\; \mathbf{S}\,] = \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T \tag{3}$$

where the $L \times JT$ matrix $\mathbf{D}^T = [\,\mathbf{C}^T \;\; \mathbf{C}^T \;\; \ldots \;\; \mathbf{C}^T\,]$. Note that the columns of $\mathbf{H}$ represent the spatial structure of the signal while the columns of $\mathbf{D}$ represent the temporal structure.

We assume that the noise matrices $\mathbf{N}_j$, $j = 1, 2, \ldots, J$ are normally distributed, zero mean, independent of one another, with unknown spatial covariance matrix $\mathbf{R}_N$, and temporal covariance matrix $\mathbf{R}_T$. This model implies that the ith column and the kth row of $\mathbf{N}_j$ has covariance matrices $[\mathbf{R}_T]_{ii}\mathbf{R}_N$ and $[\mathbf{R}_N]_{kk}\mathbf{R}_T$ where $[\mathbf{R}_T]_{ii}$ and $[\mathbf{R}_N]_{kk}$ are the ith and kth diagonal elements of $\mathbf{R}_T$ and $\mathbf{R}_N$, respectively. We assume that $\mathbf{R}_T$ is known or estimated (see, e.g., [2] and [3]) prior to estimation of $\mathbf{S}$. The derivation that follows assumes $\mathbf{R}_T = \mathbf{I}$ and is applicable to the case where $\mathbf{R}_T \neq \mathbf{I}$ by transforming $\mathbf{X}_j$ to $\mathbf{X}_j\mathbf{R}_T^{-1/2}$. Hence, the probability density function for the data matrix $\mathbf{X}$ may be written

$$f(\mathbf{X}; \mathbf{R}_N, \mathbf{H}, \boldsymbol{\theta}) = (2\pi)^{-(NTJ/2)}(\det \mathbf{R}_N)^{-(TJ/2)} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{R}_N^{-1}\mathbf{Q}\right)\right\} \tag{4}$$

where $\mathbf{Q} = (\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)^T$.

III. Maximum Likelihood Parameter Estimation

The maximum-likelihood criterion chooses the estimates that are most likely given the data [12]. The MLEs of $\mathbf{H}$ and $\boldsymbol{\theta}$ are obtained by finding $\mathbf{H}$ and $\boldsymbol{\theta}$ that maximize (4) for the given data $\mathbf{X}$. Unfortunately, the noise covariance matrix $\mathbf{R}_N$ is a nuisance parameter that must also be estimated in order to estimate $\mathbf{H}$ and $\boldsymbol{\theta}$. Certain portions of the derivation in Section III-A may be found in the growth curve or statistical signal processing literature (see, e.g., [4]–[7] and references therein). We provide key steps in the derivation to enable interpretation of the results in light of the evoked response paradigm.

A. Derivation of Parameter Estimates

The MLE for $\mathbf{R}_N$ is given as a function of the unknown $\mathbf{H}$ and $\boldsymbol{\theta}$ by [13, Theorem 3.1.5]

$$\hat{\mathbf{R}}_N = \frac{1}{TJ}(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)^T = \frac{1}{TJ}\mathbf{Q}. \tag{5}$$

Replacing $\mathbf{R}_N$ in (4) with $\hat{\mathbf{R}}_N$ and simplifying we find that the MLEs of $\mathbf{H}$ and $\boldsymbol{\theta}$ are obtained by minimizing $\det \mathbf{Q}$ over $\mathbf{H}$ and $\boldsymbol{\theta}$. First, we find the MLE of the signal amplitude $\boldsymbol{\theta}$ as a function of the unknown $\mathbf{H}$ by considering the following minimization problem

$$\min_{\boldsymbol{\theta}} |(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)^T|. \tag{6}$$

[1] In general, $\mathbf{H}$ is $N \times 3$ dimensional. However, the magnetic field generated by a radially oriented dipole located inside a spherically symmetric volume conductor is zero, which forces $\mathbf{H}$ to be rank deficient in MEG applications. We assume that the linearly dependent column has been removed.

Define the unitary transformation $\mathbf{T}\mathbf{T}^T = \mathbf{T}^T\mathbf{T} = \mathbf{I}$, where[2] $\mathbf{T} = [\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1/2}; \bar{\mathbf{D}}]$ and the columns of the $JT \times JT - L$ matrix $\bar{\mathbf{D}}$ are an orthonormal basis for the null-space of the columns of $\mathbf{D}$. That is $\mathbf{D}^T\bar{\mathbf{D}} = \mathbf{0}$ and $\bar{\mathbf{D}}^T\bar{\mathbf{D}} = \mathbf{I}$. Thus, the minimization problem (6) may be rewritten as

$$\min_{\boldsymbol{\theta}} |(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)\mathbf{T}\mathbf{T}^T(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)^T|. \tag{7}$$

Next, define the following quantities:

$$\begin{aligned} \mathbf{X}_{\mathbf{D}} &= \mathbf{X}\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1/2} &(8)\\ \mathbf{X}_{\bar{\mathbf{D}}} &= \mathbf{X}\bar{\mathbf{D}} &(9)\\ \mathbf{Q}_{\bar{\mathbf{D}}} &= \mathbf{X}_{\bar{\mathbf{D}}}\mathbf{X}_{\bar{\mathbf{D}}}^T &(10)\\ \tilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}(\mathbf{D}^T\mathbf{D})^{1/2}. &(11) \end{aligned}$$

Note that the columns of $\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1/2}$ are an orthonormal basis for the space spanned by the columns of $\mathbf{D}$. Hence, $\mathbf{X}_{\mathbf{D}}$ and $\mathbf{X}_{\bar{\mathbf{D}}}$ represents the coordinates of the data with respect to the space defined by the columns of $\mathbf{D}$ and $\bar{\mathbf{D}}$ respectively. Since the rows of the signal $\mathbf{S}$ lie in the space spanned by the rows of $\bar{\mathbf{D}}^T$, we observe that $\mathbf{X}_{\mathbf{D}}$ contains signal, while $\mathbf{X}_{\bar{\mathbf{D}}}$ is signal free and contains noise alone. Simplifying (7) using the definitions (8)–(11) yields

$$\min_{\tilde{\boldsymbol{\theta}}} \left|\mathbf{Q}_{\bar{\mathbf{D}}} + (\mathbf{X}_{\mathbf{D}} - \mathbf{H}\tilde{\boldsymbol{\theta}})(\mathbf{X}_{\mathbf{D}} - \mathbf{H}\tilde{\boldsymbol{\theta}})^T\right|. \tag{12}$$

Now use a square root factorization of $\mathbf{Q}_{\bar{\mathbf{D}}}$ to rewrite (12) in the form

$$\begin{aligned} &\min_{\tilde{\boldsymbol{\theta}}} |\mathbf{Q}_{\bar{\mathbf{D}}}| \left|\mathbf{I} + \mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}(\mathbf{X}_{\mathbf{D}} - \mathbf{H}\tilde{\boldsymbol{\theta}})(\mathbf{X}_{\mathbf{D}} - \mathbf{H}\tilde{\boldsymbol{\theta}})^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}\right|\\ &\quad = \min_{\tilde{\boldsymbol{\theta}}} \left|\mathbf{I} + \mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}(\mathbf{X}_{\mathbf{D}} - \mathbf{H}\tilde{\boldsymbol{\theta}})(\mathbf{X}_{\mathbf{D}} - \mathbf{H}\tilde{\boldsymbol{\theta}})^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}\right| \quad (13) \end{aligned}$$

where the equality in (13) results because $\det \mathbf{Q}_{\bar{\mathbf{D}}}$ is independent of $\boldsymbol{\theta}$ and $\mathbf{H}$ and thus can be dropped from the minimization problem. Note that $\mathbf{Q}_{\bar{\mathbf{D}}}$ is the sum of the outer product of $JT - L$ independent random vectors. Thus, $\mathbf{Q}_{\bar{\mathbf{D}}}$ is invertible and $\mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}$ exists with probability one provided $JT - L \geq N$ [13].

Consider a general minimization problem of the form

$$\min_{\mathbf{G}} |\mathbf{I} + (\mathbf{A} - \mathbf{B}\mathbf{G})(\mathbf{A} - \mathbf{B}\mathbf{G})^T|. \tag{14}$$

It can be shown that the minimum is attained when $\mathbf{G} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}$; that is, $\mathbf{G}$ is the least squares solution to $\mathbf{A} - \mathbf{B}\mathbf{G} = \mathbf{0}$ [4]. Identifying $\mathbf{A} = \mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}\mathbf{X}_{\mathbf{D}}$ and $\mathbf{B} = \mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}\mathbf{H}$ in (13) we obtain the estimate for $\tilde{\boldsymbol{\theta}}$

$$\hat{\tilde{\boldsymbol{\theta}}} = \left(\mathbf{H}^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1}\mathbf{X}_{\mathbf{D}}. \tag{15}$$

Substituting $\hat{\tilde{\boldsymbol{\theta}}}$ into the cost function of (13), we find the MLE for $\mathbf{H}$ by solving

$$\begin{aligned} \min_{\mathbf{H}} \Big|&\mathbf{I} + \mathbf{X}_{\mathbf{D}}^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1}\mathbf{X}_{\mathbf{D}}\\ &- \mathbf{X}_{\mathbf{D}}^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1}\mathbf{H}\left(\mathbf{H}^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1}\mathbf{X}_{\mathbf{D}}\Big|. \quad (16) \end{aligned}$$

Now define

$$\tilde{\mathbf{H}} = \mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}\mathbf{H}\left(\mathbf{H}^T\mathbf{Q}_{\bar{\mathbf{D}}}^{-1}\mathbf{H}\right)^{-1/2} \text{ and } \tilde{\mathbf{X}}_{\mathbf{D}} = \mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}\mathbf{X}_{\mathbf{D}},$$

so that (16) can be rewritten as

$$\min_{\tilde{\mathbf{H}}} \left|\mathbf{I} + \tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{X}}_{\mathbf{D}} - \tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T\tilde{\mathbf{X}}_{\mathbf{D}}\right|. \tag{17}$$

Also note that $\tilde{\mathbf{H}}^T\tilde{\mathbf{H}} = \mathbf{I}$, that is, the columns of $\tilde{\mathbf{H}}$ are an orthonormal basis for the space spanned by the columns of $\mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}\mathbf{H}$. Next, let $\boldsymbol{\Gamma} = \mathbf{I} + \tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{X}}_{\mathbf{D}}$ and factor the determinant to write

$$\begin{aligned} &\left|\boldsymbol{\Gamma}^{1/2}\left(\mathbf{I} - \boldsymbol{\Gamma}^{-1/2}\tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1/2}\right)\boldsymbol{\Gamma}^{1/2}\right|\\ &\qquad = \left|\mathbf{I} - \boldsymbol{\Gamma}^{-1/2}\tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1/2}\right||\boldsymbol{\Gamma}|. \quad (18) \end{aligned}$$

Lastly, using the identity $|\mathbf{I} - \mathbf{A}\mathbf{A}^T| = |\mathbf{I} - \mathbf{A}^T\mathbf{A}|$ we may rewrite the minimization problem (17) in the equivalent form

$$\min_{\tilde{\mathbf{H}}} \left|\mathbf{I} - \tilde{\mathbf{H}}^T\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{H}}\right|. \tag{19}$$

In the case $P = 1$—for example, if the signal is dipolar with constant, unknown moment orientation—then $\tilde{\mathbf{H}}$ is an $N \times 1$ vector and (19) involves minimizing the determinant of the scalar $1 - \tilde{\mathbf{H}}^T\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{H}}$. The solution is to choose $\tilde{\mathbf{H}}$ as the eigenvector corresponding to the largest eigenvalue of $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$. In the case $P \geq 2$, the solution for $\tilde{\mathbf{H}}$ is a bit more complicated. Note that since $\left|\mathbf{I} - \tilde{\mathbf{H}}^T\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{H}}\right| = \left|\mathbf{I} - \mathbf{U}\tilde{\mathbf{H}}^T\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{H}}\mathbf{U}^T\right|$ where $\mathbf{U}$ is any $P \times P$ unitary matrix, then, without loss of generality, we may assume $\tilde{\mathbf{H}}$ diagonalizes $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$. That is, $\tilde{\mathbf{H}}$ contains P of the N eigenvectors of the matrix $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$. From this observation it follows that the MLE of $\tilde{\mathbf{H}}$ is given by the P eigenvectors of $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$ corresponding to the P largest eigenvalues. This corresponds to the solution for the reduced rank regression problem of [7].

Finally, we note that the eigenvalues of $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$ are upper bounded by one and thus the cost function in (19) is nonnegative. To see this, let the singular value decomposition of $\tilde{\mathbf{X}}_{\mathbf{D}}$ be $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, and let σ_i denote the ith diagonal element of $\boldsymbol{\Sigma}$. Using $\boldsymbol{\Gamma} = \mathbf{I} + \tilde{\mathbf{X}}_{\mathbf{D}}^T\tilde{\mathbf{X}}_{\mathbf{D}}$, substituting $\tilde{\mathbf{X}}_{\mathbf{D}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, and simplifying gives $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T = \mathbf{U}\boldsymbol{\Sigma}(\mathbf{I} + \boldsymbol{\Sigma}^2)^{-1}\boldsymbol{\Sigma}\mathbf{U}^T$. Hence, the eigenvalues of $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$ are $\sigma_i^2/\left(1 + \sigma_i^2\right) < 1, i = 1, 2, \ldots, N$. Furthermore, the corresponding eigenvectors of $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$ are the left singular vectors of $\tilde{\mathbf{X}}_{\mathbf{D}}$ contained in the columns of $\mathbf{U}$.

Given the MLEs $\hat{\tilde{\mathbf{H}}}$ and $\hat{\tilde{\boldsymbol{\theta}}}$, we obtain the MLE for the signal component of the data, $\hat{\mathbf{S}}$, as

$$\hat{\mathbf{S}} = \mathbf{Q}_{\bar{\mathbf{D}}}^{1/2}\hat{\tilde{\mathbf{H}}}\hat{\tilde{\mathbf{H}}}^T\tilde{\mathbf{X}}_{\mathbf{D}}(\mathbf{D}^T\mathbf{D})^{-1/2}\mathbf{C}^T. \tag{20}$$

B. Interpretation

In order to develop insight into the nature of the solution (20), consider a simpler maximum-likelihood estimation problem in which $\mathbf{R}_N$ in (4) is known. In this case, the goal is to solve

$$\min_{\mathbf{H},\boldsymbol{\theta}} \operatorname{tr}\left[\mathbf{R}_N^{-1}(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)(\mathbf{X} - \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T)^T\right]. \tag{21}$$

[2] Alternatively, $\mathbf{T}$ may also be defined by appropriately partitioning the complete set of the left singular vectors of $\mathbf{D}$.

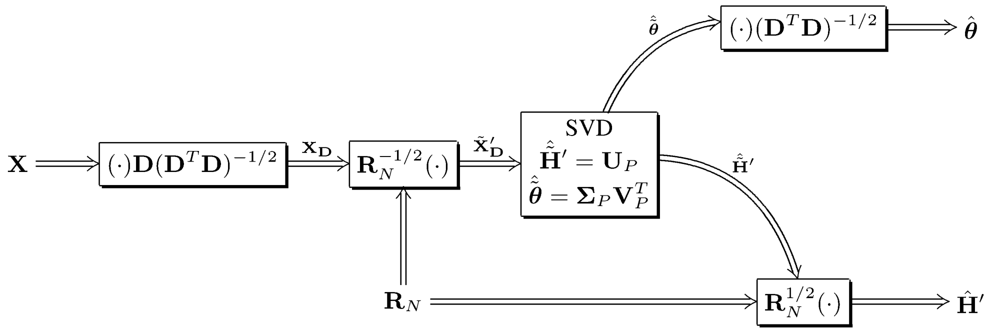

$\hat{\tilde{\boldsymbol{\theta}}}$
$(\cdot)(\mathbf{D}^T\mathbf{D})^{-1/2}$
$\hat{\boldsymbol{\theta}}$
$\mathbf{X}$
$(\cdot)\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1/2}$
$\mathbf{X_D}$
$\mathbf{R}_N^{-1/2}(\cdot)$
$\tilde{\mathbf{X}}'_{\mathbf{D}}$
SVD
$\hat{\tilde{\mathbf{H}}}' = \mathbf{U}_P$
$\hat{\tilde{\boldsymbol{\theta}}} = \boldsymbol{\Sigma}_P\mathbf{V}_P^T$
$\hat{\tilde{\mathbf{H}}}'$
$\mathbf{R}_N$
$\mathbf{R}_N^{1/2}(\cdot)$
$\hat{\mathbf{H}}'$

Fig. 1. Algorithm steps in the case of known $\mathbf{R}_N$.

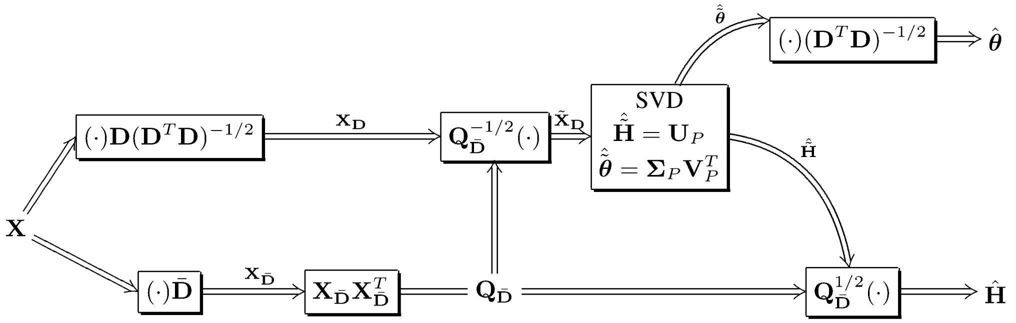

$\hat{\tilde{\boldsymbol{\theta}}}$
$(\cdot)(\mathbf{D}^T\mathbf{D})^{-1/2}$
$\hat{\boldsymbol{\theta}}$
$(\cdot)\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1/2}$
$\mathbf{X_D}$
$\mathbf{Q}_{\bar{\mathbf{D}}}^{-1/2}(\cdot)$
$\tilde{\mathbf{X}}_{\mathbf{D}}$
SVD
$\hat{\tilde{\mathbf{H}}} = \mathbf{U}_P$
$\hat{\tilde{\boldsymbol{\theta}}} = \boldsymbol{\Sigma}_P\mathbf{V}_P^T$
$\hat{\tilde{\mathbf{H}}}$
$\mathbf{X}$
$(\cdot)\bar{\mathbf{D}}$
$\mathbf{X}_{\bar{\mathbf{D}}}$
$\mathbf{X}_{\bar{\mathbf{D}}}\mathbf{X}_{\bar{\mathbf{D}}}^T$
$\mathbf{Q}_{\bar{\mathbf{D}}}$
$\mathbf{Q}_{\bar{\mathbf{D}}}^{1/2}(\cdot)$
$\hat{\mathbf{H}}$

Fig. 2. Algorithm steps in the case of unknown $\mathbf{R}_N$.

Factoring $\mathbf{R}_N^{-1} = \mathbf{R}_N^{-1/2}\mathbf{R}_N^{-1/2}$, using the identity $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$, and inserting the transformation $\mathbf{T}$ as in (7), we rewrite (21) as

$$\min_{\mathbf{H},\boldsymbol{\theta}} \operatorname{tr}\left[\left(\mathbf{R}_N^{-1/2}\mathbf{X}\mathbf{T} - \mathbf{R}_N^{-1/2}\mathbf{H}\boldsymbol{\theta}\mathbf{D}^T\mathbf{T}\right) \times \left(\mathbf{R}_N^{-1/2}\mathbf{X}\mathbf{T} - \mathbf{R}_N^{-1/2}\mathbf{H}\boldsymbol{\theta}\mathbf{D}^T\mathbf{T}\right)^T\right]. \quad (22)$$

Expand $\mathbf{T}$ and define $\tilde{\mathbf{X}}'_{\mathbf{D}} = \mathbf{R}_N^{-1/2}\mathbf{X}\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1/2}$, $\tilde{\mathbf{H}}' = \mathbf{R}_N^{-1/2}\mathbf{H}$, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{D}^T\mathbf{D})^{1/2}$ to write the minimization problem in (22) as

$$\min_{\tilde{\mathbf{H}}',\boldsymbol{\theta}} \operatorname{tr}\left[\left(\tilde{\mathbf{X}}'_{\mathbf{D}} - \tilde{\mathbf{H}}'\tilde{\boldsymbol{\theta}}\right)\left(\tilde{\mathbf{X}}'_{\mathbf{D}} - \tilde{\mathbf{H}}'\tilde{\boldsymbol{\theta}}\right)^T\right] + \operatorname{tr}\left[\mathbf{R}_N^{-1/2}\mathbf{X}_{\bar{\mathbf{D}}}\mathbf{X}_{\bar{\mathbf{D}}}^T\mathbf{R}_N^{-1/2}\right]. \quad (23)$$

The first term in (23) is just the Frobenius norm squared of the difference $\tilde{\mathbf{X}}'_{\mathbf{D}} - \tilde{\mathbf{H}}'\tilde{\boldsymbol{\theta}}$, while the second term is independent of $\tilde{\mathbf{H}}'$ and $\tilde{\boldsymbol{\theta}}$. Hence, the maximum-likelihood problem is equivalent to finding the best rank P approximation (in the Frobenius norm sense) to $\tilde{\mathbf{X}}'_{\mathbf{D}}$. It is well known that the solution is obtained by constructing $\tilde{\mathbf{H}}'\tilde{\boldsymbol{\theta}}$ from the P largest singular values and corresponding singular vectors of $\tilde{\mathbf{X}}'_{\mathbf{D}}$ [14]. That is if $\tilde{\mathbf{X}}'_{\mathbf{D}} = \mathbf{U}'\boldsymbol{\Sigma}'\mathbf{V}'^T$ is the singular value decomposition of $\tilde{\mathbf{X}}'_{\mathbf{D}}$ then we may set $\tilde{\mathbf{H}}' = \mathbf{U}'_P$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\Sigma}'_P\mathbf{V}'^T_P$ where $\mathbf{U}'_P$ and $\mathbf{V}'_P$ are the first P columns of $\mathbf{U}'$ and $\mathbf{V}'$, respectively, and $\boldsymbol{\Sigma}'_P$ is the upper left $P \times P$ block of $\boldsymbol{\Sigma}'$. Equivalently, $\tilde{\mathbf{H}}'$ and $\tilde{\boldsymbol{\theta}}$ are obtained from a P-term principal component expansion of $\tilde{\mathbf{X}}'_{\mathbf{D}}$.

The set of operations involved in estimating $\mathbf{H}$ and $\boldsymbol{\theta}$ for the case when $\mathbf{R}_N$ is known is illustrated in Fig. 1. First, we identify the component of the data that lies in the space spanned by $\mathbf{D}$. Next, this component is whitened using the known noise covariance matrix. Note that if $\mathbf{X} = \mathbf{H}\boldsymbol{\theta}\mathbf{D}^T + \mathbf{N}$ then after whitening $\tilde{\mathbf{X}}'_{\mathbf{D}} = \tilde{\mathbf{H}}'\tilde{\boldsymbol{\theta}} + \tilde{\mathbf{N}}_{\mathbf{D}}$, where $\tilde{\mathbf{N}}_{\mathbf{D}} = \mathbf{R}_N^{-1/2}\mathbf{N}\mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1/2}$ is spatially white and has statistically independent columns. Since $\tilde{\mathbf{X}}'_{\mathbf{D}}$ contains a low-rank signal term plus independent, identically distributed noise, the principal component expansion of the next step provides the optimum rank P approximation. Finally, in the last step we obtain estimates $\hat{\mathbf{H}}$ and $\hat{\boldsymbol{\theta}}$ by spatially coloring the column space of $\tilde{\mathbf{X}}'_{\mathbf{D}}$ and correcting for the factor $(\mathbf{D}^T\mathbf{D})^{-1/2}$.

In the estimation problem considered in Section III-A, $\mathbf{Q}_{\bar{\mathbf{D}}}$ takes the role of $\mathbf{R}_N$ in whitening the data matrix $\mathbf{X_D}$. Transformation of $\mathbf{X}$ on the right by $\bar{\mathbf{D}}$ does not change the spatial covariance associated with the columns of $\mathbf{X}$ because $\bar{\mathbf{D}}$ has orthonormal columns. Thus, $\mathbf{Q}_{\bar{\mathbf{D}}}$ is a sum of outer products of independent, signal-free data vectors and represents an estimate of $\mathbf{R}_N$. At first glance it appears that the estimate for $\hat{\tilde{\mathbf{H}}}'$ (constructed from the SVD of $\tilde{\mathbf{X}}'_{\mathbf{D}}$) differs from the estimate of $\hat{\tilde{\mathbf{H}}}$ (constructed from the eigendecomposition of $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$). However, since the eigenvectors of $\tilde{\mathbf{X}}_{\mathbf{D}}\boldsymbol{\Gamma}^{-1}\tilde{\mathbf{X}}_{\mathbf{D}}^T$ correspond to the left singular vectors of $\tilde{\mathbf{X}}_{\mathbf{D}}$, we see that $\hat{\tilde{\mathbf{H}}}'$ and $\hat{\tilde{\mathbf{H}}}$ may be estimated in an analogous manner. Thus, we may interpret the MLEs derived in Section III-A in terms of the principal component representation for an appropriately whitened data matrix. The temporal structure of the signal represented by the columns of $\mathbf{D}$ enables extraction of the signal-free data that is used to obtain a noise covariance matrix estimate for the whitening step.

Fig. 2 depicts MLEs for the case of unknown $\mathbf{R}_N$, analogous to Fig. 1. As in Fig. 1, we first identify the component of the data that lies in the space spanned by $\mathbf{D}$. However, in addition we determine the component that lies in the null space of the signal, spanned by the columns of $\bar{\mathbf{D}}$. From the signal-free data $\mathbf{X}_{\bar{\mathbf{D}}}$, we estimate a spatial covariance matrix $\mathbf{Q}_{\bar{\mathbf{D}}}$ which is used

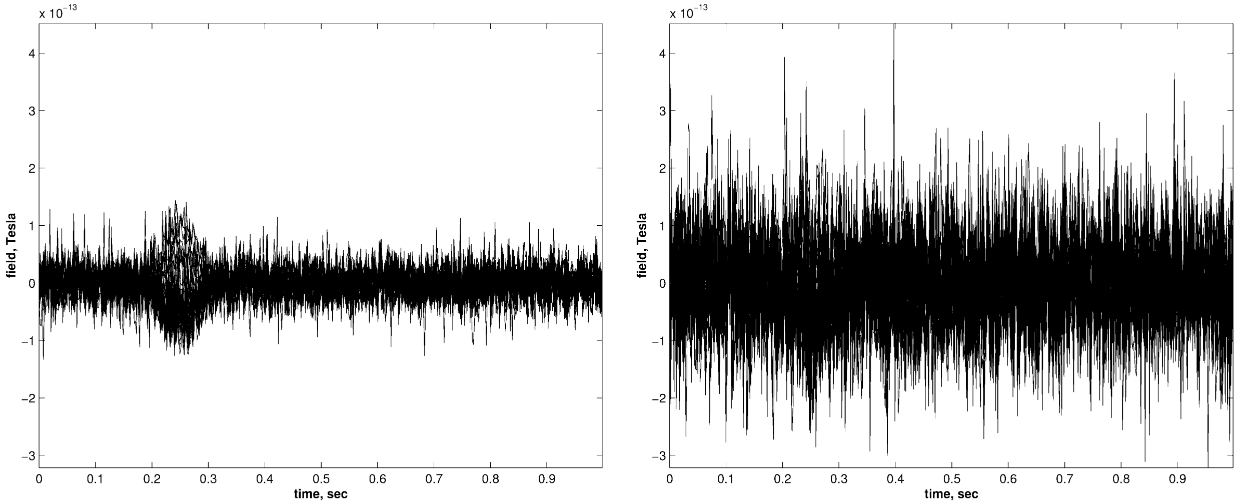


Fig. 3. Synthetic averaged signal estimate. Left: high SNR (13.2 dB), right: low SNR (3.6 dB). In the simulation, the signal amplitude is constant and the noise amplitude varies.

to whiten the signal component $\mathbf{X}_{\mathbf{D}}$. The remaining steps in Fig. 2 are identical to those in Fig. 1 with $\mathbf{Q}_{\bar{\mathbf{D}}}$ replacing $\mathbf{R}_N$.

C. Effect of Temporal Noise Correlation Errors

The actual temporal noise correlation structure may differ from that assumed or estimated from the data. In this section, we provide a qualitative analysis of the effects of temporal noise correlation matrix errors on the algorithm presented in Section III-A. It suffices to consider the case where the assumed noise correlation matrix is $\mathbf{I}$ and the true underlying noise correlation matrix is $\mathbf{R}_T \neq \mathbf{I}$. We assume that T is large with respect to the temporal correlation extent of the noise and the noise is temporally stationary, in which case the DFT matrix $\mathbf{G}$ diagonalizes $\mathbf{R}_T$. That is, $\mathbf{R}_T \approx \mathbf{G}\mathbf{P}^2\mathbf{G}^H$, where $\mathbf{P}^2$ is a diagonal matrix, containing the power spectral density of the noise [15]. We further assume that the error between the assumed or estimated and true noise correlation matrices is relatively small, which implies that each diagonal element of $\mathbf{P}^2$ is on the order of one.

Consider replacing $\mathbf{X}_j$ with frequency domain data $\mathbf{X}_j\mathbf{G}$. The data transformed in this way has unknown spatial covariance $\mathbf{R}_N$ and diagonal "temporal" covariance $\mathbf{P}^2$. Thus, the ith column of $\mathbf{X}_j$ has covariance matrix $\mathbf{P}^2_{ii}\mathbf{R}_N$ and each column of $\mathbf{X}_j$ has identical spatial covariance except for the constant $\mathbf{P}^2_{ii}$. This implies that operations based on the columns of $\mathbf{X}_i$ will be minimally affected by unknown temporal noise coloration. For example, $\mathbf{Q}_{\bar{\mathbf{D}}}$, defined in (10), involves a sum of outer products of a subset of the columns (those associated with frequencies outside the band of interest) of $\mathbf{X}_j$. The presence of unknown noise coloration changes the relative amplitude of each column and consequently changes the overall amplitude of $\mathbf{Q}_{\bar{\mathbf{D}}}$. However, this unknown factor does not effect the ability of $\mathbf{Q}_{\bar{\mathbf{D}}}$ to spatially whiten the columns of $\mathbf{X}_{\mathbf{D}}$.

Next, consider obtaining $\hat{\tilde{\mathbf{H}}}$ and $\hat{\tilde{\boldsymbol{\theta}}}$ from the SVD of $\tilde{\mathbf{X}}_{\mathbf{D}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$. We obtain $\hat{\tilde{\mathbf{H}}}$ as the columns of $\mathbf{U}$ associated with significant singular values. Equivalently, $\hat{\tilde{\mathbf{H}}}$ may be obtained from the eigenvectors of $\tilde{\mathbf{X}}_{\mathbf{D}}\tilde{\mathbf{X}}_{\mathbf{D}}^H$, which is a sum of outer products of columns of $\tilde{\mathbf{X}}_{\mathbf{D}}$. Note that $\tilde{\mathbf{X}}_{\mathbf{D}} = \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}} + \tilde{\mathbf{N}}$, where the columns of the spatially whitened noise $\tilde{\mathbf{N}}$ have covariance proportional to $\mathbf{I}$. On average, $E\left\{\tilde{\mathbf{X}}_{\mathbf{D}}\tilde{\mathbf{X}}_{\mathbf{D}}^H\right\} = \tilde{\mathbf{H}}\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^H\tilde{\mathbf{H}}^T + \gamma\mathbf{I}$. Thus, unknown noise coloration reflected by the constant γ has minimal impact on estimation of $\tilde{\mathbf{H}}$ because $\tilde{\mathbf{H}}$ is a basis for the spatial component of the mean and the spatial component of the noise is white.

On the other hand, $\hat{\tilde{\boldsymbol{\theta}}}$ is obtained from the eigenvectors of the product $\tilde{\mathbf{X}}_{\mathbf{D}}^H\tilde{\mathbf{X}}_{\mathbf{D}} = \tilde{\boldsymbol{\theta}}^H\tilde{\boldsymbol{\theta}} + \tilde{\mathbf{N}}^H\tilde{\mathbf{N}}$, since $\tilde{\mathbf{H}}^H\tilde{\mathbf{H}} = \mathbf{I}$. Assuming $\mathbf{C}$ represents a frequency band of interest, then the columns of $\tilde{\mathbf{N}}$ are frequency components of the spatially whitened noise, averaged over epochs in the band of interest and $E\{\tilde{\mathbf{N}}^H\tilde{\mathbf{N}}\} = \tilde{\mathbf{P}}^2$, a diagonal matrix, containing the noise power spectral density on the band. In this case, a noise component may dominate the actual signal component in $\tilde{\boldsymbol{\theta}}$ and the estimate of $\tilde{\boldsymbol{\theta}}$ becomes contaminated by noise. This swapping of noise and signal components is likely when the SNR is low after averaging over all epochs. Note that noise coloration outside the band of interest does not contribute to errors in $\tilde{\boldsymbol{\theta}}$. Thus, if the noise has significant coloration and is of comparable power relative to the mean, the estimate of $\tilde{\boldsymbol{\theta}}$ will be contaminated. As $\tilde{\boldsymbol{\theta}}$ captures the temporal evolution of the signal, it seems logical that the temporally colored noise would warp its estimate most significantly.

IV. Results

In this section, we demonstrate the effectiveness of the maximum-likelihood approach using both simulated and real MEG data.

A. Synthetic Data

In order to facilitate comparison of the simulated and real data, we generate simulated data using the sensor configuration of the 37-channel Magnes II system (Biomagnetic Technologies, Inc.) in our lab. Each simulation run consists of 100 one-second-long epochs, with 37 channels and 521 time samples per epoch. This data corresponds to a typical set of 100 epochs of evoked response MEG data. The data is generated assuming that the signal originates from a fixed dipole located 2 cm below the surface and 2 cm off the radial axis of a spherical

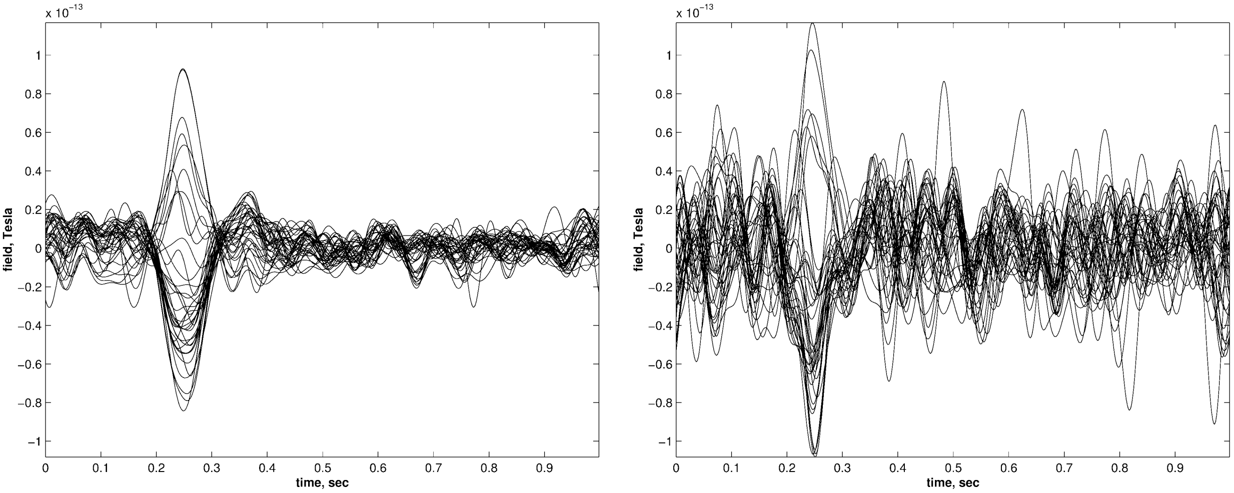


Fig. 4. Filtered 1–20 Hz (fourth-order zero-phase butterworth filter) and averaged synthetic data. Left: high SNR, right: low SNR.

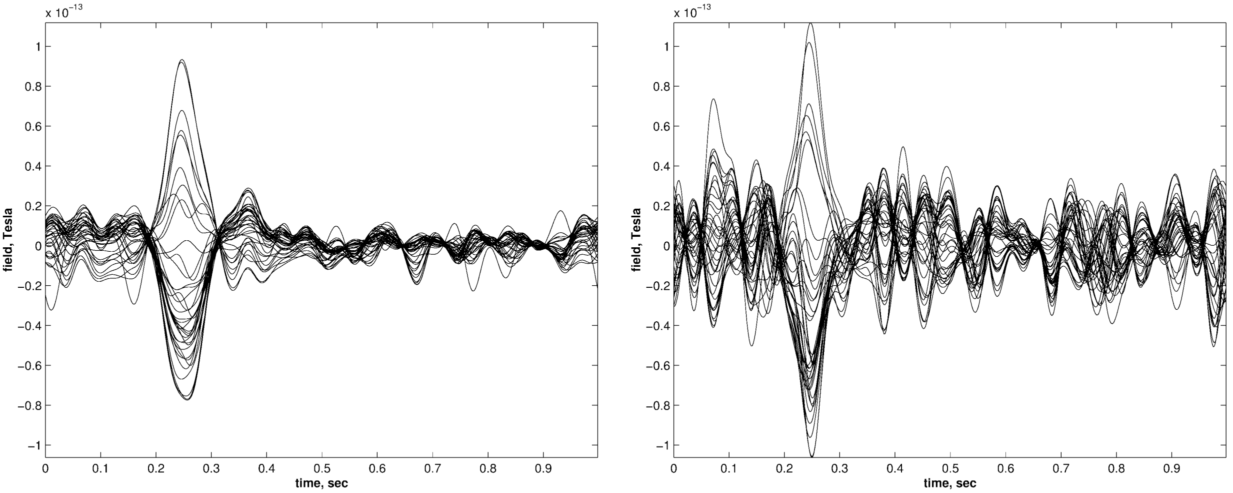


Fig. 5. Rank 3 principal component representation for the filtered and averaged synthetic data in Fig. 4. Left: high SNR, right: low SNR.

head model with 12-cm radius.[3] The time evolution of the signal for all epochs is chosen as a Gaussian pulse of 70 ms full width at half maximum, with the maximum at 250 ms relative to the start of the epoch. Spatially colored noise of different levels was added to this signal. The noise is generated by spatially coloring white noise to match a noise covariance matrix estimated from real subject data. The temporal noise covariance matrix is unity.

Fig. 3 shows the 37-channel signal estimate obtained by averaging 100 epochs of high (13.2 dB) and low (3.6 dB) SNR simulated data, where SNR is defined as $\mathrm{tr}\left(\mathbf{S}^T\mathbf{R}_N^{-1}\mathbf{S}\right)$. This definition accounts for the total energy across time and space in the signal and the spatial coloration of the noise. If the noise is spatially white, i.e., $\mathbf{R}_N = \sigma^2\mathbf{I}$, then SNR is $\mathrm{tr}(\mathbf{S}^T\mathbf{S}/\sigma^2)$, the ratio of the total energy in the signal to the noise variance. Note that the presence of the signal term is completely obscured by noise in the low SNR example even after averaging 100 epochs. Fig. 4 shows the 37-channel data after filtering the averaged data with a fourth-order zero-phase Butterworth filter having passband 1–20 Hz. The presence of a signal is barely evident above the noise in the low SNR example. Data of this quality would typically be discarded in traditional MEG analysis and not used for further processing, such as source localization.

Fig. 5 demonstrates a rank three principal component representation of the averaged and filtered dataset shown in Fig. 4. The principal component representation offers some improvement; however, it is problematic with spatially colored noise and low SNR, since one or more principal components may be noise components. Whitening of the noise is necessary to prevent this potential confusion between signal and noise subspaces. Fig. 6 illustrates the result of first whitening the data using a noise covariance matrix constructed from 50 samples of prestimulus data and then applying a rank three principle component representation. Whitening leads to additional improvement in signal quality.

Fig. 7 depicts the MLE of the 37-channel signal obtained from (20). In this case, we chose $\mathbf{C}$ to be a basis for signals bandlimited on 1–20 Hz, with $L = 40$. The results in Fig. 7 assume $P = 3$. Rank 3 approximations are used as they provide empirically satisfying results for real evoked response data. For comparison purposes we chose P to be the same for the simulated data even though the simulation model was rank 1. The MLE offers significant reductions in noise level relative to filtering and

[3] The rank of the simulation model is 1 in this case.

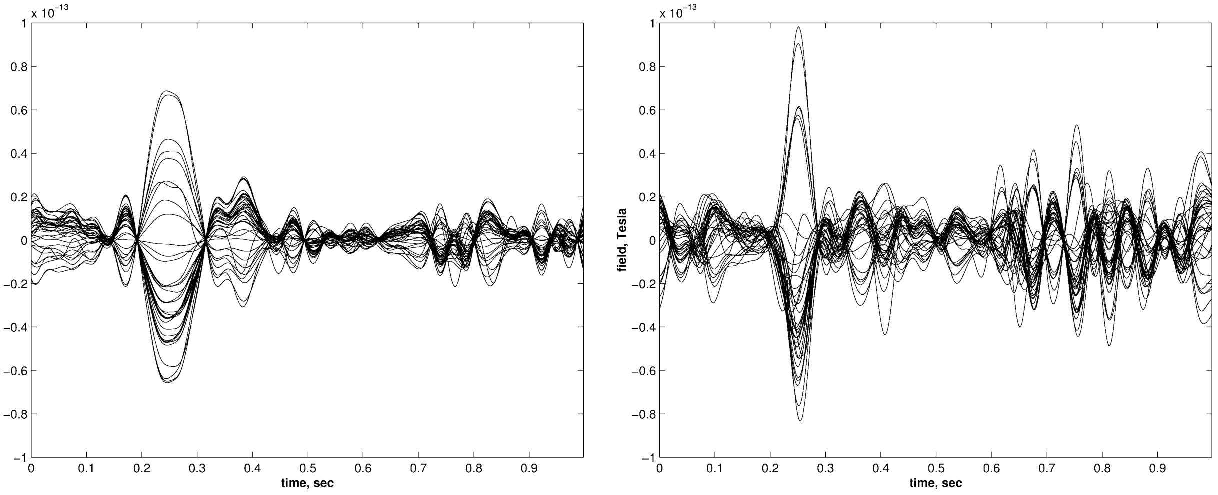


Fig. 6. Rank 3 principal component representation for the synthetic data in Fig. 4 after whitening using 50 prestimulus data vectors. Left: high SNR, right: low SNR.

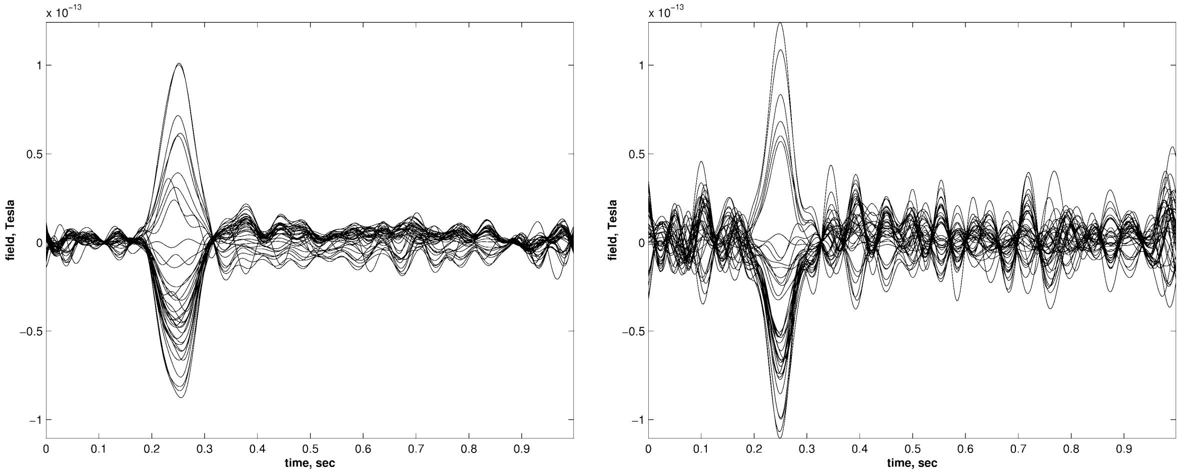


Fig. 7. MLEs of the synthetic signal using a rank three $\tilde{\mathbf{H}}$. Left: high SNR, right: low SNR.

averaging, particularly in the low SNR example, and modest visual improvement relative to prestimulus-based whitening. The quality of the low SNR estimate is now sufficient for postprocessing in a typical MEG application.

Fig. 8 compares the mean squared error (MSE) of the rank three maximum-likelihood signal estimate to traditional filtering and averaging, and prestimulus-based whitening followed by PCA as a function of SNR. The MSE is computed as the average over 100 independent simulations of the Frobenius norm squared of the difference between the estimated and true spatio-temporal signal. Each independent simulation contained 100 epochs. The Frobenius norm of the true spatio-temporal signal is one. Three cases of prestimulus-based whitening and PCA are shown, assuming either 50, 100 or 200 prestimulus data vectors are available to estimate the spatial noise covariance matrix used to whiten the remainder of the data record. Fig. 8 also includes the theoretical lower bound on the MSE obtained by assuming the noise covariance matrix is known, that is, using the algorithm considered in Section III-B and depicted in Fig. 1. Clearly, the proposed maximum-likelihood algorithm offers superior quality signal estimates at all SNR levels compared to these other methods and is nearly coincident with the lower bound attained with the known noise covariance matrix. This is because it exploits the structure in $\mathbf{D}$ to obtain the maximum number of independent, signal-free data vectors for noise covariance matrix estimation, which, in this scenario, leads to a very accurate spatial noise covariance matrix estimate. Note that the performance of the prestimulus-based whitening method improves as the number of prestimulus data vectors used to estimate the noise covariance matrix increases. The difference between the MSE and the bound of the known covariance matrix case is the performance loss associated with estimating the noise covariance matrix. When little data is available for prewhitening, the prestimulus-based whitening method does not offer significant improvement over traditional filtering and averaging.

B. Application to Real Data

The MEG data for this part of the analysis is recorded with a 37-channel SQUID magnetometer (Magnes II, Biomagnetic Technologies) in a magnetically shielded room. Several types of MEG data are analyzed: adult auditory evoked responses,

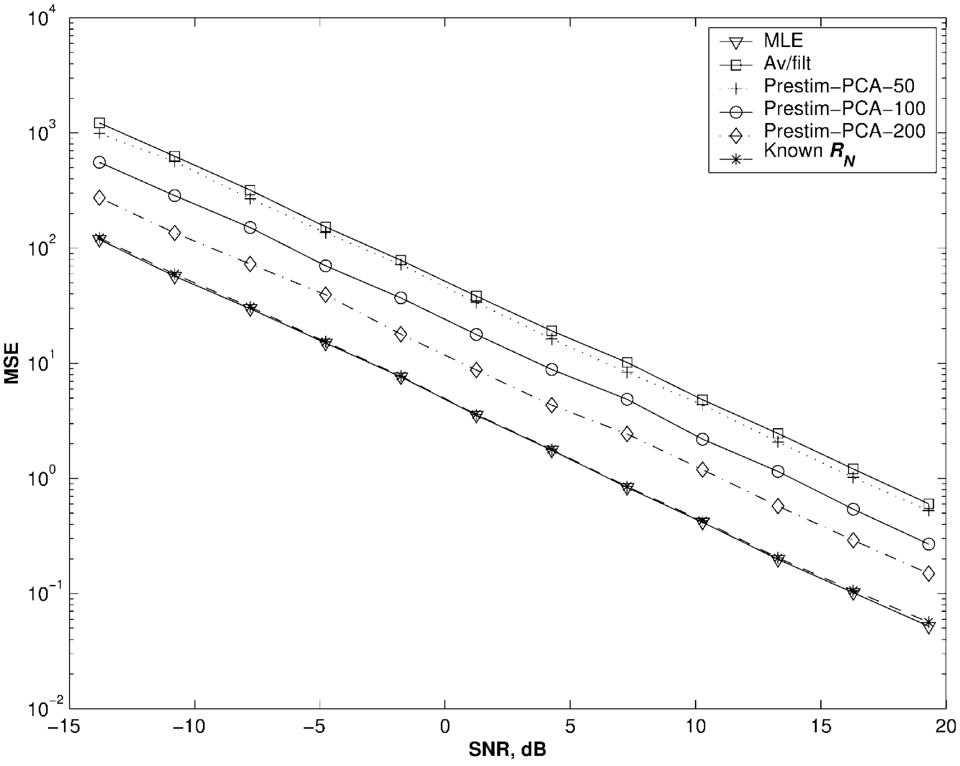


Fig. 8. Comparison of the MSE for different processing methods (MLE), filtering/averaging (Av/filt), prestimulus-based whitening and PCA with 50 (prestim-PCA-50), 100 (prestim-PCA-100) and 200 (prestim-PCA-200) points available for the noise covariance matrix estimate) versus SNR. Curve for the case of known $\mathbf{R}_N$ is also shown as a theoretical lower bound. A rank three approximation for the estimated signal is used in each case.

visual motion-defined contrast evoked potentials and fetal auditory evoked responses. The temporal noise covariance matrix is assumed to be the identity in all of the real data experiments. In the adult evoked data example, 100 one-second-long epochs are collected per run with sampling rate of 520.8 Hz. The auditory evoked response data is obtained by presenting 50 ms 1-kHz tones to the subject at 60-dB SPL above hearing threshold with 1.5-s average interstimulus interval. The results for the adult auditory evoked data are depicted in Fig. 9. Note that the MLEs offer enough SNR improvement so that the slow auditory evoked response is clearly identifiable in the record, in contrast to averaging and filtering.

The motion defined contrast experiment is conducted as follows. An object consisting of brush strokes (a bird) or dots (a rectangle) is presented over a background composed of identical, randomly distributed brush strokes or dots, respectively. The object blends perfectly with the background when stationary and is visible only when the object is coherently moving [16]. A control recording in which every element of the image is moving independently and randomly is made for comparison to the experiment involving coherent motion.

Figs. 10 and 11 depict the signal estimates for coherent motion and the control, respectively, using both averaging/filtering and the MLE. While both coherent motion and control stimuli elicit a strong response with latency 150–180 ms relative to onset of the object movement, only the coherent motion case exhibits a late response in the range 250–450 ms. This difference is difficult to identify using traditional filtering and averaging (left panels of Figs. 10 and 11) due to the low SNR of these signals. However, the MLEs (right panels of Figs. 10 and 11) clearly show the late time activity and may enable further analysis of the activity on an epoch-by-epoch basis.

Finally, we apply the proposed algorithm to the extremely challenging case of fetal auditory evoked response data. The stimuli are 1500 Hz 50 ms tone bursts of intensity 100 dB with 3.5-s average interstimulus interval. Approximately 300 epochs are collected per run. We assume the response of interest lies in the range 100 to 300 ms post stimulus. Fig. 12 clearly illustrates the advantages of the maximum-likelihood algorithm. First, in this case it is practically impossible to determine the signal-free portion of the data prior to processing. Second, the noise statistics often change over the acquisition interval due to the nonstationarity of fetal and maternal heart interference. Consequently, prestimulus-based whitening techniques are less effective. The top panel of Fig. 12 illustrates the signal estimate obtained by using the first 100 ms of data to estimate a spatial noise covariance matrix which is used to whiten the remainder of the record. A rank three principal component representation of the whitened data is used as the signal estimate. In contrast, the maximum-likelihood approach (bottom panel of Fig. 12) estimates the noise covariance matrix using the actual data of interest (100–300 ms) and provides significant SNR improvement. Although a sphere-based dipolar forward model is not applicable to fetal data, we expect the auditory evoked response to exhibit phase reversal across the sensor array, as illustrated in the bottom panel of Fig. 12, due to the localized nature of the source beneath the array. The maximum-likelihood method effectively separates the maternal and fetal heart interference from the fetal auditory evoked response, while residual maternal heart activity predominates with the prestimulus-based whitening method.

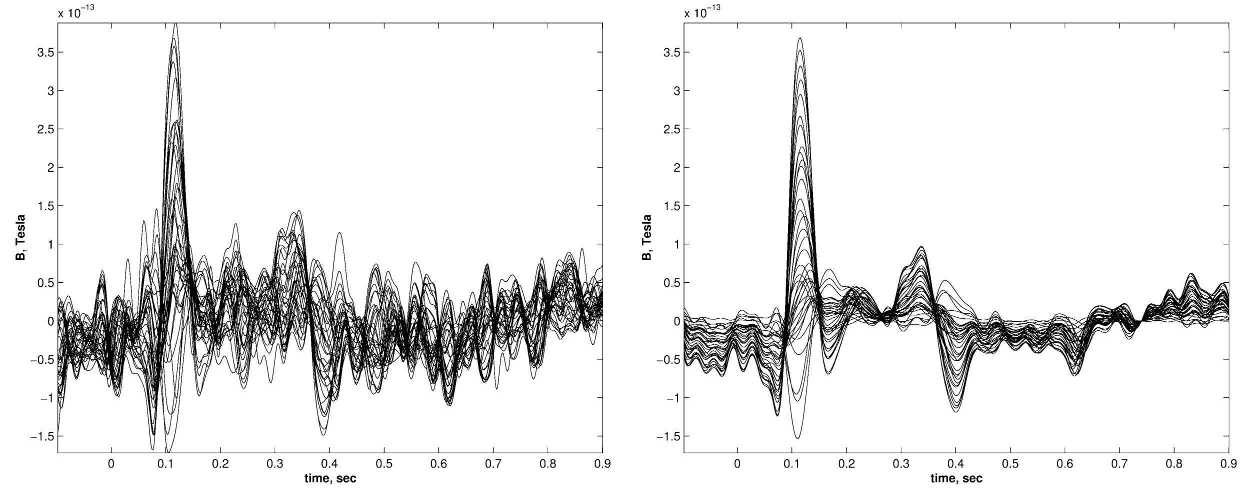


Fig. 9. Improvement of SNR in the auditory evoked response experiment. Left: conventional filtered and averaged epoch; right: MLE.

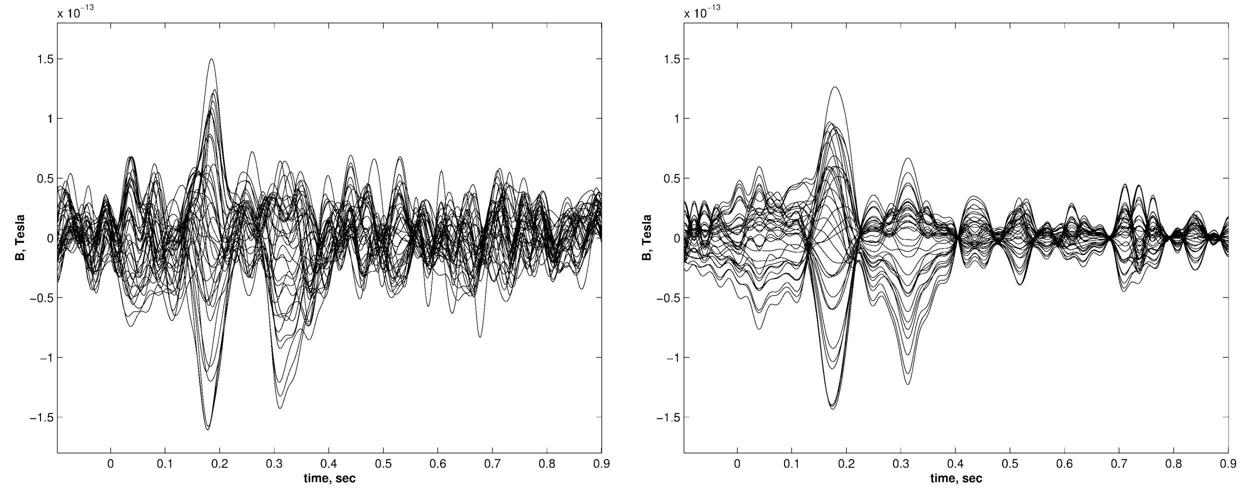


Fig. 10. Improvement of SNR in the coherent visual motion experiment. Left: conventional filtered and averaged epoch; right: MLE.

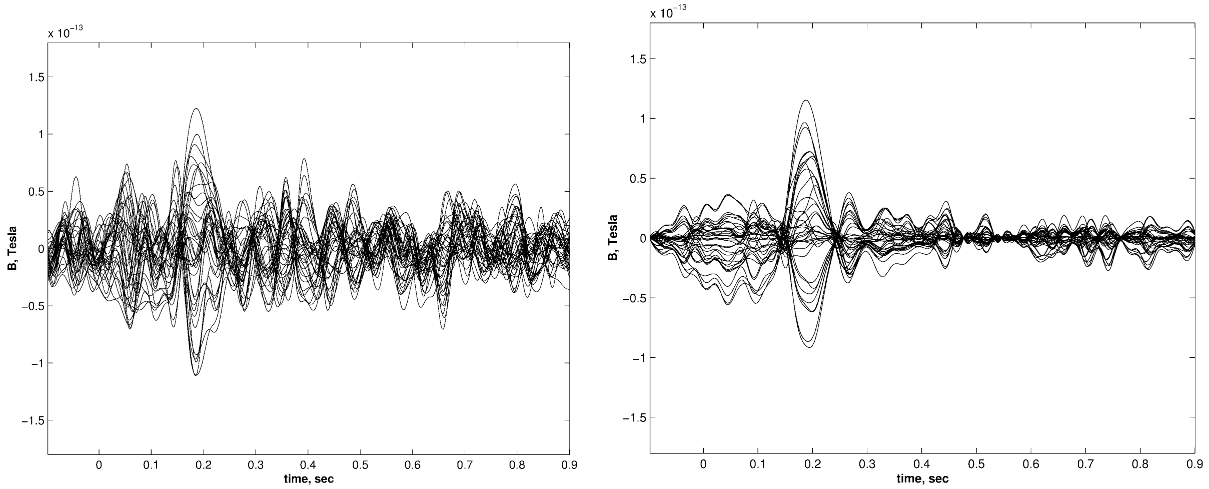


Fig. 11. Improvement of SNR in the coherent visual motion experiment for the control stimulus. Left: conventional filtered and averaged epoch; right: MLE.

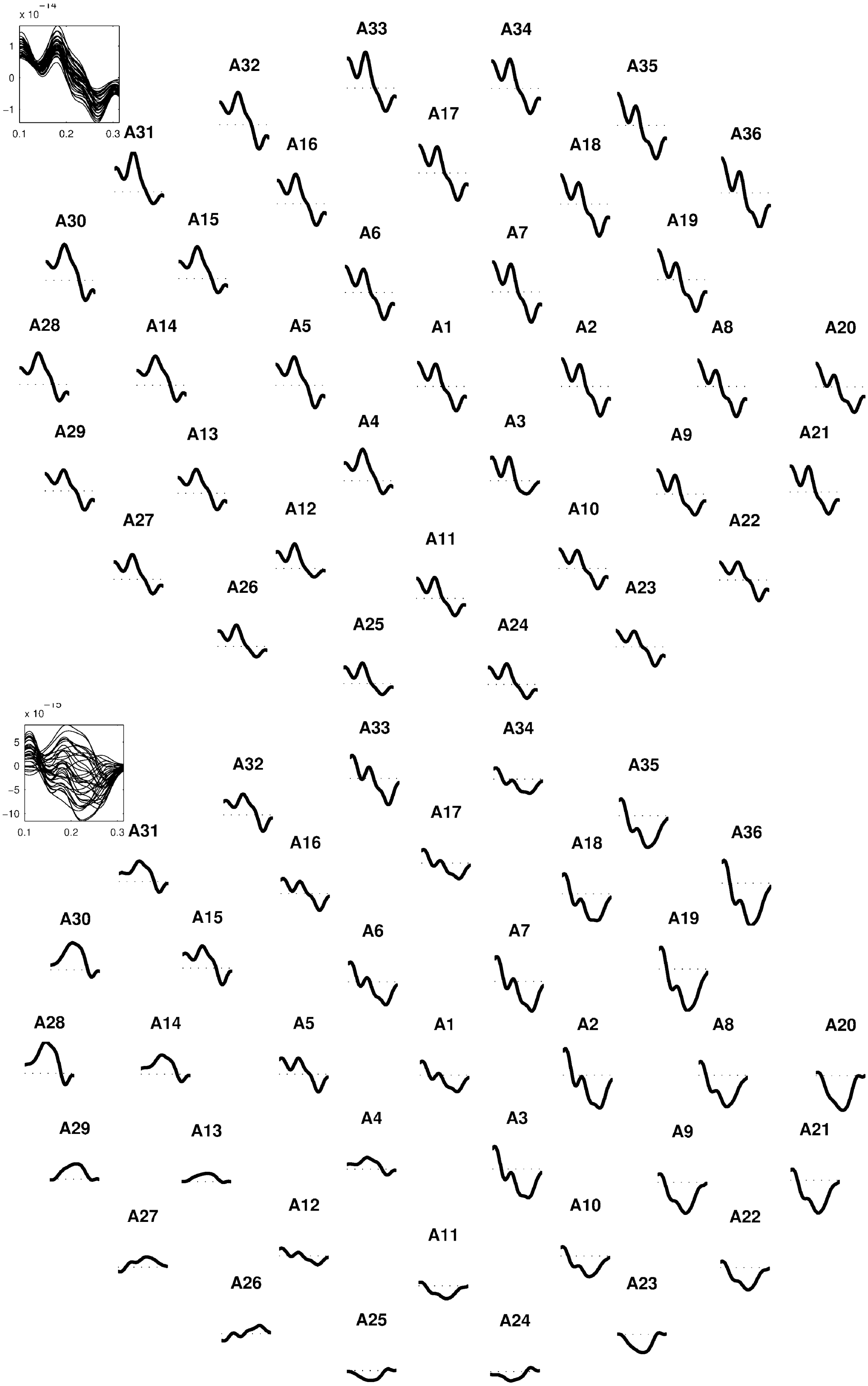


Fig. 12. Improvement of SNR and removal of maternal and fetal heart signals in the fetal auditory evoked response experiment. Top: prestimulus whitened followed by PCA processed data is clearly contaminated with fetal and maternal heart interference due to its nonstationarity. Bottom: MLE displays signal phase inversion across channels of the array as expected for a focal source. Rank three approximation was used in both cases.

V. Discussion

MLEs are known to be asymptotically optimal, that is, they are asymptotically unbiased and asymptotically attain the Cramer-Rao lower bound on estimator variance [17]. While the number of data records is always finite in any real application, the number of data samples available in the evoked response paradigm (JTN) is typically quite large relative to the number of free parameters being estimated (approximately $N^2/2 + NP + PL$), so we expect the maximum-likelihood approach to provide near optimal performance.

The estimation problem addressed in this paper exploits the known temporal structure that results because of the repetition of the signal in each epoch and the known signal bandwidth. Other known temporal characteristics can also be incorporated into the framework presented in this paper. The spatial structure of the signal of interest is assumed unknown, but low-rank, which allows this method to be applied in situations where the activity of interest is not dipolar or where the forward model is unknown, such as in fetal experiments. We have shown that the maximum-likelihood signal estimate may be obtained by whitening the component of the data with the known temporal structure using a spatial noise covariance matrix estimated from the component of the data that lies in the space orthogonal to the known temporal signal structure, followed by principal component expansion. This interpretation suggests that the maximum-likelihood approach is similar to use of prestimulus data for whitening the average across epochs followed by principle component expansion. Indeed, if the noise is stationary, then asymptotically the maximum-likelihood and prestimulus-based whitening approaches provide equivalent performance.

However, the maximum-likelihood approach provides significant advantages over prestimulus-based whitening when the noise is nonstationary because it estimates the whitening transformation from the same section of the data as that containing the signal. Thus, it does not require the spatial noise statistics to be the same in the prestimulus interval as in the interval containing the signal. This benefit is clearly illustrated in the fetal auditory evoked response example (Fig. 12). The noise is nonstationary in this case because of the fetal and maternal cardiac interference. The maximum-likelihood method effectively suppresses the noise because the noise statistics are estimated from the time interval containing the signal. In contrast, use of prestimulus noise statistics is not effective at suppressing the noise in the time interval of interest.

The maximum-likelihood method is also advantageous when limited data is available, because it estimates the whitening transformation using the maximum possible number of independent, signal-free data vectors, and thus it obtains the best possible estimate of the noise covariance matrix. Given J epochs containing T time samples, there are a total of JT independent spatial data vectors available for estimating the spatial statistics of the noise. Since the signal lies in the L-dimensional space spanned by the columns of $\mathbf{D}$, there are a total of $JT - L$ signal-free data components available for constructing $\mathbf{Q}_{\bar{\mathbf{D}}}$. All of these components are used by the maximum-likelihood method. In typical evoked response applications both J and T are on the order of 100, which means that on the order of 10 000 independent data vectors are available for forming the N by N matrix $\mathbf{Q}_{\bar{\mathbf{D}}}$. Even if the noise is stationary so that prestimulus data provides a valid estimate of the noise covariance matrix for whitening, the prestimulus approach uses substantially less data for estimating the noise covariance matrix than the maximum-likelihood approach. If T_p is the number of available prestimulus time samples, then the prestimulus whitening method uses JT_p independent data vectors. In contrast, the maximum-likelihood method uses $JT_p + JT - L$ independent data vectors, where in this case T denotes the number of samples in the poststimulus interval. Thus, the maximum-likelihood method provides better whitening, especially when JT_p is small, and consequently provides better quality signal estimates.

In general, good whitening requires that the number of independent data vectors used to estimate the noise covariance matrix be much larger than the dimension of the noise covariance matrix (N). This is illustrated in Fig. 8, which assumes $N = 37$. Use of 50 prestimulus data vectors does not provide significant performance improvement over simple averaging. However, significant improvement over averaging occurs when 200 prestimulus data vectors are utilized. The maximum-likelihood method uses $JT - L = 52\,060$ data vectors, which is three orders of magnitude greater than N, and effectively obtains the same performance as the known noise covariance matrix case.

Several special cases of the maximum-likelihood method are worth consideration. If there is no temporal information (e.g., bandwidth) available concerning the signal of interest, then we set $\mathbf{C} = \mathbf{I}$ and have $L = T$ so the number of data vectors used to form $\mathbf{Q}_{\bar{\mathbf{D}}}$ is $(J-1)T$. In this case, $\mathbf{X}_{\mathbf{D}}$ is N by T and simply contains the average of all epochs of data. The matrix $\mathbf{X}_{\bar{\mathbf{D}}}$ is N by $(J-1)T$ and contains the $J-1$ possible linearly independent weighted combinations of epochs where the weights sum to zero. In this case, the maximum-likelihood algorithm only exploits the assumption that the signal of interest is repeated in each epoch. At the other extreme, if only a single epoch of data is available, then $\mathbf{D} = \mathbf{C}$ and the method relies entirely on the temporal structure of the signal as represented by the basis vectors in $\mathbf{C}$. The noise statistics used for whitening are estimated from the components of the data in the space orthogonal to $\mathbf{C}$. If $\mathbf{C}$ spans a frequency band, then the noise statistics are estimated from frequency components in the data that lie outside the signal frequency band. There are $T - L$ data vectors available for constructing $\mathbf{Q}_{\bar{\mathbf{D}}}$ when $J = 1$ and at a minimum we require $T - L \geq N$ to obtain a nonsingular $\mathbf{Q}_{\bar{\mathbf{D}}}$. Effective whitening requires $T - L \gg N$, which suggests that single epoch analysis is limited to cases where the number of time samples is large relative to the number of sensors and the number of independent temporal components L is small.

The experimental results suggest that assuming the temporal noise covariance matrix is identity leads to an effective improvement in signal quality. Different results are likely if the noise is strongly correlated in time, such as when noise is due to strong alpha rhythm. Unknown temporal noise coloration affects the maximum-likelihood method proposed here and prestimulus-based whitening followed by principal component expansion in a similar fashion. The algorithm presented here is best suited for scenarios in which the noise is strongly correlated in the spatial dimension and offers significant advantages relative to prestimulus-based spatial whitening methods.

The framework presented in this paper for exploiting the temporal structure of evoked response data may be used to detect the presence of certain components in the data by following the detection strategies presented in [4]–[7]. Although we have assumed in this paper that the rank (P) of $\mathbf{H}$ is known or determined empirically, performing hypothesis testing on the rows of $\boldsymbol{\theta}$ to determine which are nonzero as discussed in [4] and [5] may yield an automated method for choosing P. Finally, source localization may also be performed by applying any published localization algorithm (e.g., [18] and [19]) after preprocessing with the maximum-likelihood method presented in this paper to improve the SNR.

Appendix

The temporal basis matrix $\mathbf{C}$ for the data in this paper has been constructed as follows, borrowing from the analysis in [20]. Consider a data vector $\mathbf{x}$ with all of its energy concentrated in the normalized frequency band $f_1 \leq |f| \leq f_2$. A low-rank approximation for $\mathbf{x}$ is $\tilde{\mathbf{x}} = \mathbf{U}\boldsymbol{\beta}$, where $\mathbf{U}$ is an $N \times p$ matrix with orthonormal columns and $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients. The normalized mean squared error of this approximation is

$$e^2 = E\{|\mathbf{x} - \tilde{\mathbf{x}}|^2\} = \mathrm{tr}\,\mathbf{R} - \mathrm{tr}\,\mathbf{U}^H\mathbf{R}\mathbf{U} \qquad (24)$$

where $\mathbf{R} = E\{\mathbf{x}\mathbf{x}^H\}$. It is well known that choosing the columns of $\mathbf{U}$ as the eigenvectors of $\mathbf{R}$, corresponding to the p largest eigenvalues minimizes the error.

In general, the matrix $\mathbf{R}$ is unknown. However, the error in (24) is upper bounded by assuming $\mathbf{R}$ corresponds to a bandpass white noise. Define $\mathbf{R}_C$

$$\mathbf{R}_C = \int_{f_1}^{f_2} \mathbf{c}(f)\mathbf{c}^H(f)df + \int_{-f_1}^{-f_2} \mathbf{c}(f)\mathbf{c}^H(f)df \qquad (25)$$

where $\mathbf{c}(f) = [1\; e^{j2\pi f}\; e^{j4\pi f}\; \ldots\; e^{j2(T-1)\pi f}]^T$. The eigenvectors of $\mathbf{R}_C$ corresponding to significant eigenvalues provide an efficient basis for signals that are bandlimited to $f_1 \leq |f| \leq f_2$. Thus, choosing $\mathbf{C}$ as eigenvectors of $\mathbf{R}_C$ is optimal in the mean squared error sense for a given frequency band of interest.

The number of basis vectors is approximately given by the time-bandwidth product $2(f_2 - f_1)T$. For example, for the one-second-long epochs used in the simulations, sampling frequency of 520.8 Hz and frequency band of 1–20 Hz, we get $L = 2(f_2 - f_1)T = 2(20 - 1)/520.8 \times 521 \approx 38$. Note that normalized frequency is used in this formula. This is approximately equal to $L = 40$ used in the paper. The latter number is obtained as the number of vectors required for the normalized error, $e^2/\mathrm{tr}\,\mathbf{R}_C$ to be less than 1%.

Acknowledgment

The authors would like to thank M. M. Maier for helping in the collection of fetal auditory data.

References

[1] J. Knuutila *et al.*, "Characterization of brain noise using a high sensitivity 7-channel magnetometer," in *Proc. 6th Int. Conf. Biomag.*, Tokyo, Japan, 1987, pp. 186–189.

[2] J. C. de Munck, H. M. Huizenga, L. J. Waldrop, and R. M. Heethaar, "Estimating stationary dipoles from MEG/EEG data contaminated with spatially and temporally correlated background noise," *IEEE Trans. Signal Processing*, vol. 50, pp. 1565–1572, July 2002.

[3] F. Bijma, J. C. de Munck, H. M. Huizenga, and R. M. Heethaar, "A mathematical approach to the temporal stationarity of background noise in MEG/EEG measurements," *NeuroImage*, vol. 20, pp. 233–243, 2003.

[4] E. Kelly and K. Forsythe, "Adaptive Detection and Parameter Estimation for Multidimensional Signal Models," Lincoln Laboratories, Lexington, MA, Tech. Rep. 848, 1989.

[5] M. Srivastava, *Methods of Multivariate Statistics*. NewYork: Wiley, 2002, ch. 10.

[6] A. Dogandžić and A. Nehorai, "Estimating evoked dipole responses in unknown spatially correlated noise with EEG/MEG arrays," *IEEE Trans. Signal Processing*, vol. 48, pp. 13–25, Jan. 2004.

[7] A. Dogandžić and A. Nehorai, "Generalized multirate analysis of variance," *IEEE Signal Processing Mag.*, vol. 20, pp. 39–54, Sept. 2003.

[8] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[9] K. Sekihara, D. Poeppel, A. Marantz, H. Koizumi, and Y. Miyashita, "Noise covariance incorporated MEG-MUSIC algorithm: A method for multiple-dipole estimation tolerant of the influence of background brain activity," *IEEE Trans. Biomed. Eng.*, vol. 44, pp. 839–847, Sept. 1997.

[10] K. Sekihara, "Adaptive beamformer source reconstruction," presented at the *Conf. ISBET 2003*, Santa Fe, NM, 2003.

[11] M. Spencer, R. Leahy, J. Mosher, and P. Lewis, "Adaptive filters for monitoring localized brain activity from surface potential time series," in *Conf. Rec. 26th Annu. Asilomar Conf. Signals, Systems, and Computers*, 1992, pp. 156–161.

[12] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Reading, MA: Addison-Wesley, 1991.

[13] R. Muirhead, *Aspects of Multivariate Statistical Theory*. NewYork: Wiley, 1982.

[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: The John Hopkins University Press, 1983.

[15] R. M. Gray, "On the asymptotic eigenvalue distribution of toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, 1972.

[16] D. Regan, "Form from motion parallax and form from luminance contrast: Vernier discrimination," *Spatial Vis.*, vol. 1, no. 4, pp. 305–318, 1986.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.

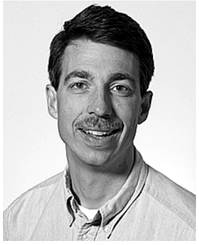
[18] J. Mosher and R. Leahy, "Recursive MUSIC: A framework for EEG and MEG source localization," *IEEE Trans. Biomed. Eng.*, vol. 45, pp. 1342–1354, Nov. 1998.

[19] B. D. Van Veen, W. van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering," *IEEE Trans. Biomed. Eng.*, vol. 44, pp. 867–880, Sept. 1997.

[20] B. D. Van Veen and L. L. Scharf, "Estimation of structured covariance matrices and multiple window spectrum analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1467–1472, Aug. 1990.

Boris V. Baryshnikov was born in Novgorod, Russia. He received the undergraduate degree with honors in physics/biophysics in 1997 from Moscow State University, Moscow, Russia. In 2000 and 2001, he received the M.S. degrees in medical physics and computer sciences, respectively, from the University of Wisconsin, Madison, WI. In 2004, he received the Ph.D. degree in medical physics under the supervision of Prof. R. T. Wakai in the biomagnetism laboratory at the University of Wisconsin-Madison.

His research interests include biomedical signal processing with emphasis on practical applications and development of algorithms and tools for data analysis and visualization.



Barry D. Van Veen (S'81–M'86–SM'97–F'02) was born in Green Bay, WI. He received the B.S. degree from Michigan Technological University, Houghton, MI, in 1983 and the Ph.D. degree from the University of Colorado, Boulder, in 1986, both in electrical engineering. He was an ONR Fellow while working on the Ph.D. degree.

In the spring of 1987, he was with the Department of Electrical and Computer Engineering at the University of Colorado-Boulder. Since August of 1987, he has been with the Department of Electrical and Computer Engineering at the University of Wisconsin-Madison and currently holds the rank of Professor. His research interests include signal processing for sensor arrays, adaptive filtering, wireless communications, and biomedical applications of signal processing. He coauthored, with S. Haykin, *Signals and Systems*, (New York: Wiley, 1st ed. 1999, 2nd ed. 2003).

Dr. Van Veen was a recipient of a 1989 Presidential Young Investigator Award from the National Science Foundation and a 1990 IEEE Signal Processing Society Paper Award. He served as an associate editor for the IEEE Transactions on Signal Processing, on the IEEE Signal Processing Society's Technical Committee on Statistical Signal and Array Processing, and on the Sensor Array and Multichannel Technical Committee. He received the Holdridge Teaching Excellence Award from the ECE Department at the University of Wisconsin in 1997.

Ronald T. Wakai was born in East Orange, NJ, in 1958. He received the B.A. degree with honors in physics from Cornell University, Ithaca, NY, in 1980 and the Ph.D. degree in physics from the University of Illinois, Urbana, in 1987.

Since then, he has been with the Department of Medical Physics at the University of Wisconsin, Madison, WI, where he is currently a Professor. His research interests include basic and technical aspects of fetal biomagnetism and adult MEG.