Space-Time Sparse Reconstruction for Magneto-/Electroencephalography

Andrew Bolstad *,1, Barry Van Veen, Robert Nowak

Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53715 USA

Abstract

This article presents a new spatio-temporal method for M/EEG source reconstruction based on the assumption that only a small number of events, localized in space and/or time, are responsible for the measured signal. Each space-time event is represented using a basis function expansion which reflects the most relevant (or measurable) features of the signal. The basis function expansion effectively models sources which span local regions of cortex and occupy specific frequency bands and/or time windows. This model of neural activity leads naturally to a Bayesian likelihood function which balances the model fit to the data with the complexity of the model, where the complexity is related to the number of included events. A novel Expectation-Maximization algorithm which maximizes the likelihood function is presented. The new method is shown to be effective on several MEG simulations of neurological activity as well as data from a self-paced finger tapping experiment.

Key words:

M/EEG source localization, Spatio-temporal priors, M/EEG Inverse problem

1 Introduction

The ill-posed nature of the M/EEG inverse problem is widely recognized (Baillet et al., 2001) and unique solutions are usually obtained by placing prior expectations on the nature of the brain activity associated with the measured data. In this paper we present an approach for solving the inverse problem that assumes the underlying brain activity is due to a small number or sparse set of space-time events. Each space-time event corresponds to activity that is concentrated in space and time. Sets of spatial and temporal basis functions are employed to represent the activity due to each space-time event. We seek to explain the measured data in terms of a small number of space-time events. This objective leads to a convex optimization problem, which we solve via a novel expectation-maximization (EM) algorithm.

Common approaches to the M/EEG inverse problem may be classified as either parametric or imaging methods (Baillet et al., 2001). Parametric or scanning methods use a small number of dipoles (Scherg and von Cramon, 1985),

* Corresponding author.
Email address: akbolstad@wisc.edu (Andrew Bolstad).
URL: homepages.cae.wisc.edu/∼bolstad (Andrew Bolstad).

[1] This work supported in part by the NIBIB under NIH award R21EB005473 and MIT Lincoln Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

multipoles (Mosher et al., 1999; Nolte and Curio, 2000), or cortical patches (Limpiti et al., 2006; Wagner et al., 2002, 2000) and scan over locations to find the best set of sources to represent the data. Examples of scanning methods include MUSIC (Mosher et al., 1992), beamforming (Van Veen et al., 1997), and maximum likelihood estimation (Dogandžić and Nehorai, 2000). One persistent challenge for scanning methods is choosing the number of sources in the solution.

The method proposed in this paper may be classified as an imaging technique. Imaging methods tessellate the brain using a large number of dipoles and attempt to simultaneously solve for all dipole amplitudes. This leads to the following linear model:

$$\mathbf{Y} = \mathbf{HX} + \mathbf{N} \tag{1}$$

where $\mathbf{Y}$ is an M x T matrix of spatial-temporal measurements from M sensors at T time points, $\mathbf{H}$ is an M x $3D$ (or M x D in the case of known dipole moment orientation) lead field matrix relating sensor measurements to D discrete dipole locations (voxels) on the cortex, $\mathbf{X}$ is the $3D$ x T signal matrix representing the unknown time courses and orientations of each dipole on the cortical surface, and $\mathbf{N}$ is a matrix of additive noise. The goal of imaging techniques is to estimate each element of $\mathbf{X}$ simultaneously given $\mathbf{Y}$ and $\mathbf{H}$. Typically M is on the order of several hundred, while D exceeds several thousand, so Eq. (1) is severely underdetermined. Furthermore, the effective rank of $\mathbf{H}$ is often significantly less than M.

The problem of nonuniqueness with imaging techniques is often reduced by

use of a Bayesian prior to favor some solutions over others. The prior is incorporated as a penalty term in a cost function which is minimized to estimate the neural activity. Under the assumption of white noise, these cost functions take the form:

$$\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}) \tag{2}$$

We use vectors $\mathbf{y}$ and $\mathbf{x}$ in Eq. 2 since the optimization, in most previous work, is solved separately at each time point, and solutions are often found only at specific time points selected by the user. The first term of Eq. (2) ensures that solutions explain the data well, while f acts as a penalty term. The constant λ balances the contribution of both terms. Note that Eq. (2) can be viewed as regularized inversion of the lead field matrix $\mathbf{H}$. A common choice of penalty is simply $f(\mathbf{x}) = \|\mathbf{W}\mathbf{x}\|_2^2$, which corresponds to a Gaussian prior on $\mathbf{x}$. In its simplest form, the minimum norm estimate (MNE) of cortical activity uses $\mathbf{W} = \mathbf{I}$ and is equivalent to Tikhonov regularization. To eliminate a bias toward deep sources, $\mathbf{W}$ can be a diagonal matrix whose elements are the norm of the corresponding column of $\mathbf{H}$ (Baillet et al., 2001). We refer to this choice of $\mathbf{W}$ as the *lead field normalizing matrix* since the matrix $\tilde{\mathbf{H}} = \mathbf{H}\mathbf{W}^{-1}$ consists of columns whose norm is one. The MNE has the simple closed form solution $\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{W}^T\mathbf{W})^{-1}\mathbf{H}^T\mathbf{y}$, but tends to severely blur activity across the cortex. The well known LORETA technique sets $\mathbf{W}$ to an approximation of the Laplacian operator (Pascual-Marqui et al., 1994). Like the MNE, LORETA gives blurry, low resolution reconstructions of neuronal activity. In an effort to reduce the spatial blurring of the MNE and LORETA, the FOCUSS algorithm reinforces strong sources while deempha-

sizing weak ones by iteratively updating $\mathbf{W}$. The diagonal elements of $\mathbf{W}$ are set equal to the estimated power of each source given by the previous solution (Gorodnitsky and Rao, 1997). FOCUSS does produce more focal estimates, but is sensitive to the initialization of $\mathbf{W}$ as well as to noise (Baillet et al., 2001). Both LORETA and FOCUSS can incorporate lead field normalization into $\mathbf{W}$ as well.

Another method aimed at producing focal solution to the inverse problem is the minimum current estimate (MCE) (Matsuura and Okabe, 1995, 1997; Uutela et al., 1999). In this case, the penalty in Eq. (2) is $f(\mathbf{x}) = \|\mathbf{W}\mathbf{x}\|_1$ with $\mathbf{W}$ either the identity or the lead field normalizing matrix. This penalty corresponds to an exponential prior on $\mathbf{x}$. Recently ℓ_1 norm regularization has received much attention in many fields because of its ability to recover sparse solutions (i.e., solutions in which most components are zero) using computationally tractable methods; ℓ_1 regularized estimates are solutions of convex optimization problems. The criticism of ℓ_1 regularization is that it produces solutions which are "spiky" in space and inconsistent with physiological expectations. A more recent twist on ℓ_1 regularization in MEG is the VESTAL approach (Huang et al., 2006). Here dipole orientations are not fixed, but constrained to an octant via $\mathbf{W}$, a diagonal sign matrix. Singular value decomposition of both the lead field matrix $\mathbf{H}$ and the data matrix $\mathbf{Y}$ are used to stabilize the solution and mollify the "spiky" nature of minimum ℓ_1 solutions. Researchers have also proposed ℓ_p norm regularization with $0 < p < 1$ and $1 < p < 2$ as well (Auranen et al., 2005; Jeffs et al., 1987). The optimization problem is not convex for $0 < p < 1$. In general these optimization problems are more difficult to solve than $p = 1, 2$.

The "space-time sparse" (STS) regularization we present here is closely re-

lated to ℓ_1 minimization. The STS approach seeks a solution composed of a small number of "space-time events" (STEs) which we define below. STS differs from existing ℓ_1 approaches in that each STE is represented by a group of basis functions. We use spatial basis functions to represent extended activity on the cortex and also eliminate the need to disambiguate neighboring dipolar sources, and temporal basis functions to confine signals to specific frequency bands of interest and limit their duration. The use of basis functions allows variation in the activity of the implicit dipoles involved in a single STE. Use of spatial and temporal bases in the STS regularizer effectively combines the spatial and temporal aspects of the neural activity estimation problem. The STS formulation automatically determines both where and when significant activity occurs. Hence, the detection of STEs eliminates the need for users to specify times in the measured signal where significant activity is believed to exist. We present STS regularization as well as a novel Expectation-Maximization (EM) algorithm for solving the resulting optimization problem. A simple criterion for guiding the selection of an appropriate regularization parameter λ is incorporated into the optimization procedure. The effectiveness of the method is illustrated on simulations as well as human subject data from a self-paced finger tapping experiment.

The remainder of the paper is organized as follows. In Materials and Methods, we first introduce the concept of spatial-temporal basis function expansions of $\mathbf{X}$ in Eq. (1) and use the basis functions to define STEs. Next, we present the STS regularization technique for solving the inverse problem, as well as an EM algorithm for solving the STS regularized least squares problem. The Results section describes several MEG simulations, including a set of randomized source simulations, and application of STS to MEG recordings of a subject

performing self-paced finger tapping. We conclude with a discussion of the STS method and possible extensions.

The following notation is used throughout. Matrices and vectors are represented by upper and lower case boldface letters, respectively. Matrix inversion is represented by superscript -1, while transposition is represented by superscript T. Classic vector p norms are denoted $\|\cdot\|_p$, and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. We denote the Kronecker product using the $\otimes$ operator while vec () concatenates columns of a matrix to produce a column vector. For a review of algebra involving $\otimes$ and vec (), see Brewer (1978).

2 Materials and Methods

2.1 Spatio-Temporal Dictionaries for the Inverse Problem

The STS approach to the M/EEG inverse problem relies on representing the cortical signal $\mathbf{X}$ as a linear combination of physiologically meaningful basis functions from carefully crafted spatial and temporal dictionaries. In general, a spatio-temporal expansion of $\mathbf{X}$ can be written:

$$\mathbf{X} = \mathbf{S}\boldsymbol{\Theta}\mathbf{T}^T \tag{3}$$

where the columns of $\mathbf{S}$ are a dictionary containing spatial bases, the columns of $\mathbf{T}$ are a dictionary containing temporal bases, and the $(i,j)^{th}$ element of coefficient matrix $\boldsymbol{\Theta}$ specifies the contribution associated with the i^{th} spatial basis modulated by the j^{th} temporal basis. If both $\mathbf{S}$ and $\mathbf{T}$ are invertible then there is a one-to-one mapping between $\boldsymbol{\Theta}$ and $\mathbf{X}$, that is $\boldsymbol{\Theta} = \mathbf{S}^{-1}\mathbf{X}\mathbf{T}^{-T}$.

However, $\mathbf{S}$ and $\mathbf{T}$ may be chosen to intentionally restrict the signal to lie in specific spatial and/or temporal subspaces in order to incorporate prior expectations into the solution of the M/EEG inverse problem.

Given a spatio-temporal expansion in the form of Eq. (3), many methods can be used to solve the reformulated problem for $\mathbf{\Theta}$. In particular, regularized least squares problems analogous to Eq. (2) can be formulated with some function of $\mathbf{\Theta}$ as the regularization term. However, the power of a spatio-temporal expansion comes from the ability to design the bases and the penalty term concurrently. Incorporating prior expectations into both the dictionaries and the criterion for choosing $\mathbf{\Theta}$ offers potential for improved quality solutions.

Our approach assumes that M/EEG signals are the result of a small number of space-time events (STEs) which occur over limited time spans and occupy contiguous, local areas of cortex. This assumption is reflected in the following spatio-temporal expansion:

$$\mathbf{X} = \mathbf{S}\mathbf{\Theta}\mathbf{T}^T = \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_2 & \dots & \mathbf{S}_I \end{bmatrix} \begin{bmatrix} \mathbf{\Theta}_{1,1} & \mathbf{\Theta}_{1,2} & \dots & \mathbf{\Theta}_{1,J} \\ \mathbf{\Theta}_{2,1} & \mathbf{\Theta}_{2,2} & \dots & \mathbf{\Theta}_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Theta}_{I,1} & \mathbf{\Theta}_{I,2} & \dots & \mathbf{\Theta}_{I,J} \end{bmatrix} \begin{bmatrix} \mathbf{T}_1^T \\ \mathbf{T}_2^T \\ \vdots \\ \mathbf{T}_J^T \end{bmatrix} = \sum_{i=1}^{I} \sum_{j=1}^{J} \mathbf{S}_i \mathbf{\Theta}_{i,j} \mathbf{T}_j^T \tag{4}$$

Here the spatial and temporal dictionaries, as well as the coefficient matrix, have been partitioned into space-time events. The columns of each submatrix $\mathbf{S}_i$ are an orthogonal basis for measurable activity spanning a local region of cortex. Likewise, the columns of each $\mathbf{T}_j$ are an orthogonal basis for a

temporal subspace, such as a frequency band or time interval. Cortical activity that lies in the vector space defined by a single spatial submatrix $\mathbf{S}_i$ and a single temporal submatrix $\mathbf{T}_j$ defines a single STE and is represented by the coefficients in $\mathbf{\Theta}_{i,j}$. We seek parsimonious sets of spatial and temporal bases to describe candidate STEs and formulate a criterion for finding solutions that involve a small number of non-zero $\mathbf{\Theta}_{i,j}$.

This philosophy is not limited to a specific set of STEs. The temporal bases in $\mathbf{T}_j$ may be chosen using wavelets, Gabor functions, Fourier bases, etc., to span time intervals and bandwidths consistent with the temporal charactersitics of the signals of interest. Spatial bases $\mathbf{S}_i$ for events that activate a limited area of cortex include cortical patches (Limpiti et al., 2006; Wagner et al., 2002) and multipolar expansions (Mosher et al., 1999; Nolte and Curio, 2000). The spatial bases are selected consistent with the spatial attributes of interest.

Our examples in this paper employ spatial event bases derived from cortical patches. This approach results in event bases that are locally supported and particularly parsimonious. We describe our basis construction procedure here to exemplify several key attributes of appropriate bases. Assume the i^{th} cortical patch contains a collection of dipoles whose lead fields and amplitudes are concatenated into a patch lead field matrix $\mathbf{H}_i$ and amplitude vector $\boldsymbol{\alpha}_i$, respectively. Hence, the signal $\mathbf{y}_i$ measured at the sensors due to the i^{th} patch is $\mathbf{y}_i = \mathbf{H}_i\boldsymbol{\alpha}_i$. The bases representing events in this patch should only describe the space of possible activity distributions $\boldsymbol{\alpha}_i$ that contribute to the measurement. To this end, take the singular decomposition of $\mathbf{H}_i = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where the columns of $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix of singular values ordered from largest to smallest. Typically $\mathbf{H}_i$ is low rank (Limpiti et al., 2006) so a small number of the

singular values are significant.

The columns of $\mathbf{V}$ are a natural basis for the patch signal $\boldsymbol{\alpha}_i$. Suppose $\boldsymbol{\alpha}_i = \mathbf{v}_k$ where $\mathbf{v}_k$ is the k^{th} column of $\mathbf{V}$. The corresponding signal measured at the sensor is:

$$\mathbf{y}_i = \mathbf{u}_k \sigma_k \tag{5}$$

where $\mathbf{u}_k$ is the k^{th} column of $\mathbf{U}$ and σ_k is the k^{th} largest singular value. Since each $\mathbf{u}_k$ is unit norm, the strength of the measured signal is given by σ_k, and we see that spatial bases $\mathbf{v}_k$ associated with small singular values make small contributions to the measured signal. Given the typically low signal to noise ratio of M/EEG, the measurable spatial component of an STE from the i^{th} patch is adequately represented by the right singular vectors corresponding to significant singular values. Hence, we construct the spatial basis $\mathbf{S}_i$ from the "significant" right singular vectors associated with the i^{th} patch.

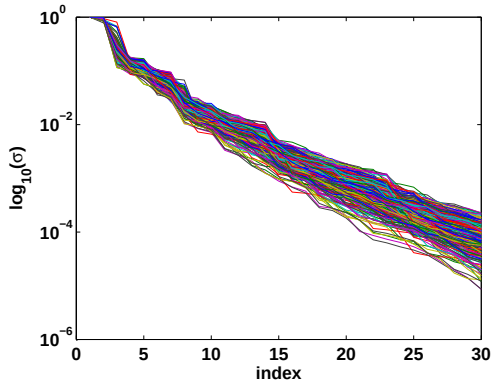


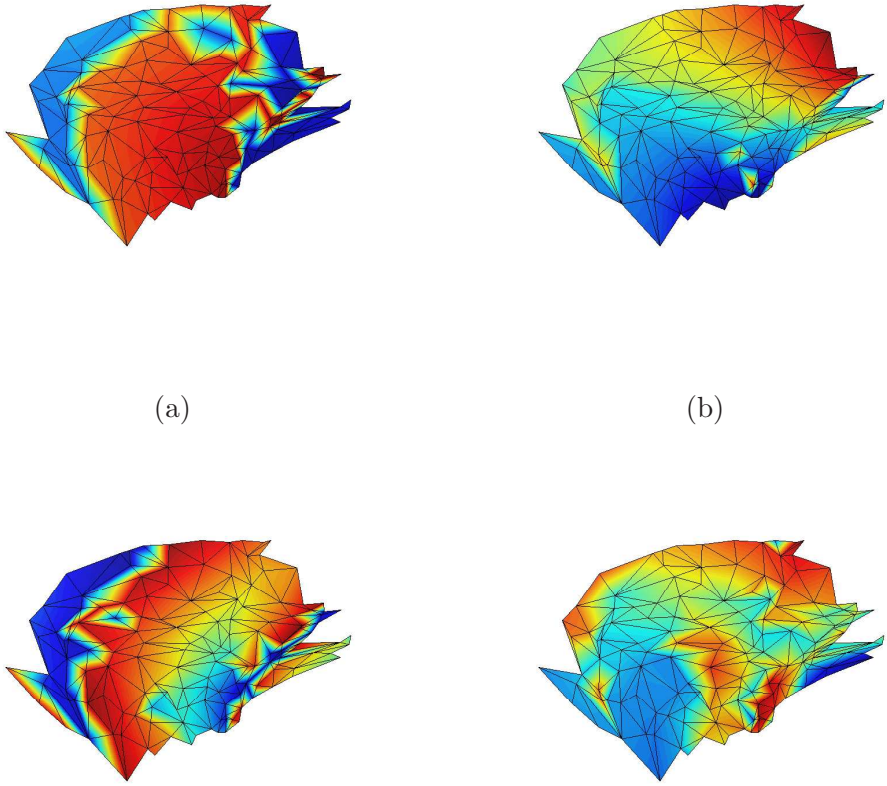
Fig. 1. Normalized singular values associated with 393 patch lead field matrices. The singular values decay rapidly in every patch.

Figure 1 depicts the singular values for all 393 lead field matrices corresponding to patches of geodesic radius 20 mm tiling both hemispheres of a subject

with 50% overlap. The lead fields are computed for the 275 channel CTF MEG system. Note that in each case the 7^{th} and higher singular values are an order of magnitude smaller than the largest singular value. Furthermore, the first two or three singular values are five times larger than the higher order singular values. Figure 2 illustrates the spatial pattern associated with several right singular vectors for a typical patch by depicting each element of $\mathbf{v}_k$ at the corresponding dipole location. The singular vectors associated with the largest singular values tend to represent low resolution features within the patch and the level of detail or fine structure represented by a singular vector increases as the corresponding singular values decrease. Selecting the number of significant singular values and vectors thus involves a tradeoff between representing the details of the spatial distribution of activity within the patch and having adequate SNR to estimate the detail. Similar tradeoffs exist with other approaches to choosing spatial bases. For example, the contributions of higher order terms in a multipole expansion also decay rapidly, so the SNR dictates the number of terms that can be reliably estimated and used as spatial bases.

2.2 Space-Time Sparse Regularization

The motivation behind Space-Time Sparse (STS) regularization is the assumption that the activity of interest is a collection of a small number of space-time events which occur over limited time spans and occupy local areas of cortex. This assumption, combined with appropriate spatial and temporal dictionaries, implies that the true underlying signal should have $\mathbf{\Theta}_{i,j} = \mathbf{0}$ for the majority of pairs (i, j). To encourage such a sparse solution, we seek to solve an optimization problem which balances measurement error with the



(a) (b)

(c) (d)

Fig. 2. Color cortical activity pattern for singular vectors $\mathbf{v}_k$ of a representative patch. (a) $\mathbf{v}_1$, largest singular value. (b) $\mathbf{v}_3$, third largest singular value. (c) $\mathbf{v}_5$, fifth largest singular value. (d) $\mathbf{v}_7$, seventh largest singular value. The activity patterns associated with larger singular values tend to represent low resolution features within the patch, while those associated with smaller singular values represent fine structure.

number of nonzero blocks $\mathbf{\Theta}_{i,j}$. One way to do this is to penalize the least squares problem based on the number of STEs. That is, solve:

$$\min_{\mathbf{\Theta}} \|\mathbf{Y} - \mathbf{HS\Theta T}^T\|_F^2 + \lambda \sum_{i,j} \mathbf{1}_{\{\mathbf{\Theta}_{i,j} \neq 0\}} \tag{6}$$

where $\mathbf{1}_{\{E\}}$ is the indicator function of event E and the parameter $\lambda > 0$ balances error with the complexity (or non-sparsity) of the solution. Equation (6) represents a highly nonconvex problem requiring an exhaustive search of all possible combinations of active blocks $\mathbf{\Theta}_{i,j}$. Exhaustive search is computationally prohibitive even for relatively small problems. Thus we seek a suitable proxy for the regularizer $\sum_{i,j} \mathbf{1}_{\{\mathbf{\Theta}_{i,j} \neq 0\}}$.

The regularization term in Eq. (6) can be viewed as the ℓ_0 pseudonorm of a vector whose entries are the ℓ_p norms of each block, $\|\mathrm{vec}\,(\mathbf{\Theta}_{i,j})\,\|_p$, for any $p > 0$. The nonconvex nature of the problem is due to the ℓ_0 pseudonorm. Many recent results have suggested that relaxing the nonconvex ℓ_0 pseudonorm of a vector to the convex ℓ_1 norm in regularized least squares problems provides a very good approximation to the desired regularization (Candes and Tao, 2005; Donoho, 2006; Tropp, 2006). In the spirit of these results, we relax the ℓ_0 block penalty to an ℓ_1 block penalty:

$$\min_{\mathbf{\Theta}} \|\mathbf{Y} - \mathbf{HS\Theta T}^T\|_F^2 + \lambda \sum_{i,j} \|\mathrm{vec}\,(\mathbf{\Theta}_{i,j})\,\|_p \tag{7}$$

We still must select an appropriate p to capture the magnitude of the contribution of each block. Setting $p \geq 1$ results in a convex optimization problem. The "obvious" choices for a convex optimization are $p = 1$, 2, or ∞. If $p = 1$, then the block structure of Eq. (4) is lost and the penalty in Eq. (7) becomes $\|\mathrm{vec}\,(\mathbf{\Theta})\,\|_1$. This encourages a solution which is sparse in the number of nonzero coefficients in $\mathbf{\Theta}$, but not necessarily sparse in terms of the number of STEs. We prove in (Bolstad et al., 2007) that when one STE is present, the $p = 2$ case with lead field normalization (see Section 2.6) provides a higher probability of activating the correct STE than the $p = \infty$ case. Using $p = 2$ in

Eq. (7) and noting that $\|\text{vec}\left(\boldsymbol{\Theta}_{i,j}\right)\|_2 = \|\boldsymbol{\Theta}_{i,j}\|_F$ leads to estimate $\hat{\mathbf{X}} = \mathbf{S}\hat{\boldsymbol{\Theta}}\mathbf{T}^T$ where:

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \|\mathbf{Y} - \mathbf{HS}\boldsymbol{\Theta}\mathbf{T}^T\|_F^2 + \lambda \sum_{i,j} \|\boldsymbol{\Theta}_{i,j}\|_F \tag{8}$$

This objective function is convex since it is a sum of convex functions. At least one minimizer exists since the objective function is positive (i.e., bounded below) and convex for all $\boldsymbol{\Theta}$. The next section describes a novel EM algorithm for finding $\hat{\boldsymbol{\Theta}}$.

2.3 Expectation-Maximization Algorithm

The gradient of $\sum_{i,j} \|\boldsymbol{\Theta}_{i,j}\|_2$ is discontinuous so classical gradient-based minimization approaches to Eq. (8) can not be used without modification. Instead, we have developed an Expectation-Maximization (EM) algorithm which treats the objective of Eq. (8) as a negated Bayesian log-likelihood function where the penalty can be interpreted as the logarithm of a prior distribution on $\boldsymbol{\Theta}$. EM finds the maximum a posteriori (MAP) estimate of $\boldsymbol{\Theta}$. The approach described here is a generalization of the EM algorithm developed by Figueiredo and Nowak (2003). The pioneering work on EM was published by Dempster, Laird, and Rubin (Dempster et al., 1977).

The EM algorithm monotonically increases the Bayesian likelihood $p(\mathbf{Y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})$ over $\boldsymbol{\Theta}$. Assuming white Gaussian noise and the density $p(\boldsymbol{\Theta}) \propto \exp\left\{-\frac{\lambda}{2\sigma^2}\sum_{i,j}\|\boldsymbol{\Theta}_{i,j}\|_F\right\}$, the negated Bayesian log-likelihood is proportional to $\frac{\|\mathbf{Y}-\mathbf{HS}\boldsymbol{\Theta}\mathbf{T}^T\|_F^2}{2\sigma^2} + \frac{\lambda}{2\sigma^2}\sum_{i,j}\|\boldsymbol{\Theta}_{i,j}\|_F$. Multiplying through by $2\sigma^2$, we see that the EM algorithm is monotonically decreasing the function $\|\mathbf{Y}-\mathbf{HS}\boldsymbol{\Theta}\mathbf{T}^T\|_F^2 +$

$\lambda \sum_{i,j} \|\mathbf{\Theta}_{i,j}\|_F$, which is the objective of Eq. (8).

The key aspect of solving an estimation problem via the EM algorithm is to identify a so-called "hidden data set" which, if known, would make the problem easy to solve. EM alternates between solving the easy problem given an estimate of the hidden data and updating the estimate of the hidden data. We identify hidden data $\mathbf{Z}$ for the STS problem based on the decomposition:

$$\mathbf{Z} = \mathbf{\Theta} + \alpha \mathbf{N}_1 \tag{9}$$
$$\mathbf{Y} = \mathbf{HSZT}^T + \mathbf{N}_2 \tag{10}$$

Here the noise matrix $\mathbf{N}$ of Eq. (1) is decomposed into two independent random matrices $\mathbf{N}_1$ and $\mathbf{N}_2$ with distributions $\text{vec}\,(\mathbf{N}_1) \sim \mathcal{N}(\mathbf{0}, \mathbf{I} \otimes \mathbf{I})$ and $\text{vec}\,(\mathbf{N}_2) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I} - \alpha^2\mathbf{TT}^T \otimes \mathbf{HSS}^T\mathbf{H}^T)$. The scalar α^2 must satisfy $\alpha^2 < \sigma^2\|\mathbf{TT}^T\|^{-1}\|\mathbf{HSS}^T\mathbf{H}^T\|^{-1}$ so that a valid covariance matrix exists for $\mathbf{N}_2$. With this decomposition, estimating $\mathbf{\Theta}$ given $\mathbf{Z}$ amounts to a classic denoising problem (Donoho, 1995) in which the density of the desired signal $\mathbf{\Theta}$ is characterized by the penalty function $\sum_{i,j} \|\mathbf{\Theta}_{i,j}\|_F$. Equation 10 is then used to update the estimate of $\mathbf{Z}$.

The EM algorithm alternates between an Expectation Step (E-Step) and a Maximization Step (M-Step). The E-Step calculates the Q-function: $Q(\mathbf{\Theta}|\hat{\mathbf{\Theta}}^{(k-1)}) \equiv E[\log p(\mathbf{Y}, \mathbf{Z}|\mathbf{\Theta})|\mathbf{Y}, \hat{\mathbf{\Theta}}^{(k-1)})]$ where the expectation is over $\mathbf{Z}$ and $\hat{\mathbf{\Theta}}^{(k-1)}$ is the estimate from the previous iteration. The M-Step maximizes $Q(\mathbf{\Theta}|\hat{\mathbf{\Theta}}^{(k-1)}) + \log p(\mathbf{\Theta})$ over $\mathbf{\Theta}$ to produce $\hat{\mathbf{\Theta}}^{(k)}$ [2]. This produces a series of

[2] In ML estimation, the M-Step is to maximize the Q-function alone; however, in the case of MAP estimation, the log likelihood of Θ is included as well (see (Dempster et al., 1977)).

iterates $\hat{\mathbf{\Theta}}^{(k)}$ which are guaranteed to converge to a stationary point provided that $Q(\mathbf{A}|\mathbf{B})$ is continuous in both $\mathbf{A}$ and $\mathbf{B}$ (Wu, 1983).

In the E-Step, we calculate the Q-function using the measured data and the estimate $\hat{\mathbf{\Theta}}^{(k-1)}$:

$$\begin{aligned} Q(\mathbf{\Theta}|\hat{\mathbf{\Theta}}^{(k-1)}) &= E[\log p(\mathbf{Y}, \mathbf{Z}|\mathbf{\Theta})|\mathbf{Y}, \hat{\mathbf{\Theta}}^{(k-1)})] \\ &= K - \frac{\|\mathbf{\Theta} - \hat{\mathbf{Z}}^{(k)}\|_F^2}{2\alpha^2} \end{aligned} \tag{11}$$

where K is not a function of $\mathbf{\Theta}$ and:

$$\begin{aligned} \hat{\mathbf{Z}}^{(k)} &= E[\mathbf{Z}|\mathbf{Y}, \hat{\mathbf{\Theta}}^{(k-1)}] \\ &= \hat{\mathbf{\Theta}}^{(k-1)} + \frac{\alpha^2}{\sigma^2}\mathbf{S}^T\mathbf{H}^T(\mathbf{Y} - \mathbf{HS}\hat{\mathbf{\Theta}}^{(k-1)}\mathbf{T}^T)\mathbf{T} \end{aligned} \tag{12}$$

Thus, specifying $Q(\mathbf{\Theta}|\hat{\mathbf{\Theta}}^{(k-1)})$ amounts to calculating $\hat{\mathbf{Z}}^{(k)}$ using Eq. (12).

The M-Step is a generalization of the M-Step for an ℓ_1 penalized least squares problem (see Figueiredo and Nowak (2003)) since the penalty in Eq. (8) is a generalization of the ℓ_1 norm. We seek to solve the following problem:

$$\begin{aligned} \hat{\mathbf{\Theta}}^{(k)} &= \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\hat{\mathbf{\Theta}}^{(k-1)}) + \log p(\mathbf{\Theta}) \\ &= \arg\min_{\mathbf{\Theta}} \left\{ \frac{\|\mathbf{\Theta} - \hat{\mathbf{Z}}^{(k)}\|_F^2}{2\alpha^2} + \frac{\lambda}{2\sigma^2} \sum_{i,j} \|\mathbf{\Theta}_{i,j}\|_F \right\} \end{aligned} \tag{13}$$

This is the maximum a posteriori estimate of $\mathbf{\Theta}$ given $\hat{\mathbf{Z}}^{(k)}$ assuming $p(\mathbf{\Theta}) \propto \exp(-\frac{\lambda}{2\sigma^2} \sum_{i,j} \|\mathbf{\Theta}_{i,j}\|_F)$. The key feature facilitating the solution of Eq. (13) is the separability of the problem into blocks corresponding to the blocks of Eq. (4):

$$\frac{\|\mathbf{\Theta} - \hat{\mathbf{Z}}^{(k)}\|_F^2}{2\alpha^2} + \frac{\lambda}{2\sigma^2}\sum_{i,j}\|\mathbf{\Theta}_{i,j}\|_F = \sum_{i,j}\frac{\|\mathbf{\Theta}_{i,j} - \hat{\mathbf{Z}}_{i,j}^{(k)}\|_F^2}{2\alpha^2} + \frac{\lambda}{2\sigma^2}\|\mathbf{\Theta}_{i,j}\|_F \quad (14)$$

We solve Eq. (13) by finding the $\mathbf{\Theta}_{i,j}$ that minimizes $\|\mathbf{\Theta}_{i,j} - \hat{\mathbf{Z}}_{i,j}^{(k)}\|_F^2 + \frac{\alpha^2\lambda}{\sigma^2}\|\mathbf{\Theta}_{p,q}\|_F$ for every pair (i,j). The single block problem for the $(i,j)^{th}$ block is rewritten in simplified notation using $\mathbf{z} = \text{vec}\left(\hat{\mathbf{Z}}_{i,j}^{(k)}\right)$ and $\boldsymbol{\theta} = \text{vec}\left(\mathbf{\Theta}_{i,j}\right)$ as $f(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \mathbf{z}\|_2^2 + \frac{\alpha^2\lambda}{\sigma^2}\|\boldsymbol{\theta}\|_2$. Since $f(\boldsymbol{\theta})$ is strictly convex, it has a unique minimizer $\boldsymbol{\theta}^*$. We show in Appendix A that:

$$\boldsymbol{\theta}^* = \begin{cases} (1 - \frac{\alpha^2\lambda}{2\sigma^2\|\mathbf{z}\|_2})\mathbf{z}, \text{ if } \|\mathbf{z}\|_2 > \frac{\alpha^2\lambda}{2\sigma^2} \\ \\ \mathbf{0}, \qquad\qquad \text{otherwise} \end{cases} \quad (15)$$

Thus we summarize the EM algorithm as:

For $k = 1, 2, \ldots,$

$$\begin{aligned} \hat{\mathbf{Z}}^{(k)} &= \hat{\mathbf{\Theta}}^{(k-1)} + c\mathbf{S}^T\mathbf{H}^T(\mathbf{Y} - \mathbf{H}\mathbf{S}\hat{\mathbf{\Theta}}^{(k-1)}\mathbf{T}^T)\mathbf{T} \\ \hat{\mathbf{\Theta}}^{(k)} &= \arg\min_{\mathbf{\Theta}}\left\{\|\mathbf{\Theta} - \hat{\mathbf{Z}}^{(k)}\|_F^2 + c\lambda\sum_{i,j}\|\mathbf{\Theta}_{i,j}\|_F\right\} \end{aligned} \quad (16)$$

where $c = \|\mathbf{T}\mathbf{T}^T\|^{-1}\|\mathbf{H}\mathbf{S}\mathbf{S}^T\mathbf{H}^T\|^{-1}$ since we choose α^2 at the upper bound $\alpha^2 = \sigma^2\|\mathbf{T}\mathbf{T}^T\|^{-1}\|\mathbf{H}\mathbf{S}\mathbf{S}^T\mathbf{H}^T\|^{-1}$.

As described in (Wu, 1983), continuity of the Q function and the existence of a lower bound to Eq. (8) implies convergence to a stationary point. Since Eq. (8) is convex, the only stationary point is the global minimum, so $\hat{\mathbf{\Theta}}^{(k)}$ converges to the global minimizer. Various stopping criteria may be used to terminate the algorithm, such as lack of sufficient change in the objective function or

satisfaction of the subgradient condition, i.e., the subdifferential of Eq. (8) at $\hat{\mathbf{\Theta}}$ contains $\mathbf{0}$ within a set tolerance.

An advantage of our EM algorithm approach is that it provides an easy means to calculate solutions $\hat{\mathbf{\Theta}}(\lambda)$ for a wide variety of λ. Since a small change in λ doesn't typically produce a large change in $\hat{\mathbf{\Theta}}(\lambda)$, the solution $\hat{\mathbf{\Theta}}(\lambda_1)$ can be used to initialize the algorithm when solving for $\hat{\mathbf{\Theta}}(\lambda_2)$. If $\lambda_2 \approx \lambda_1$, then this initialization point will be close to $\hat{\mathbf{\Theta}}(\lambda_2)$. Furthermore, we show in Appendix B that $\hat{\mathbf{\Theta}}(\lambda) = \mathbf{0}$ for $\lambda \geq \lambda_{MAX}$, where:

$$\lambda_{MAX} = 2\| \max_{i,j} \mathbf{S}_i^T \mathbf{H}^T \mathbf{Y} \mathbf{T}_j \|_2 \tag{17}$$

depends on the data $\mathbf{Y}$. Thus, we solve for a range of λ by setting $\hat{\mathbf{\Theta}}(\lambda_{MAX}) = \mathbf{0}$, and then solve for $\lambda = (1-\epsilon)\lambda_{MAX}$ using $\mathbf{0}$ as the initial guess $\hat{\mathbf{\Theta}}^{(0)}$. Next we use $\hat{\mathbf{\Theta}}((1-\epsilon)\lambda_{MAX})$ as a starting point to find $\hat{\mathbf{\Theta}}((1-2\epsilon)\lambda_{MAX})$ and continue in this manner until we arrive at the desired λ. Selection of λ is discussed in the next section.

2.4 Penalty Weight Parameter Selection

The quality of the estimate $\hat{\mathbf{\Theta}}$ given by Eq. (8) depends on the penalty weight parameter λ. Intuitively, large λ corresponds to heavy reliance on the prior and gives a very sparse (or all zero) estimate, while small λ corresponds to heavy reliance on the noisy data and gives solutions with activity spread all over the cortical surface. Clearly, both of these extremes are undesirable. In general, user expertise may provide the best results for selecting λ, and in many cases it may be desirable to consider a range of solutions from multiple values of

λ. However, automatic selection of λ is necessary for objective performance evaluation and to provide a starting point for user based selection.

We have developed a simple heuristic to automate selection of λ based on the number of nonzero $\mathbf{\Theta}_{i,j}$, that is, the cardinality of $\hat{\mathcal{A}}(\lambda) = \{(i,j)|\hat{\mathbf{\Theta}}_{i,j}(\lambda) \neq \mathbf{0}\}$. The intuition behind our heuristic is the following. Suppose no signal is present, $\mathbf{Y} = \mathbf{N}$, and we solve Eq. (8). Since white noise is spread roughly equally through all subspaces, no single block $\mathbf{\Theta}_{i,j}$ can account for a significant portion of the measured data $\mathbf{Y}$. Thus, as λ decreases, the number of blocks in $\hat{\mathcal{A}}(\lambda)$ increased rapidly to reduce the error term in Eq. (8). In contrast, when a signal described by a small number of STEs is present, the signal portion of the measured data can be explained by the corresponding blocks, and initially the cardinality of $\hat{\mathcal{A}}(\lambda)$ increases relatively slowly as λ decreases. Once the number of active blocks is able to explain the signal component of the data, then additional blocks attempt to represent noise and the number of active blocks will increase rapidly as λ is decreased further. This behavior is illustrated in Fig. 3. Note that the number of STEs increases rapidly as λ decreases for the noise only case, while it initially increases very slowly when signal components are present.

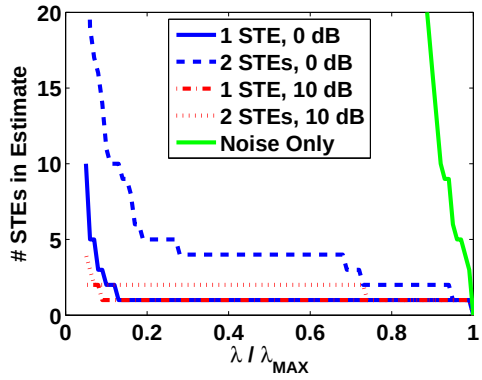


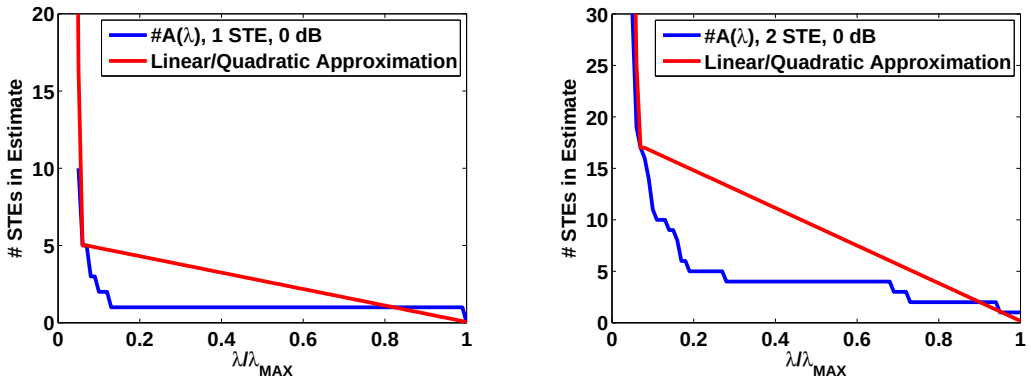
Fig. 3. Characteristic behavior of the number of STEs identified by STS vs. λ/λ_{MAX}.

These observations suggest a simple strategy analogous to the L-curve method of regularization in which we select the value of λ where the number of STEs begins to increase rapidly with further decrease in λ. We implement one such strategy by fitting the cardinality of $\hat{\mathcal{A}}(\lambda)$, which we denote $\sharp A(\lambda)$, with two curve segments: a linear function for large λ and quadratic function for small λ. The linear function connects the points $(\lambda_{MAX}, 0)$ and $(\lambda, \sharp A(\lambda))$. The quadratic function is uniquely defined by two fixed endpoints since it is purely quadratic (i.e., has no linear term). The endpoints used for the quadratic are $(\lambda, \sharp A(\lambda))$ and $(0, \sharp \hat{A}(0))$, where $\sharp \hat{A}(0)$ is a rough estimate of $\sharp A(0)$ found by dividing the rank of the forward model by the number of bases in each block. This is the number of active blocks required to produce an estimate which fits any measured data exactly. We vary λ to find the value that minimizes the squared error between $\sharp A(\lambda)$ and the piecewise approximation. This value of λ is used to identify the point at which the transition between fitting signal and noise occurs.

This method of λ selection requires that $\sharp A(\lambda)$ be calculated for small enough λ such that the characteristic shape of Fig. 3 is observed before applying the heuristic. Two examples of the heuristic λ selection are shown in Fig. 4. These two examples, as well as the additional examples in the following section, suggest that the λ selected with this heuristic is typically smaller than the "optimal" λ. Hence, this heuristic provides a good lower bound.

2.5 Application to Non-white Noise

A key assumption behind the STS method as presented so far is that the signal is contaminated with white noise. This is rarely the case in real problems.



(a) (b)

Fig. 4. Example of heuristic applied to 0 dB simulated data with (a) 1 STE and (b) 2 STEs.

We accommodate colored noise by first whitening the data with the noise covariance matrix. Suppose the noise in Eq. (1) is distributed $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_N)$. Taking the matrix square root $\mathbf{R}_N = \mathbf{R}_N^{1/2}\mathbf{R}_N^{T/2}$, the whitened data is:

$$\mathbf{Y}_w = \mathbf{R}_N^{-1/2}\mathbf{Y} = \mathbf{R}_N^{-1/2}\mathbf{H}\mathbf{X} + \mathbf{R}_N^{-1/2}\mathbf{N} = \mathbf{H}_w\mathbf{X} + \mathbf{N}_w \tag{18}$$

where $\mathbf{N}_w \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. STS is applied to the whitened system. In practice $\mathbf{R}_N$ is usually unknown, so whitening requires that $\mathbf{R}_N$ be estimated from available data.

2.6 Lead Field Normalization

Lead field normalization is employed to eliminate bias toward sources closer to the sensors. After whitening the data and expanding the signal with STS basis functions, our signal model is:

$$\mathbf{Y}_w = \mathbf{H}_w \mathbf{S} \mathbf{\Theta} \mathbf{T}^T + \mathbf{N}_w \tag{19}$$

We normalize the lead fields using diagonal matrix $\mathbf{W}$ whose entries equal the ℓ_2 norm of the corresponding column of $\mathbf{H}_w\mathbf{S}$. Thus we have:

$$\mathbf{Y}_w = \mathbf{H}_w \mathbf{S} \mathbf{W}^{-1} \mathbf{W} \mathbf{\Theta} \mathbf{T}^T + \mathbf{N}_w \tag{20}$$

and the EM algorithm solves for $\mathbf{\Theta}' = \mathbf{W}\mathbf{\Theta}$. Here the columns of the matrix $\mathbf{H}_w\mathbf{S}\mathbf{W}^{-1}$ are unit norm due to normalization. The columns of $\mathbf{T}$ may be normalized as well, though typically the $\mathbf{T}_j$ are constructed so that $\mathbf{T}_j^T\mathbf{T}_j = \mathbf{I}$ and thus normalization of $\mathbf{T}$ is not necessary.

2.7 Debiasing

If it is known a priori which blocks of Eq. (4) are nonzero and which are zero, then a minimum norm solution to the inverse problem takes the form:

$$\begin{aligned} &\min \|\mathbf{Y} - \mathbf{H}\mathbf{S}\mathbf{\Theta}\mathbf{T}^T\|_F^2 + \gamma\|\mathbf{\Theta}\|_F^2 \\ &\text{s.t. } \mathbf{\Theta}_{i,j} = \mathbf{0},\ \forall (i,j) \notin \mathcal{A} \end{aligned} \tag{21}$$

where the set of nonzero or "active" blocks are denoted by $\mathcal{A} = \{(i,j) | \mathbf{\Theta}_{i,j} \neq \mathbf{0}\}$. This problem can be solved in closed form by building an appropriate matrix from the columns of $\mathbf{HS}$ and $\mathbf{T}$ corresponding to active blocks. The minimum norm criterion ensures a unique solution even if this matrix is numerically rank deficient. Note that the blurring normally associated with minimum norm solutions is not as problematic here since we are restricting the solution to active blocks.

The STS regularized problem is designed to select a space-time sparse solution from the set of all solutions with comparable residual noise, and thus the nonzero blocks of the solution to Eq. (8) are an estimate of the active set: $\hat{\mathcal{A}} = \{(i,j)|\hat{\mathbf{\Theta}}_{i,j} \neq \mathbf{0}\} \approx \mathcal{A}$. This estimate can be used as the active set in Eq. (21) to produce a refined estimate of $\mathbf{\Theta}$. This process is referred to as debiasing in ℓ_1 regularized inverse problems since it removes some of the bias introduced by the penalty term in Eq. (8). The debiasing step may be performed without lead field normalization. Also, different bases expansions for the active blocks may be used in debiasing than those used for the original STS regularization. The debiasing step is not necessary for the STS regularization approach, especially if source localization is the primary goal.

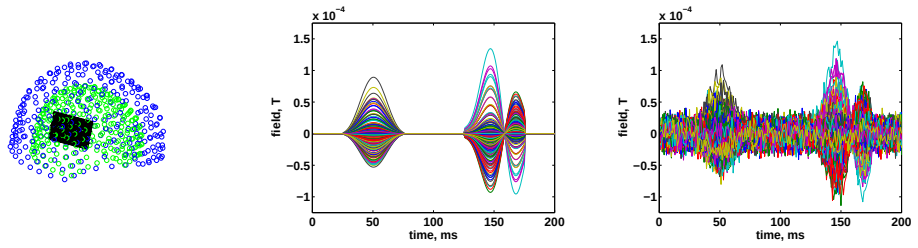
3 Results

We illustrate specific examples of STS reconstruction in this section. We begin by comparing the STS approach to standard ℓ_1 regularization on a "brain" with simplified geometry to illustrate the benefits of the STS approach. Next, we describe the method we use for whitening multi epoch data in more realistic scenarios. We then demonstrate STS reconstruction on realistic, simulated MEG signals. These simulations show that the method works well even though the decomposition in Eq. (4) might not exactly match the signal and when the noisy data is whitened using an estimated noise covariance matrix. We also illustrate the characteristics of the regularization parameter selection heuristic described in Section 2.4. Finally, we apply STS reconstruction to measurements collected during a self-paced finger tapping experiment.

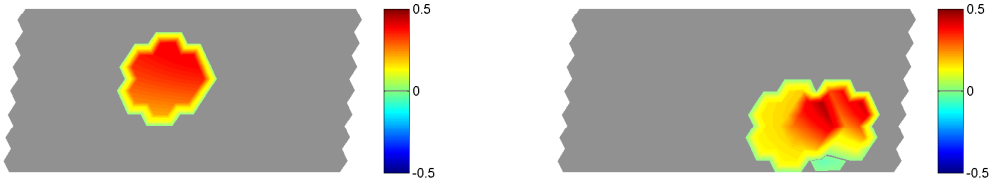
3.1 Two Dimensional Example

Simultaneous space and time ℓ_1 regularized inversion of Eq. (1) is computationally intractable for typical cortical surface tesselations, so, we replace the realistic cortical surface with a planar surface consisting of 397 dipoles. We calculate MEG lead fields generated by these dipoles for the sensor configuration shown in Fig. 5(a). The spatial STS basis functions consist of the cortical patch bases described in Section 2.1 with three bases per patch and 50% overlap. We use seven overlapping time windows of 50 ms, each with seven temporal basis functions spanning 0 – 40 Hz for the temporal STS bases. We simulate a space-time sparse signal consisting of three active STEs. Figure 5(b) – 5(e) shows the spatial activation pattern of the signal as well as the simulated measurements.

White noise is added to these measurements at three SNRs: -5, 0, and 5 dB. In all our simulations, we define SNR $= \|\mathbf{HX}\|_F^2 / E\{\|\mathbf{N}\|_F^2\}$. Figure 5(c) illustrates a representative noisy data set with SNR = 0 dB. At each noise level, we simulate ten trials and calculate the STS solution and the ℓ_1 regularized inversion of Eq. (1) using penalty $\|\mathrm{vec}\,(\mathbf{X})\,\|_1$. The debiased STS solution is calculated with a fixed, non-optimal $\lambda = 0.5\lambda_{MAX}$. The lowpass temporal bases used in the STS approach automatically filter out high frequency noise from the signal estimate. To give the ℓ_1 approach the same advantage, we first project the measured data $\mathbf{Y}$ onto the subspace spanned by $\mathbf{T}$. Next, we minimize $\|\mathbf{Y} - \mathbf{HX}\|_F^2 + \lambda\|\mathrm{vec}\,(\mathbf{X})\,\|_1$ over a large range of λ and calculate de-biased estimates $\mathbf{X}_{\ell_1}(\lambda)$. We then select the clairvoyant λ which minimizes the mean squared error: $\mathbf{X}_{\ell_1} = \min_\lambda \|\mathbf{X}_{\ell_1}(\lambda) - \mathbf{X}_{true}\|_F^2$ where $\mathbf{X}_{true}$ is the true signal $\mathbf{X}$. Thus we are comparing the best possible ℓ_1 estimate (in a



(a) (b) (c)

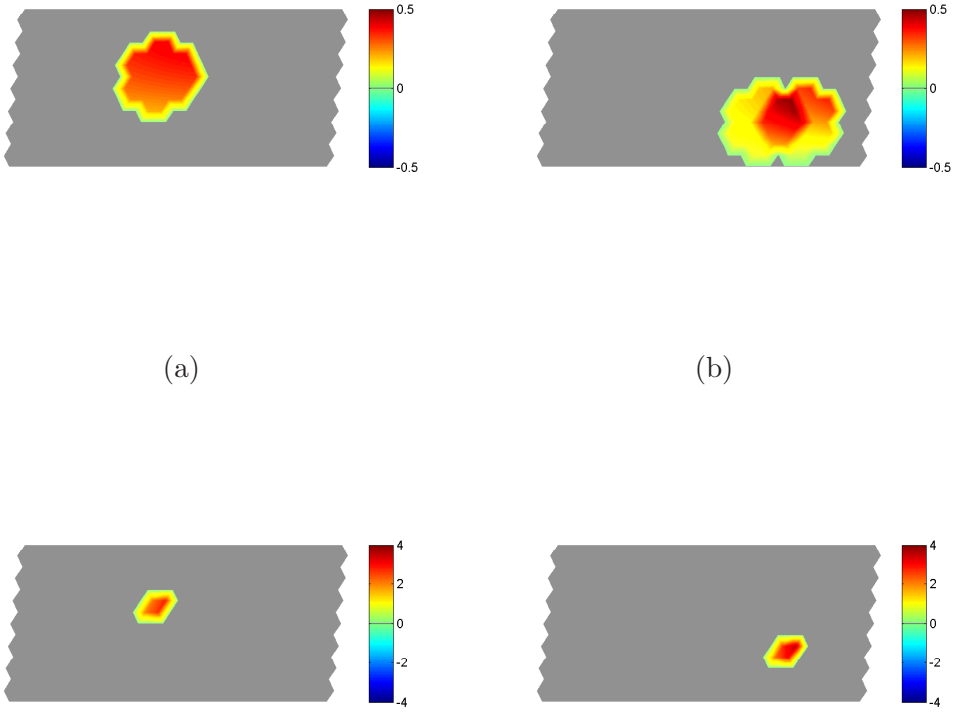


(d) (e)

Fig. 5. Two dimensional example. The cortical geometry is restricted to a plane to visualize differences between STS and ℓ_1 approaches. (a) Location of plane relative to MEG sensors. (b) Butterfly plot of noise free simulated measurements. (c) Example of noisy data for SNR=0 dB. (d) Spatial activation pattern at 50 ms. (e) Spatial activation pattern at 147 ms.

mean squared error sense) to an STS estimate with arbitrary regularization parameter.

The STS reconstruction outperforms optimal ℓ_1 reconstruction at all noise levels. The mean squared error optimal ℓ_1 estimate is the all zero solution, $\mathbf{X}_{\ell_1} = \mathbf{0}$, for all trials due to the low rank lead field matrix and high noise level. The MSE is simply $\|\mathbf{X}\|_F^2 = 131$. The STS estimate, on the other hand, produced estimates with average MSEs of 12.3 for the -5 dB case, 8.05 for the



(a) (b)

(c) (d)

Fig. 6. Comparison of the STS and ℓ_1 regularized inversion (with debiasing) on a simplified cortical geometry. (a) STS, 50 ms (b) STS, 147 ms (c) ℓ_1, 50 ms (d) ℓ_1, 147 ms.

0 dB case, and 3.53 for the 5 dB case. Figure 6 compares the STS estimate to an ℓ_1 estimate with a hand-picked λ for SNR = 5 dB. Figure 6 shows near perfect reconstruction using the STS approach, while the ℓ_1 reconstruction (with $\lambda = .2\lambda_{MAX}$)) produces estimates which are spiky (too focal) with incorrect magnitude.

The remaining results in this paper are based on whitening the data using the sample covariance matrix estimated from a multi epoch data set. Other methods, such as using resting state or prestimulus measurements, may be used if multiple epochs are not available. Suppose we have Q epochs of data $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(Q)}$. Assuming that the signal repeats in each epoch, we remove the mean from each epoch to approximate Q realizations of noise: $\tilde{\mathbf{N}}^{(q)} = \mathbf{Y}^{(q)} - Q^{-1} \sum_{q'=1}^{Q} \mathbf{Y}^{(q')}$. M/EEG data is typically lowpass filtered during data collection and thus successive time samples in the $\tilde{\mathbf{N}}^{(q)}$ are not independent. Hence, we select a subset of R time points, i.e., columns, from each realization $\tilde{\mathbf{N}}^{(q)}$ which are separated by the coherence time of the low pass filter. Denote these columns by $\tilde{\mathbf{n}}^{(q,r)}$, $r = 1, 2, \ldots, R$. The spatial sample covariance matrix of the noise is then estimated as:

$$\hat{\mathbf{R}}_N = (QR)^{-1} \sum_{q,r} \tilde{\mathbf{n}}^{(q,r)} (\tilde{\mathbf{n}}^{(q,r)})^T \tag{22}$$

We use $\hat{\mathbf{R}}_N$ in Eq. (18) to approximately whiten the spatial component of the noise. Note that the noise is not temporally white since the data is typically lowpass or bandpass filtered. However, noise coloration due to temporal filtering may be neglected because the temporal bases comprising $\mathbf{T}$ typically span the same bandwidth as the filtered data. In the results presented here we ignore possible temporal noise correlation within the processing band.

3.3 Space-Time Basis Function Selection for Examples

In the remaining examples we illustrate the attributes of our STS reconstruction approach using STEs defined by cortical patches and fixed time intervals having a common bandwidth. As noted previously there are many possible ways to define STEs. We choose patches on the cortical surface extracted from an MRI of the subject. Both hemispheres are tiled with patches of geodesic radii of either 10 mm or 20 mm. Note that geodesic distances calculated using the edges of triangles representing the surface are greater than the true geodesic distance. Figures 7 and 10 illustrate typical patches. We choose three spatial bases per patch for each STE since the majority of measurable activity in any patch is represented by the first three singular vectors, as indicated in Fig. 1.

The temporal events divide the total time extent of interest into seven time windows that overlap by 50%. The bases for each of these seven events are wavelet bases that span the same frequency band as the signal. We have found that wavelet bases result in smooth transitions at the window edges. We use 32 wavelets per time interval, so the number of coefficients associated with each STE is 96.

3.4 Simulated Example 1

We test the STS approach using simulated data from a 275 channel CTF MEG system. The simulated cortical signal consists of three STEs: activation of a patch in both the left and right primary auditory cortex with a slight offset in time followed by a patch of activity in the motor cortex. All three

simulated patches have geodesic radii of roughly 10 mm. The measurements are simulated using Eq. (1), with $\mathbf{N}$ consisting of colored noise drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{I} \otimes \mathbf{R})$ distribution where $\mathbf{R}$ was calculated from the prestimulus portion of an MEG experiment. The noise level was set to achieve an average SNR of -6 dB. Figure 7 illustrates the time courses and spatial extents of the three space-time events, as well as the noisy measurements.

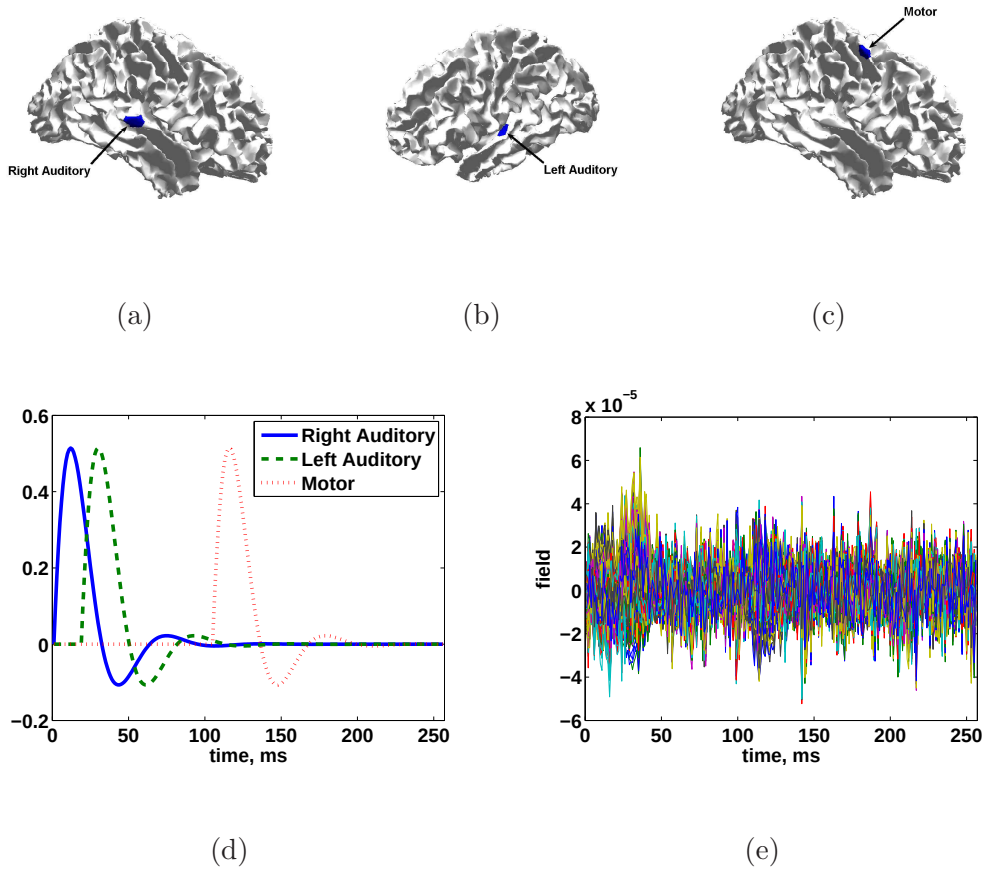
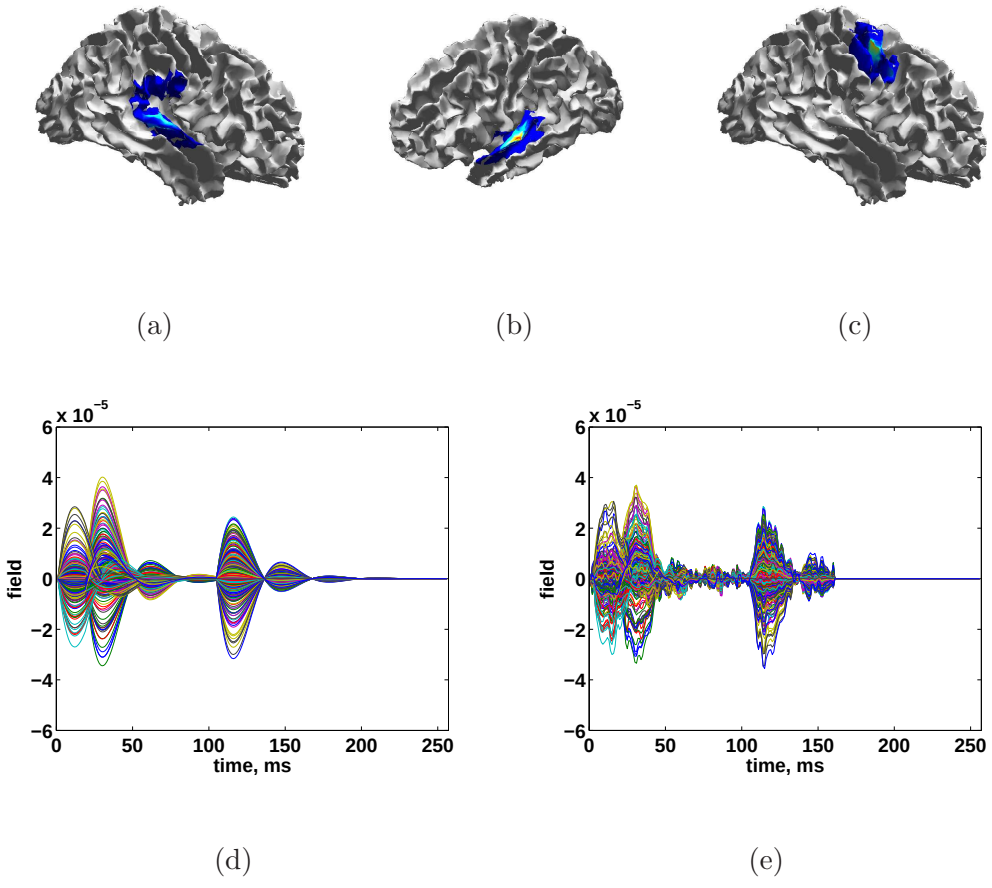


Fig. 7. Simulated three source scenario. True source location and spatial extent of sources at times (a) 12, (b) 30, and (c) 116 ms. (d) True source time courses. (e) Simulated signals in noise for SNR = -6 dB.

After whitening the data using the estimated noise covariance matrix as described in Section 2.5, we apply STS reconstruction to the simulated signal using the basis functions described in Section 3.3 assuming 20 mm geodesic

radii patches and wavelet temporal bases spanning windows of 64? ms. This results in 393 spatial patches with 3 bases each and 7 overlapping time windows with 32 bases each for a total of 2751 candidate STEs. Note that the basis functions used to perform STS reconstruction differ from the actual spatial and temporal characteristics of the signal. The patches used in the STS reconstruction are larger than the true signal patches and do not use the same center locations as the true signal patches. Furthermore, the temporal evolution of the simulated signal is not constructed from wavelet bases.



(a) (b) (c)

(d) (e)

Fig. 8. Signal estimate from de-biased STS: Reconstructed cortical activity at (a) 12, (b) 30, and (c) 116 ms. (d) Noise free sensor measurements (e) De-biased STS reconstruction of sensor measurements.

The heuristic of Section 2.4 is used to select λ. The result of debiasing the STS

reconstruction for this λ is shown in Fig. 8. All three STEs are recovered at the correct locations in space and time. Activity is spread spatially around the true location. Some of this spreading is due to the use of larger patches in the STS reconstruction, while some spreading is inherent to the ill-posed nature of the inverse problem. We also point out that some temporal spreading occurs as a result of choosing fixed time windows. This is evident between 97 and 105 ms, where the motor cortex exhibits low-level activity in the reconstruction, though the true motor activity starts at 106 ms. The fixed time window in which the motor cortex is active begins at 97 ms.

3.5 Simulated Example 2

Next a series of simulations is conducted to test how well the STS procedure reconstructs random combinations of STEs. We use the same basis decomposition as in the first example above (393 spatial patches, 7 time windows) as well as the same forward model; however, for these simulations three active blocks are selected uniformly at random from the available 2751 blocks for each trial. The coefficients of the activity for each active block are i.i.d. standard Gaussian random variables. Here there is no mismatch – the simulated STEs are contained in the set of STEs used for reconstruction. This facilitates comparison of the STS reconstruction to the true signal. We also use white noise and 0 dB SNR. Table 3.5 summarizes the results of 20 different simulations.

False negatives are true STEs which are not identified by the STS reconstruction, while false positives are STEs identified by the STS reconstruction that do not overlap in both space and time with a true STE. Active STEs identi-

Table 1

Number of false negative and false positive STEs for each simulation in Example 2 using the heuristic λ selection from Sec. 2.4 and the optimal λ. Each trial contained three randomly selected true STEs out of 2751 potential STEs.

Trial	Heuristic λ		Optimal λ	
	False Positives	λ/λ_{max}	False Positives	λ/λ_{max}
1	2	.09	1	.40-.62
2	5	.08	0	.32-.35
3	2	.07	0	.17-.25
4	3	.08	3	.18
5	4	.10	0	.41-.74
6	1	.06	0	.14-.15
7	2	.09	0	.22-.8
8	1	.07	0	.13-.19
9	8	.09	1	.34-.68
10	6	.08	1	.18-.19,.21-.26
11	12	.08	1	.20-.24
12	10	.08	2	.32-.55
13	2	.07	0	.10-.22
14	1	.08	0	.25
15	5	.08	1	.16-.29,.31
16	3	.08	1	.22-.23,.28-.30
17	5	.09	0	.29-.66
18	2	.08	0	.32-.44
19	10	.07	3	.38-.43
20	6	.07	3	.23-.25

fied in the STS reconstruction which overlap true STEs in space and time are ignored. For the columns labeled "Heuristic λ," we select λ using the heuristic described in Section 2.4. In all trials, the λ selected by the heuristic had no false negatives. This suggests that the heuristic may be used as a lower bound to aid user selection of λ. The range specified as "Optimal λ" are those values of λ that achieve no false negatives with the minimum number of overlapping and false positive STEs. This "optimum" value of λ is chosen based on knowledge of the true source activity and is included to illustrate characteristics of the heuristic λ and the STS algorithm. Figures 9 and 10 depict the true and reconstructed activity for trials 6 and 11 using the heuristic λ. Trial 6 is one of the best trials (1 false positive) while Trial 11 is one of the worst (12 false positives). In Trial 6, the heuristic λ selection produces a single false positive, but in the de-biased reconstruction this patch has low amplitude (Fig. 9(f)). In trials 9-12 and 19, the heuristic λ performs significantly worse than the optimal, which suggests there may be value in considering more sophisticated schemes for selecting λ. The results of Trial 11, shown in Fig. 10, indicate that although there are a significant number of false positives, the excess activity is in the general vicinity of the true activity.

Conducting simulations with random STEs provides some insight into how STS performs for a variety of signal conditions. In general, the STS approach does a good job identifying active space-time patches of cortical activity. The optimal λ performs very well for all trials with only three or fewer false positives. These false positives are typically spatial patches very close to true sources, but not overlapping them. In some cases when activity occurs on the interhemispheric fissure, STS identifies activity on the opposite hemisphere, immediately across from the true source. Such errors in spatial localizations

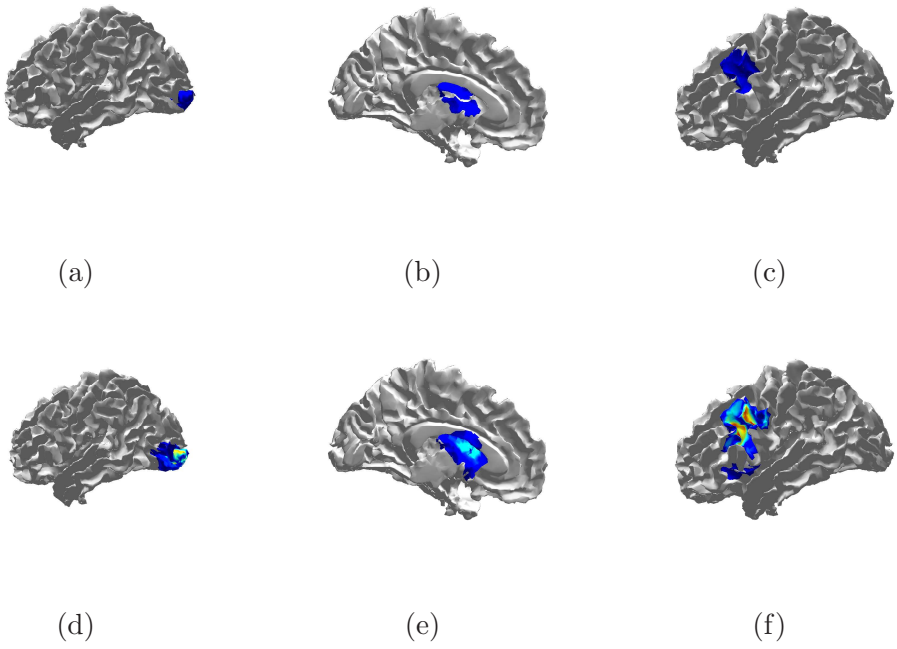
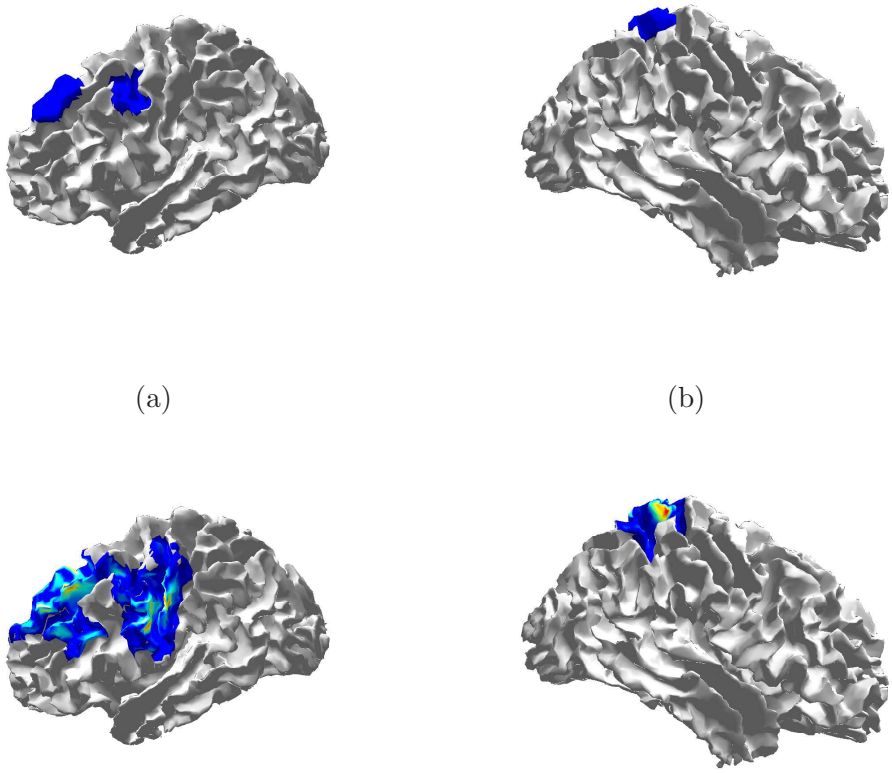


Fig. 9. True signal and de-biased STS reconstruction of Trial 6 with heuristic λ. (a) Location of true source activity between samples 1 and 64. (b) Location of source activity between samples 161 and 224. (b) Location of source activity between samples 193 and 256. (d) De-biased STS reconstruction of activity at sample 30. (e) De-biased STS reconstruction of activity at sample 200. (f) De-biased STS reconstruction of activity at sample 200 illustrating the false positive patch.

are a result of the low rank, noisy M/EEG inverse problem.

The majority of the simulations reported in Table 3.5 involve temporal overlap of spatially distinct STEs; only Trials 8, 13, and 17 contain STEs with no temporal overlap. In most cases, distinct STEs overlap by 50% in time. In Trials 1, 10, 11, 19, and 20, however, two distinct STEs occupy the exact same time window. Hence, the STS reconstruction is able to identify simultaneously occurring, spatially distinct sources. If two STEs overlap in time and are spatially close, false positives tend to occur between them, though the amplitudes of the false positive STEs are typically small compared to the true sources.



(a) (b)

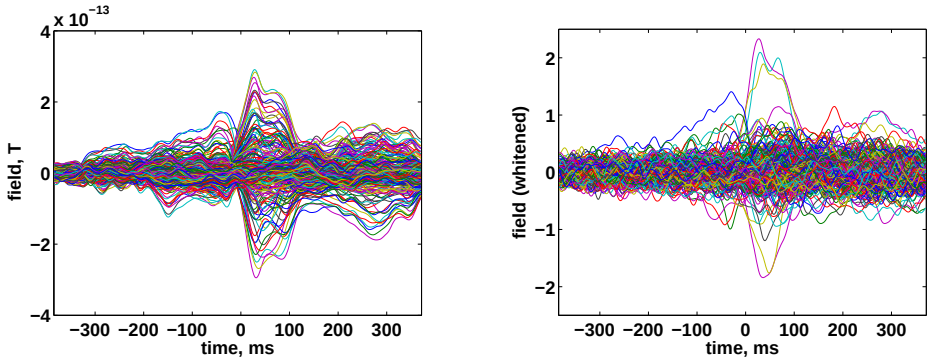
(c) (d)

Fig. 10. Excess spatial spreading associated with heuristic λ being smaller than optimal in Trial 11. (a) Location of true source activity between samples 161 and 224. (b) Location of source activity between samples 193 and 256. (c) De-biased STS reconstruction of activity at sample 200. (d) De-biased STS reconstruction of activity at sample 250.

3.6 Self-Paced Finger Tapping Experiment

We also applied STS to MEG measurements of a subject performing self-paced finger tapping. The data were collected with a Magnes 3600 whole head system at a sampling rate of 678 Hz and band pass filtered to lie between 2 and 40 Hz. The measured data were grouped into 80 epochs of 755 ms duration (512 samples) referenced to the tap at time 0 detected via a switch closure. These

epochs were then used to perform whitening on the average of all epochs as described in Section 3.2. The averaged sensor measurements before and after whitening are shown in Fig. 11. Cortical patches of 20 mm geodesic radii are used to construct the spatial bases $\mathbf{S}_i$. Each of the seven temporal events span 189 ms (128 samples).

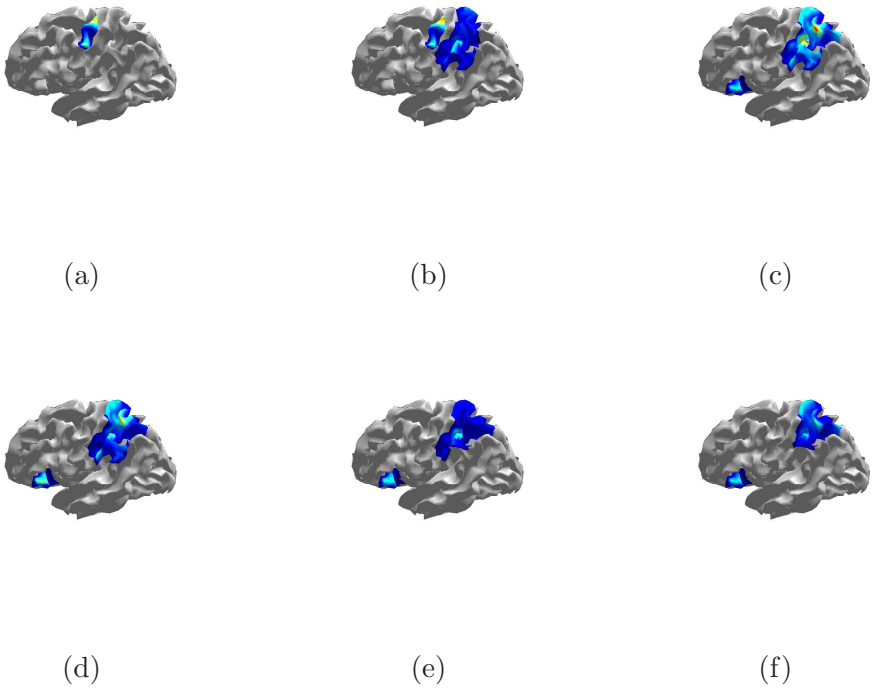


(a) (b)

Fig. 11. Recorded MEG from self-paced finger taping. (a) Averaged data. (b) Whitened average data.

The de-biased reconstruction shown in Fig. 12 is calculated using the heuristic of Section 2.4 to select λ. The estimate of activity begins about 200 ms pre-tap and persists through the end of the measurement period. The earliest activity occurs in the left motor cortex, which corresponds with the action of initiating the finger tap. The next region to become active spans the somatosensory and posterior parietal cortex. The somatosensory activity likely reflects the sensation of pushing down the switch. The posterior parietal cortex is involved in spatial relationships, body image, and movement planning (Bear et al., 2001). The parietal lobe activity begins at low levels about 100 ms before the switch closes and continues throughout the measurement period. It is possible that the early onset of activity in the parietal lobe is a result of temporal spreading throughout an active time window. Beginning at about 0 ms the

STS reconstruction includes a small patch on the left frontal lobe. Some weak activity is also identified in the right hemisphere, primarily in the parietal lobe. This is consistent with earlier studies' finding that the right parietal cortex plays a role in spatial processing (Culham et al., 2006).



(a) (b) (c)

(d) (e) (f)

Fig. 12. De-biased STS reconstruction of self-paced finger taping with heuristic λ. (a) -105 ms. (b) -31 ms. (c) 28 ms. (d) 80 ms. (e) 176 ms. (f) 264 ms.

4 Discussion

The STS approach searches for a small number of STEs spanning an extended area of cortex and limited time duration which best explain the measurements. By minimizing the number of events rather than the number of active dipoles, STS solutions avoid "spiky" spatial and temporal artifacts typical of ℓ_1 methods. The basis function representation of each event reduces the number of

parameters by only representing activity that is measurable at the M/EEG sensors. Thus, our approach incorporates the inherent limitations of the forward problem to formulate a more tractable inverse problem. The temporal basis functions naturally constrain the solution based on features of interest, such as the time-frequency characteristics. Use of both spatial and temporal bases allows STS to automatically determine both where and when significant activity occurs. The STS criterion involves the weighted combination of the squared fitting error and a penalty designed to encourage solutions consisting of a small number of STEs. The resulting cost function is convex and is minimized with a novel EM algorithm. The EM algorithm efficiently identifies a parameterized family of solutions that balance sparsity against squared error.

We propose choosing the penalty weight by evaluating the number of STEs in the solution as the weight decreases from the maximum value. The number of STEs increases relatively slowly as the weight decreases when unrepresented signal components are present, and then increases very rapidly as the weight decreases when the STEs in the solution begin to represent noise. Hence, the weight value at which the number of STEs begins to increase rapidly is a good choice. We present one automated approach for finding this value of rapid increase based on fitting linear and quadratic terms to the number of STEs. Our simulations indicate that this method produces a good lower bound on the weight since it generally selects more STEs than would be identified with the oracle or optimal weight. Erring on the side of a few extraneous STEs is preferable to missing true sources of activity, especially since post processing debiasing generally assigns small amplitudes to extraneous STEs.

The STS approach differs significantly from recent work on ℓ_1 regularized least squares. In most existing ℓ_1 approaches, the regularization term penalizes every

active dipole via $\|\text{vec}\,(\mathbf{X})\,\|_1$ or, if applied in the context of a basis expansion (Eq. (4)), penalizes every active coefficient via $\|\text{vec}\,(\mathbf{\Theta})\,\|_1$. In contrast, STS gives a smaller penalty to active coefficients in the same block than to active coefficients in different blocks. The results of this difference are threefold. First, given our criteria for choosing spatial and temporal bases for each STE, it favors solutions which are connected spatially and vary smoothly over time as opposed to the "spiky", disconnected results often seen in standard ℓ_1 minimization. Second, it allows a wider range of activity within a block for only a small additional penalty. Third, it promotes solutions consisting of a few active spatio-temporal blocks, rather than a few active coefficients or dipoles.

The mesostate model recently proposed by Daunizeau and Friston (2007) also treats cortical activity as a collection of a small number of events, or meso-sources, which are identified through a Variational Bayesian (VB) inversion. The mesostate approach differs in a number of ways from the STS approach. First, the maximum number of meso-sources must be known a priori. In (Daunizeau and Friston, 2007), this number is between one and eight. Second, the VB inversion "switches off" mesostates with little or no evidence. The STS approach, on the other hand, works by "switching on" STEs with strong evidence from a large number of candidate STEs. Third, each meso-source is described by a mean time course and spatial location. Dipoles comprising a meso-source are modeled as perturbations of this mean activity. An STE, on the other hand, describes any measurable activity occurring in a local region of space and time, and thus is not limited to representation of average activity.

Several block penalty methods similar to our STS approach have been proposed (Malioutov et al., 2005; Tropp, 2006; Turlach et al., 2005; Yuan and Lin, 2006), though none of them in the context of M/EEG. All of these aim

to find solutions which are sparse spatially, but not temporally, i.e., are sparse in only one of the dimensions. Also, all but Yuan and Lin (2006) construct blocks containing only a single spatial basis repeated over multiple time samples. This corresponds to defining blocks based on single rows of $\mathbf{X}$. Malioutov et al. (2005) penalize the sum of the ℓ_2 norms of each row of the signal matrix in direction of arrival (DOA) estimation for radar. Second order cone (SOC) programming is used to perform the optimization. Both Turlach et al. (2005) and Tropp (2006) use the sum of ℓ_∞ norms as a penalty, but again consider only single rows of the signal matrix as blocks and do not employ a temporal decomposition. Turlach et al. (2005) solve a constrained version of the STS problem via an interior point method. Tropp (2006) does not specify an optimization algorithm, but instead provides useful analysis of the properties of minimizers. Yuan and Lin (2006) suggest essentially fixed point iteration on the nonlinear functions comprising the subgradient condition for a minimizer of Eq. (8). However, they indicate that computational complexity limits their approach to small problems. None of these earlier works consider the cost function as a Bayesian likelihood, nor does there appear to be any published Expectation-Maximization approach to these "block-sparse" problems.

The choice $p = 2$ in Eq. (7) results in a convex optimization and appears to yield very good results. However, modifying the M-Step in Eq. (16) slightly allows solution of Eq. (7) for other values of p as well. In this case Eq. (13) is written:

$$\hat{\mathbf{\Theta}}^{(k)} = \arg\min_{\mathbf{\Theta}} \left\{ \|\mathbf{\Theta} - \hat{\mathbf{Z}}^{(k)}\|_F^2 + c\lambda \sum_{i,j} \|\mathrm{vec}\left(\mathbf{\Theta}_{i,j}\right)\|_p \right\} \tag{23}$$

Exact solutions to the M-Step represented by Eq. (23) are available for the

cases $p = 0$, $p = 1$, and $p = \infty$ in addition to $p = 2$. The Generalized Expectation-Maximization approach (Gellman et al., 2003) can be used to solve Eq. (23) for other values of p. Note that the case $p = 0$ corresponds to the idealized problem Eq. (6). The M-Step for this case is a hard threshold:

$$\hat{\mathbf{\Theta}}_{i,j}^{(k)} = \begin{cases} \hat{\mathbf{Z}}_{i,j}^{(k)}, & \text{if } \|\hat{\mathbf{Z}}_{i,j}^{(k)}\|_F^2 > c\lambda \\ \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

Although this implementation will find local minima of Eq. (6), there is potentially a unique local minima for every combination of active blocks $\mathbf{\Theta}_{i,j}$, making a global minimum very difficult to find. Thus, in practice, the $p = 2$ relaxation provides better results.

Using a set of spatial basis functions based on cortical patches and wavelet temporal bases, we demonstrated the effectiveness of STS regularization on simulated cortical signals as well as data collected from a self-paced finger tapping experiment. These specific basis functions are not required to use STS. Other methods for constructing the $\mathbf{S}_i$ and $\mathbf{T}_j$ may also prove effective. As one example, short time Fourier bases could be used to isolate activity in different frequency bands over the same time window. Similarly, the spatial bases could be defined for patches associated with anatomically meaningful regions of the brain such as Brodmann areas.

An anatomical approach to patch design is used in the recently proposed MiMS technique (Cottereau et al., 2007). The spatial extent of each patch is chosen based on the local characteristics of the cortical surface. A multipolar expansion of fixed order is used to represent the activity from candidate

patch. MiMs employs a coarse to fine strategy to identify active regions. It begins with large patches, finds a subset of active patches that best fit the data using cross validation, subdivides active patches and repeats the patch selection/subdivision process until a fitting criterion is satisfied. MiMS does not address selection of temporal areas of interest.

Multiresolution spatial events can be incorporated into the STS approach by using patches of varying size. This would allow STS to select the spatial event whose extent best matches the true signal. However, this would result in different numbers of coefficients for each STE. Our presentation and examples assume that each STE involves the same number of coefficients, that is, $\sharp\boldsymbol{\Theta}_{i,j} =$ constant where $\sharp\boldsymbol{\Theta}_{i,j}$ is the product of row and column dimensions for $\boldsymbol{\Theta}_{i,j}$. If different numbers of coefficients are employed, then the penalty term of Eq. (8) should be modified to $\lambda \sum \sqrt{\sharp\boldsymbol{\Theta}_{i,j}} \| \mathrm{vec}\left(\boldsymbol{\Theta}_{i,j}\right) \|_2$ to equalize the expected noise power explained by each STE.

Among the multitude of methods for addressing the M/EEG inverse problem, there does not appear to be a single approach suitable to all applications. Each is tailored to a certain set of assumptions about the underlying brain activity. The assumptions behind STS apply to wide range of experimental paradigms, and STS complements existing techniques. Simulated and actual MEG data suggest STS is effective at localizing cortical activity in both space and time.

5 Acknowledgement

The authors thank the MEG Lab at the University of Texas Health Science Center Houston for the finger tapping data.

A Minimizer of Single Block Problem

We show that $\boldsymbol{\theta}^*$ given by Eq. (15) is the minimizer of $f(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \mathbf{z}\|_2^2 + \frac{\alpha^2\lambda}{\sigma^2}\|\boldsymbol{\theta}\|_2$. We need a few facts from the theory of subgradients (Rockafellar, 1970). A vector $\mathbf{g}$ is a subgradient of a convex function $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ if $f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_0) \geq \mathbf{g}^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ for all θ. The set of all subgradients is called the subdifferential. The subdifferential generalizes the gradient; if there is only one subgradient of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$, it is the gradient. A point $\boldsymbol{\theta}^*$ is the global minimizer of a convex function if and only if $\mathbf{0}$ is a subgradient of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$.

It is easy to see that the subdifferential of $\|\boldsymbol{\theta}\|_2$ contains only the gradient $\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}$ when $\boldsymbol{\theta} \neq \mathbf{0}$, and any vector $\mathbf{g}$ such that $\|\mathbf{g}\|_2 \leq 1$ when $\boldsymbol{\theta} = \mathbf{0}$. Now consider two cases. Case 1: Suppose $\|\mathbf{z}\|_2 > \frac{\alpha^2\lambda}{2\sigma^2}$. Then $\boldsymbol{\theta}^* = (1 - \frac{\alpha^2\lambda}{2\sigma^2\|\mathbf{z}\|_2})\mathbf{z}$ from Eq. (15). Since $\boldsymbol{\theta}^* \neq \mathbf{0}$, the subdifferential of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$ contains only the gradient: $2(\boldsymbol{\theta}^* - \mathbf{z}) + \frac{\alpha^2\lambda}{\sigma^2}\frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|_2} = -\frac{\alpha^2\lambda}{\sigma^2\|\mathbf{z}\|_2}\mathbf{z} + \frac{\alpha^2\lambda}{\sigma^2}\frac{\mathbf{z}}{\|\mathbf{z}\|_2} = \mathbf{0}$. Thus, $\boldsymbol{\theta}^*$ is a minimizer of $f(\boldsymbol{\theta})$ when $\|\mathbf{z}\|_2 > \frac{\alpha^2\lambda}{2\sigma^2}$. Case 2: Suppose $\|\mathbf{z}\|_2 \leq \frac{\alpha^2\lambda}{2\sigma^2}$. Then $\boldsymbol{\theta}^* = \mathbf{0}$ from Eq. (15). Let $\mathbf{g}^* = \frac{2\sigma^2\mathbf{z}}{\alpha^2\lambda}$ and note that $\mathbf{g}^*$ is a subgradient of $\|\boldsymbol{\theta}\|_2$ at $\mathbf{0}$ since $\|\mathbf{g}^*\|_2 \leq 1$. This means that $2(\boldsymbol{\theta}^* - \mathbf{z}) + \frac{\alpha^2\lambda}{\sigma^2}\mathbf{g}^*$ is a subgradient of $f(\boldsymbol{\theta})$ at $\mathbf{0}$. To complete the proof, note that $2(\boldsymbol{\theta}^* - \mathbf{z}) + \frac{\alpha^2\lambda}{\sigma^2}\mathbf{g}^* = -2\mathbf{z} + 2\mathbf{z} = \mathbf{0}$.

B Derivation of λ_{MAX}

We show that $\hat{\boldsymbol{\Theta}}(\lambda) = \mathbf{0}$ for $\lambda \geq \lambda_{MAX}$ given in Eq. (17). We again use the theory of subgradients (Rockafellar, 1970). We first rewrite the objective function of Eq. (8) as $f(\boldsymbol{\theta}) = \|\mathbf{y} - \sum_{i,j}(\mathbf{T}_j \otimes \mathbf{HS}_i)\boldsymbol{\theta}_{i,j}\|_2^2 + \lambda\sum_{i,j}\|\boldsymbol{\theta}_{i,j}\|_2$, where $\boldsymbol{\theta}$ is a vector concatenation of all the $\boldsymbol{\theta}_{i,j}$, $\boldsymbol{\theta} = [\,\boldsymbol{\theta}_{1,1}^T\,\boldsymbol{\theta}_{1,2}^T \cdots \boldsymbol{\theta}_{2,1}^T \cdots \boldsymbol{\theta}_{I,J}^T\,]^T$. The subdifferential of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \mathbf{0}$ contains only the vectors whose $(i,j)^{th}$

component is $-2(\mathbf{T}_j \otimes \mathbf{HS}_i)^T(\mathbf{y} - \mathbf{0}) + \lambda \mathbf{g}_{i,j}$ where $\|\mathbf{g}_{i,j}\|_2 \leq 1$ for all i, j. Let $\mathbf{g}_{i,j} = \frac{2}{\lambda}(\mathbf{T}_j \otimes \mathbf{HS}_i)^T\mathbf{y}$ and suppose $\lambda \geq 2\max_{i,j}(\mathbf{T}_j \otimes \mathbf{HS}_i)^T\mathbf{y}$. Then $\|\mathbf{g}_{i,j}\|_2 \leq 1$ for all i, j and the subdifferential of $f(\boldsymbol{\theta})$ at $\mathbf{0}$ contains the all zero vector, implying that $\boldsymbol{\theta} = \mathbf{0}$ minimizes $f(\boldsymbol{\theta})$.

References

Auranen, T., Nummenmaa, A., Hämäläinen, M., Jääskeläinen, I., Lampinen, J., Vehtari, A., Sams, M., 2005. Bayesian analysis of the neuromagnetic inverse problem with ℓ^p-norm priors. NeuroImage 26 (3), 870–884.

Baillet, S., Mosher, J., Leahy, R., 2001. Electromagnetic brain mapping. IEEE Signal Processing Magazine 18 (6), 14–30.

Bear, M., Connor, B., Paradiso, M., 2001. Neuroscience: Exploring the Brain, 2^{nd} Ed. Lippincott, Williams and Wilkins, Philadelphia, PA.

Bolstad, A., Van Veen, B., Nowak, R., August 2007. Magneto-/electroencephalography with space-time sparse priors. In: IEEE Statistical Signal Processing Workshop. Madison, WI, pp. 190–194.

Brewer, J., 1978. Kronecker products and matrix calculus in system theory. IEEE Transactions on Circuits and Systems cas-25 (9), 772–781.

Candes, E. J., Tao, T., 2005. Decoding by linear programming. IEEE Transactions on Information Theory 51 (12), 4203–4215.

Cottereau, B., Jerbi, K., Baillet, S., 2007. Multiresolution imaging of meg cortical sources using an explicit piecewise model. NeuroImage 38 (3), 439–451.

Culham, J. C., Cavina-Pratesi, C., Singhal, A., 2006. The role of parietal cortex in visuomotor control: What have we learned from neuroimaging? Neuropsychologia 44 (13), 2668–2684.

Daunizeau, J., Friston, K., 2007. A mesostate-space model for eeg and meg. NeuroImage 38 (1), 67–81.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the em algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

Dogandžić, A., Nehorai, A., 2000. Estimating evoked dipole responses in unknown spatially correlated noise with EEG/MEG arrays. IEEE Trans Signal Processing 48 (1), 13–25.

Donoho, D., 1995. De-noising by soft-thresholding. IEEE Transactions on Information Theory 41 (3), 613–627.

Donoho, D., 2006. For most large underdetermined systems of linear equations, the minimal ell-1 norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59 (6), 797–829.

Figueiredo, M., Nowak, R., 2003. An em algorithm for wavelet-based image restoration. IEEE Transactions on Image Processing 12 (8), 906–916.

Gellman, A., Carlin, J., Stern, H., Rubin, D., 2003. Bayesian Data Analysis, 2^{nd} Ed. CRC Press, Boca Raton, FL.

Gorodnitsky, I., Rao, B., 1997. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. IEEE Transactions on Signal Processing 45 (3).

Huang, M., Dale, A., Song, T., Halgren, E., Harrington, D., Podgorny, I., Canive, J., Lewis, S., Lee, R., 2006. Vector-based spatial-temporal minimum l1-norm solution for meg. NeuroImage 31, 1025–1037.

Jeffs, B., Leahy, R., Singh, M., 1987. An evaluation of methods for neuromagnetic image reconstruction. IEEE Transactions on Biomedical Engineering 34, 713–723.

Limpiti, T., Van Veen, B., Wakai, R., 2006. Cortical patch basis model for

spatially extended neural activity. IEEE Transactions on Biomedical Engineering 53 (9), 1740–1754.

Malioutov, D., Çetin, M., Wilsky, A., 2005. A sparse signal reconstruction perspective for source localization with sensor arrays. IEEE Transactions on Signal Processing 53 (8), 3010–3022.

Matsuura, K., Okabe, Y., 1995. Selective minimum-norm solutions of the biomagnetic inverse problem. IEEE Transactions on Biomedical Engineering 42 (6), 608–615.

Matsuura, K., Okabe, Y., 1997. A robust reconstruction of sparse biomagnetic sources. IEEE Transactions on Biomedical Engineering 44 (8), 720–726.

Mosher, J., Leahy, R., Shattuck, D., Baillet, S., 1999. MEG source imaging using multipolar expansions. In: Lecture Notes in Computer Science. Vol. 1613. Proc. IPMI99, Springer-Verlag GmbH.

Mosher, J., Lewis, P., Leahy, R., 1992. Multiple dipole modeling and localization from spatio-temporal MEG data. IEEE Trans. Biomed. Eng. 39 (6), 541–557.

Nolte, G., Curio, G., 2000. Current multipole expansion to estimate lateral extent of neuronal activity: A theoretical analysis. IEEE Trans Biomed. Eng. 47 (10), 1347–1355.

Pascual-Marqui, R. D., Michel, C. M., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. International Journal of Psychophysiology 18, 49–65.

Rockafellar, R. T., 1970. Convex Analysis. Princeton University Press, Princeton, New Jersey.

Scherg, M., von Cramon, D., 1985. Two bilateral sources of the late AEP as identified by a spatio-temporal dipole model. Electroencephalogr. Clin. Neurophysiol. 62, 32–44.

Tropp, J., 2006. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. Signal Processing 86, 589–602.

Turlach, B., Venables, W., Wright, S., 2005. Simultaneous variable selection. Technometrics 47 (3), 349–363.

Uutela, K., Hämäläinen, M., Somersalo, E., 1999. Visualization of magnetoencephalographic data using minimum current estimates. NeuroImage 10, 173–180.

Van Veen, B., Van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44 (9), 867–880.

Wagner, M., Fuchs, M., Kastner, J., Aug 2002. Current density reconstructions and deviation scans using extended sources. In: Proceedings of the 13th International Conference on Biomagnetism. Jena, Germany.

Wagner, M., Köhler, T., Fuchs, M., Kastner, J., Aug 2000. An extended source model for current density reconstructions. In: Nenonen, J., Ilmoniemi, R., Katila, T. (Eds.), Proceedings of the 12th International Conference on Biomagnetism. Espoo: Helsinki Univ. of Technology, Espoo, Finland, pp. 749–752.

Wu, C., 1983. On the convergence properties of the em algorithm. Annals of Statistics 11, 95–103.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society B 68, 49–67.