A spatio-temporal solution for the EEG/MEG inverse problem using group penalization methods

Tian Siva Tian* and Zhimin Li

The inverse problem encountered in electroencephalography (EEG) and magnetoencephalography (MEG) studies refers to estimating neural activity given limited scalp-recorded data. We propose a spatio-temporal solution using group penalization approaches. This proposed method is based on the assumption that the underlying sources of EEG/MEG measurements are smooth in the temporal domain, and focal in the spatial domain. It transforms the spatio-temporal problem to a high-dimensional linear regression problem with grouped predictors using a basis expansion. Then an iterative group elastic net algorithm is utilized to localize and estimate the source time courses. The proposed approach is shown to be effective on simulations and human MEG studies.

Keywords and phrases: Inverse problem, Spatio-temporal data, Group elastic net, EEG/MEG.

1. INTRODUCTION

Over the past few decades, scalp electroencephalography (EEG) and magnetoencephalography (MEG) have been widely used in detecting spontaneous or evoked-response brain activities in normal or clinical populations due to their non-invasive nature. When a subject is receiving a stimulus, neurons in the functional region of the brain, such as the somatosensory cortex or visual cortex, are activated and then generate small currents. These small current flows produce a small magnetic and a small electric field that are mutually orthogonal. EEG and MEG measure the voltage and the magnetic fields, respectively, using a set of sensors at various locations on or near the scalp of the subject.

One of the most important uses for EEG/MEG is to determine the locations of these functional regions and how they respond to stimuli. Other important uses include detecting and localizing epileptiform activities, distinguishing different neural activity patterns between different groups of patients, determining the malfunctioning brain regions for patients with mental diseases and so on. The accomplishment of these goals requires creating a sequence of accurate brain activity maps over time. That is, one wishes to find out the source to the electric/magnetic fields. This involves deriving the electric current density (Sarvas, 1987), and this problem is called the “inverse problem”.

The EEG/MEG inverse problem has no unique solution making it by definition “ill-posed” (Baillet et al., 2001). This is because there is an infinite number of electric current distributions which can lead to the same EEG/MEG measurements (Nunez, 1981; Sarvas, 1987). In addition, the current EEG/MEG devices have only a few hundred sensors, but the number of sources is much larger because each neuron can be treated as a potential source. The number of sensors is generally too insufficient to provide a precise reconstruction of the source current (Grave de Peralta Menendez and Gonzalez Andino, 1998).

In order to make the problem manageable and improve computational efficiency, many methods discretize the entire distribution, and generate a fine grid that covers the region of interest in the brain. Then mutually orthogonal electric dipoles with unknown amplitudes are attached at each grid point. Existing methods designed for solving this problem can be mainly grouped into three categories: the dipole fitting methods, the scanning methods and the imaging methods. Dipole fitting and scanning methods assume that the sources reside in only a few activated grid points, and then they attempt to decide among the possible dipole locations where the most appropriate location is for the sources so as to best describe the measured time courses. One challenge of these methods is that it is usually difficult to choose the number of sources (Bolstad et al., 2009). Examples of the dipole fitting methods include the equivalent current dipole methods (Scherg and von Cramon, 1986; Hämäläinen et al., 1993; Yamazaki et al., 2000) and some Bayesian methods (Jun et al., 2005; Sorrentino et al., 2009). Examples of the scanning methods include the multiple signal classification (MUSIC) method (Mosher et al., 1992), the beamforming method (VanVeen et al., 1997), maximum likelihood estimation (Dogandžić and Nehorai, 2000), etc.

The imaging methods are also referred to as the non-parametric methods. These methods attempt to estimate the dipole sources all together based on the assumption that the primary sources can be represented as linear combinations of neuron activities (Barlow, 1994). That is, the inverse problem can be expressed using the following nonparametric linear model:

\[ \mathbf{y}(t) = \mathbf{X}\mathbf{\beta}(t) + \mathbf{e}(t), \]
where \( y(t) = [y_1(t), y_2(t), \ldots, y_n(t)]^T \) is a set of EEG/MEG time courses measured by \( n \) sensors. The forward operator \( X \) is an \( n \times p \) matrix representing the propagation of the voltage/magnetic field. It contains information about positions and orientations of dipoles, and how they are represented at the sensor level. Most commonly, \( X \) is computed by a simple spherical head model (Mosher et al., 1999), so it is a “known” design matrix in Equation (1). The \( \beta(t) = [\beta_1(t), \beta_2(t), \ldots, \beta_p(t)]^T \) is a set of time courses representing dipole activity. As mentioned earlier, the cortical area is divided into a fine grid of small cells, say \( p/3 \) cells. Each cell is assigned a current dipole with unknown amplitudes. The dipolar moment orientation is decomposed to \( x, y, \) and \( z \) directions in a 3-D coordinate system. Therefore, the total dimension of \( X \) is \( p \), and each dimension (column) represents the lead field for one dipole at one direction. The \( e(t) = [e_1(t), e_2(t), \ldots, e_n(t)]^T \), is a set of additive noise time courses. The goal is to estimate \( \beta(t) \) given \( y(t) \) and \( X \).

Typically \( p \) can be several thousands, whereas \( n \) is at most a few hundred, i.e., \( p \gg n \). Furthermore, \( \beta(t) \) is assumed to have two characteristics: spatial focality and temporal smoothness. Spatial focality means that there should be only a small number of compact regions in the entire cortex staying active at one time point. That is, among \( p \) \( \beta_j(t) \)'s only a small portion related to the stimuli has significant amplitudes, while the rest associated with inactive regions should be normal static brain waves or noise perturbations. Temporal smoothness means that in the source regions the \( \beta_j(t) \)'s should be relatively smooth over time even though they can change rapidly.

Regularization is a commonly used technique for the problem at hand, and the solution can be expressed as:

\[
\arg \min_{\beta} \{ \| y(t) - X\beta(t) \|^2 + \lambda \text{PEN}(\beta(t)) \},
\]

where \( \text{PEN}(\cdot) \) is a penalty function and \( \lambda \geq 0 \) is a fixed penalty parameter that controls the degree of regularization.

Earlier methods solve Equation (2) separately at each time point. That is, suppose there are \( s \) measured time points. To solve Equation (2), one has to solve \( s \) optimization problems individually. One of the most well-known methods is the minimum norm estimate (MNE) (Hämäläinen and Ilmoniemi, 1994), which imposes the \( L_2 \)-penalty on \( \beta \), i.e., \( \text{PEN}(\beta) = \| \beta \|^2 \). Where \( \beta \) is \( \beta(t) \) in the matrix form. Some modifications include the weighted minimum norm estimate (WMNE) (Dale and Sereno, 1993; Iwaki and Ueno, 1998) and the low resolution electrical tomography (LORETA) (Pascual-Marqui et al., 1994). Both methods utilize a weighted \( L_2 \)-penalty, \( \text{PEN}(\beta) = \| W\beta \|^2 \), instead of the regular \( L_2 \)-penalty, where \( W \) is a \( p \)-by-\( p \) weighting matrix. One advantage of \( L_2 \) methods is that they can be very efficient due to the computational efficiency of the \( L_2 \)-norm minimization. However, applying \( L_2 \)-norm to the entire \( \beta \) space produces a well-know blurring effect on the spatial domain of the reconstructed sources.

One alternative is the \( L_1 \)-penalty, i.e., \( \text{PEN}(\beta) = \| \beta \|_1 \). [e.g., the minimum current estimate (MCE) (Matsuura and Okabe, 1995; Uutela et al., 1999; Lin et al., 2006)]. The FOCal Underdetermined System Solver (FOCUESS) (Gorodnitsky and Rao, 1997) is a recursive procedure of WMNE with an adjustable weighting matrix. Rao and Kreutz-Delgado (1999) has shown that FOCUSS is equivalent to an \( L_1 \)-norm solver with \( 0 < l \leq 1 \). However, it is claimed that FOCUSS is very sensitive to noise (Ou et al., 2009). In general, \( L_1 \) methods can produce better solutions in terms of identifying sources due to the nature of the \( L_1 \)-penalty. However, they also introduce substantial discontinuities to the temporal domain. Hence, the “spiky-looking” time courses will be observed.

A great deal of research has been devoted to prevent the emergence of spiky \( L_1 \) solutions and diffuse \( L_2 \) solutions. Remedies include many hierarchical Bayesian approaches [see, for example, (Baillet and Garnero, 1997; Sato et al., 2004; Nummenmaa et al., 2007)]. Some of them considered the spatial focality and temporal smoothness by employing sparse priors and/or smoothing priors (Friston et al., 2008; Daunizeau et al., 2006; Nummenmaa et al., 2007). Various \( L_h \)-norm methods with \( 0 < h < 1 \) and \( 1 < h < 2 \) were also discussed (Auranen et al., 2005; Jeffs et al., 1987). Ou et al. (2009) proposed an \( L_1L_2 \)-regularization, which imposed the \( L_1 \)-penalty to the spatial domain and \( L_2 \)-penalty to the temporal domain. Bolstad et al. (2009) proposed an event sparse penalty procedure, which represented \( \beta(t) \) using both spatial and temporal bases. These methods can solve the “spiky-looking” problem in the temporal domain. However, they may produce scattering estimated sources, which may deliver implausible results physiologically (Hauke et al., 2011). That is, sources are tightly grouped together and do not smoothly transit from one cell to another.

In this paper, we propose a Spatio-Temporal Penalization (STP) approach that takes into account both smoothness and focality assumptions in the temporal and spatial domains, respectively. The proposed method can potentially produce a smooth transition in the spatial domain, and hence, reduce the scattering effect. The temporal smoothness is guaranteed by the basis expansion of the source time courses. Then the temporal bases are combined with the forward operator \( X \) containing the spatial information. This combined new forward operator then contains both spatial and temporal information of the sources. The spatio-temporal problem is then transformed to a high-dimensional linear regression problem with grouped predictors. An \( L_1 \) and an \( L_2 \) penalties are placed on the grouped coefficients through an iterative group elastic net (IGEN) approach. The employment of the elastic net penalties ensures a moderate degree of the spatial focality. The application of basis representation guarantees the temporal smoothness, and the elastic net penalties ensure a moderate degree of the focality of the spatial domain.
The rest of the paper is organized as follows. We present the general scheme for the STP approach and its implementation details in Sections 2 and 3. We describe two possible basis functions for STP in Section 4. We demonstrate the effectiveness of STP on two simulated examples and a human MEG example in Sections 5 and 6. Some tuning parameters are also discussed in Section 5. Section 7 discusses potential issues for future investigation.

2. METHODOLOGY

Consider Equation (1) and we assume that \( e(t) \) is independent white noise over time. Time-dependent noise models are considered in Huizenga et al. (2002); Bijma et al. (2005), but we will leave it for future investigation. The \( \beta(t) \) can be projected onto a subspace that is spanned by its temporal orthogonal basis functions. Assuming that most energy lies in the first \( q \) basis functions, we write Equation (1) as follows:

\[
\begin{align*}
(3) \quad y_i(t) &= \sum_{j=1}^{p} x_{ij} \beta_j(t) + e_i(t) \\
&= \sum_{j=1}^{p} x_{ij} b(t)^T \eta_j + e_i(t), \quad i = 1, \ldots, n,
\end{align*}
\]

where \( b(t) = [b_1(t), b_2(t), \ldots, b_q(t)]^T \) is the \( q \)-dimensional basis of \( \beta_j(t) \) and \( \eta_j \) is its corresponding coefficient vector. Let \( x_{ijt} = x_{ij} b(t) \). Note that \( X \) contains the spatial information of sources, so \( x_{ijt} \) in turn contains both spatial and temporal information. Then Equation (3) becomes:

\[
(4) \quad y_i(t) = \sum_{j=1}^{p} x_{ijt}^T \eta_j + e_i(t) = X_{it} \eta + e_i(t), \quad i = 1, \ldots, n,
\]

where \( X_{it} = [x_{11t}^T, \ldots, x_{npt}^T] \in \mathbb{R}^{pq} \) and \( \eta = [\eta_1^T, \eta_2^T, \ldots, \eta_p^T] \in \mathbb{R}^p \). We express Equation (4) in the matrix form:

\[
(5) \quad y^* = X^* \eta + e^*,
\]

where \( y^* = \text{vec}[y(t)] = [y_1^T, \ldots, y_n^T]^T \) is an \( ns \)-vector of the new response variable. Here \( y_i, \quad i = 1, \ldots, n, \) is the \( i \)th EEG/MEG time course measured over \( s \) time points. The \( e^* = \text{vec}[e(t)] = [e_1^T, e_2^T, \ldots, e_n^T]^T \) is an \( ns \)-vector of the new white noise variable, where \( e_i, \quad i = 1, \ldots, n, \) is the \( i \)th discretized noise time course. We have \( X^* = X \otimes B \), i.e.,

\[
X^* = \begin{bmatrix} x_{11} B & \cdots & x_{1p} B \\ \vdots & \ddots & \vdots \\ x_{n1} B & \cdots & x_{np} B \end{bmatrix} = \begin{bmatrix} x_{11}^T B & \cdots & x_{1p}^T B \\ \vdots & \ddots & \vdots \\ x_{n1}^T B & \cdots & x_{np}^T B \end{bmatrix}.
\]

where \( B \in \mathbb{R}^{s \times q} \) is \( \beta(t) \)'s basis functions in the matrix form, i.e., the \((k, r)\)th element of \( B \) is \( b_r(t_k) \), and more explicitly, each block element, \( x_{ijt}^* \) (e.g., \( x_{11t}^* = x_{11} b_1(t_1), \ldots, x_{11} b_q(t_1) \)), is a \( q \)-vector. This way, the original problem has been transformed to a high-dimensional linear regression problem, where the new design matrix \( X^* \) is \( ns \)-by-\( pq \). This \( X^* \) consists of \( p \) groups, and each group contains \( q \) group members. The temporal information for \( \beta_j(t) \) resides within each group, and the original spatial information still remain between groups. Hence, \( X^* \) contains both spatial and temporal information.

We write \( X^* = [x_1^*, x_2^*, \ldots, x_p^*] \), where each column vector is a block matrix, i.e., \( x_i^* = [x_{i1}B^T, x_{i2}B^T, \ldots, x_{iq}B^T]^T \). Since \( x_i^* \) contains temporal information of the \( j \)th potential source, it is desired to treat predictors in each \( x_i^* \) as a group in the estimation procedure. We are only interested in estimating the functional forms of active \( \beta_j(t) \)'s. Hence, those inactive \( \beta_j(t) \)'s can be set to zero. That is, elements in a coefficient vector, \( \eta_j, \quad j = 1, \ldots, p \), should be either all zero or mostly nonzero. Put another way, columns in \( x_i^* \) should be either included or excluded as a group in the model. Therefore, some group penalization methods may be considered.

One solution is to use a group Lasso technique, which can shrink inactive groups to be exact zero. However, sources in active areas are highly correlated to each other, the group Lasso may tend to select only one of them and does not care which one is selected (Zou and Hastie, 2005). This is equivalent to impose an \( L_1 \)-penalty to the spatial domain, and hence, may result in the well-known “scattering” sources. Instead, we propose to use the following penalized regression model:

\[
(6) \quad \hat{\eta} = \arg \min_{\eta} \left\{ \|y^* - X^* \eta\|_2^2 + \lambda_1 \sum_{j=1}^{p} \|\eta_j\|_1 + \lambda_2 \sum_{j=1}^{p} \|\eta_j\|_2^2 \right\},
\]

where \( \lambda_1 \) and \( \lambda_2 \) are penalty parameters controlling the degrees of regularization.

Equation (6) is a group version of the elastic net problem (Zou and Hastie, 2005). It can be written as solving a group Lasso problem (Yuan and Lin, 2006):

\[
(7) \quad \hat{\eta} = \arg \min_{\eta} \left\{ \|y^* - X^* \eta\|_2^2 + \gamma \sum_{j=1}^{p} \|\eta_j\|_1 \right\},
\]

where \( \gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}} \).

The choice of \( \lambda_1 \) and \( \lambda_2 \) can be based on a cross-validation (CV) on a 2-dimensional space (Zou and Hastie, 2005). For example, we first specify a set of candidate values to \( \lambda_2 \), e.g., \( \lambda_2 \in (0, 0.01, 0.1, 1, 10) \). For each value of \( \lambda_2 \), run Equation (7) and obtain an entire path of \( \lambda_1 \). Then one can select \( \lambda_1 \) by a \( K \)-fold CV and choose the \( \lambda_2 \) that gives the smallest CV error.

*Spatio-temporal inverse solver* 523
3. ITERATIVE GROUP ELASTIC NET

There is a potential practical problem with the implementation of the group elastic net. After basis expansion, $X^*$ may become ultra-large. For example, we have $n = 247$, $p = 15,372$, $s = 228$, and $q = 7$ in our MEG example in Section 6. Then $X^*$ is a 56,316-by-107,604 matrix and $X$ is 163,920-by-107,604. Equation (7) is hard to compute and requires more memory than from any of the constituent models (Opitz and Maclin, 1999), and can be more robust to noise.

Therefore, we propose an ad-hoc iterative group elastic net (IGEN) algorithm that is more satisfactory in our settings. The idea is to resample subareas from the entire cortex and apply a group elastic net to each sample to identify active sources in that sample. Once many samples are examined, a candidate pool of possible active sources can be formed by simply putting together all historically selected sources, applying a majority vote to this candidate pool, and selecting a small portion of the sources in the pool. This algorithm is similar to an ensemble method, which uses multiple models to obtain better predictive performance than from any of the constituent models (Opitz and Maclin, 1999), and can be more robust to noise.

The following steps describe the STP approach. First, IGEN randomly selects a subset of variables from $X$ (i.e., columns of $X$) with equal probabilities. Intuitively, this step is to divide the cortical area into subareas, which can overlap with each other. Other sampling techniques may also be considered, such as even-spaced sampling, block sampling (i.e., dividing the cortex into small blocks and sampling one or a few blocks each time), and weighted sampling (i.e., assigning higher probabilities to regions of interest), but will be left to our future investigation. Second, we expand the selected subset using the method described in Section 2. Third, treating temporal coefficients for one $\beta_j(t)$ as a group, a group elastic net (GEN) is applied to identify a small number of potential active groups. Repeat these three steps a sufficiently large number of times. Ideally, one wishes that the number of iterations could be large enough to guarantee all columns in $X$ are examined multiple times. In practice, some moderately large number can be chosen to obtain a good result. We explain the choice of this number in Section 5.3. Fourth, a candidate pool is formed by putting together selected groups from all previous iterations, and a majority vote is applied to this candidate pool to further narrow down the “important groups”. That is, count the number of times a group is selected, put these counted numbers for all groups in descending order, and choose the first several groups with the largest number of times being selected in previous iterations. These groups are considered as more “important” or “active”. The idea of using a majority voting rule is based on the assumption that some real active dipoles can always be identified even in different sets of dipole samples, but some less active ones may be selected only when the more active ones are not in the sample. A majority vote can prevent selecting inactive dipoles. Then we apply GEN again to the groups resulting from the majority vote and obtain the estimated coefficient vectors. Finally, we transform everything back to the original spatio-temporal space to obtain the estimated source matrix, $\hat{\beta}$.

Denote the maximum sub-sample size for one subset of $X$ as $p_{\text{max}}$. That is, each subset can only contain less than or equal to $p_{\text{max}}$ variables of $X$. The ad-hoc IGEN algorithm can be described as follows:

1. Initialize the number of iterations $L$ and the maximum sub-sample size $p_{\text{max}}$.
2. For $l = 1$ to $L$
   (a) Draw a random sample of size $p'$ from $X$, where $p' \leq p_{\text{max}}$.
   (b) Expand the subset of data by Equation (4) or (5).
   (c) Fit GEN (6) to the expanded data from Step 2(b) and select $m^{(l)}$ groups.
3. Apply a majority vote to the $L$ sets of selected groups and obtain the $m$ most voted groups.
4. Apply GEN to these $m$ most voted groups from Step 3 to select $m_0$ groups and obtain their estimated coefficient vector, $\hat{\eta}_j$’s, $j = 1, \ldots, m_0$.
5. Calculate the estimated source time courses, $\hat{\beta} = B\hat{\eta}^T$, where $\hat{\eta}^T = [\hat{\eta}_1^T, \ldots, \hat{\eta}_{m_0}^T]^T$.

There are several tuning parameters in this algorithm. First, the ideal number of iterations $L$ should be sufficiently large so that all columns of $X$ can be examined several times with different sample combinations, especially when $p_{\text{max}}$ is relatively small. This requires $Lp_{\text{max}} \gg p$. However, we demonstrate that in practice the predictive accuracy is not affected much as long as $L$ is roughly more than twice as large as $p/p_{\text{max}}$ in Section 5.3. Second, the maximum size of a sub-sample, $p_{\text{max}}$, can depend on the computational resource. In Section 5.3 we empirically show that the prediction accuracy is not very sensitive to the choice of $p_{\text{max}}$ either, but if $p_{\text{max}}$ is too large the computational cost increases. We also illustrate the relationship between $L$ and $p_{\text{max}}$ relative to the predictive accuracy. Third, the number of groups selected by the majority voting rule can be as large as the size of the union of all selected groups, i.e., $\bigcup_{l=1}^L m^{(l)}$. But in our study, we use $m = 10\%p$ based on the assumption that only a small number of locations in the cortex are responsible for the measured signals. Fourth, $m^{(l)}$ and $m_0$ need to be justified. In Step 2(c) $m^{(l)}$ varies from iteration to iteration and it is automatically chosen by GEN based on minimizing the 10-fold cross-validated error rate on the training data. In Step 4 $m_0$ is also decided in a similar manner. Presumably, $m_0$ can be as large as $m$.

The next problem is that whether or not IGEN performs as well as the standard GEN, if not better. We compare the performance of IGEN to GEN using a simulated small data set in Section 5.
4. BASIS SELECTION

STP requires an approximation of the signal space by its temporal basis. One problem is to choose an appropriate basis. Here we examine one data-dependent basis and one data-independent basis.

The data-dependent basis is the right orthogonal matrix of the singular-value decomposition (SVD) of the EEG/MEG data. Let $Y$ be the measured EEG/MEG data in the matrix form. That is, each row of $Y$ represents a time course for all time instants. Then the SVD of $Y$ can be expressed as

$$Y = UDV^T,$$

where the orthogonal matrix $U$ can be viewed as the spatial pattern of the EEG/MEG measurements, while the orthogonal matrix $V$ can be viewed as the temporal pattern. The diagonal matrix $D$ contains the singular values in a descending order. Since $y(t)$ represents the measured time courses evoked by the signal sources, it is natural to assume that $\beta(t)$ and $y(t)$ share the same temporal bases. In particular, we approximate $\beta_j(t)$’s using the first $q$ basis from $V$, which corresponds to the first $q$ largest singular values. These $q$ bases are considered to have most of the information in active signals.

There are many options for the data-independent basis. Some commonly used ones include wavelet and Fourier bases. We demonstrate the natural cubic splines (NCS) in this paper.

One issue is to determine the value of $q$. We use a fixed number of $q = 9$ for both bases in our simulations and MEG study. Our variance analysis shows that $q = 9$ is adequate in both simulation and MEG studies in this paper, because the first nine eigenvectors contain about 80% variance of $Y$. Note that the specification of $q$ can be different for different subareas in practice, since $q$ controls the flexibility of the functions. Presumably a larger $q$ may be assigned to subareas with higher source activity. Therefore, a two-step approach may be considered. First, assign a fixed $q$ to the entire cortex, i.e., all $p \beta_j(t)$’s, and then apply STP to obtain a rough estimate of the active subareas. Second, assign a larger $q$ to the active subareas and smaller $q$ to inactive subareas, and conduct STP again to obtain a more accurate estimation. In particular, when prior knowledge is available, one can be more confident to specify a larger $q$ to the potential active regions. Intuitively, $q$ can be larger in studies with high frequency stimuli than that in studies with low frequency stimuli. To reduce computational cost, we suggest that the reassigned $q$ for the active subareas only needs to be slightly larger than the initial $q$ and the reassigned $q$ for the inactive subareas can be as low as 3.

5. SIMULATION STUDIES

In this section we demonstrate the relative performance of STP with the aforementioned two bases to two commonly used EEG/MEG inverse solvers using two simulated examples. In both simulations, we create two scenarios with different SNR levels. One is a lower noise level scenario (Scenario 1) and the other is a higher noise level scenario (Scenario 2). The two counterparts are the $L_2$-norm based MNE method (Dale and Sereno, 1993) and the $L_1$-norm based MCE method (Lin et al., 2006). In order to demonstrate the effectiveness of IGEN, we also implement the standard GEN algorithm to STP (STP-GEN) in Simulation 1.

Note that in both the synthetic and the MEG examples (i.e., examples in this section and Section 6), we use $\beta_j(t)$ and a subscript to represent the estimated $\beta_j(t)$ by different methods. For example, $\beta_{j,NCS}(t)$ represents the estimated $\beta_j(t)$ using STP with a NCS basis and $\beta_{j,SVD}(t)$ represents the estimated $\beta_j(t)$ using STP with the basis from SVD of $Y$. Similarly, $\beta_{j,MNE}(t)$ and $\beta_{j,MCE}(t)$ represent the estimated $\beta_j(t)$’s using MNE and MCE, respectively.

5.1 Simulation I

In this section we simulate a relatively small data set. We let $n = 60$, $s = 50$, $p = 100$. That is the forward operator $X$ is a 50-by-100 matrix, and each element is generated from a standard normal distribution. Among the 100 functions, the first 10 are signals and have exactly the same form. They are generated from an 11-dimensional NCS basis. Elements in the coefficient vector are generated from $N(0, 1)$. The remaining 90 functions are noise and are generated randomly from $N(0, 0.01)$. We create two scenarios: a low noise level scenario (Scenario 1) and a high noise level scenario (Scenario 2). In Scenarios 1 and 2, elements from the white noise, $c_i(t)$, are generated independently from $N(0, 0.1)$ and $N(0, 1)$, respectively. The responses, $y_C(t)$’s, are obtained by Equation (5), and they are plotted in Figure 1 (a) and (c), respectively.

Using an NCS basis to approximate $\beta_j(t)$’s matches the data generation mechanism. Therefore, STP with NCS basis should outperform STP with SVD basis. Figure 1 (b) and (d) show the true and estimated signal time courses by STP, MNE and MCE in the two scenarios. $\beta_{j,NCS}(t)$ and $\beta_{j,SVD}(t)$ estimated $\beta_j(t)$ considerably well in Scenario 1. The shapes of both $\beta_{j,NCS}(t)$’s and $\beta_{j,SVD}(t)$’s followed the shape of the signal $\beta_j(t)$. In particular, STP with NCS basis produced smoother curves. MCE also produced good estimations of $\beta_j(t)$’s. But MNE tended to underestimate the signals. The shape of $\beta_{j,MNE}(t)$ followed the signal $\beta_j(t)$ but the magnitude was much smaller. STP performed considerably well in Scenario 2 too, while MNE and MCE were unsatisfactory due to the underestimation of the signal $\beta_j(t)$.

For evaluation purposes, we use two measures to measure the performance of the inverse solvers:

Spatio-temporal inverse solver 525
Figure 1. The measured signal, $y_i(t)$’s, from hypothetical 60 sensors, and the true, $\beta(t)$, and estimated source time courses, $\hat{\beta}_j(t)$’s, by different methods in Simulation I.

- The percentage of correctly identified $\beta_j(t)$’s. That is, the percentage of the correctly identified active sources plus the percentage of the correctly identified inactive sources. We call this measure the percentage of correctly identified signals, denoted by PCIS.
- The overall mean squared error (MSE) between the true
Table 1. Comparing methods based on MSE and PCIS in Simulation I

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Criterion</th>
<th>STP (NCS)</th>
<th>STP (SVD)</th>
<th>MNE</th>
<th>MCE</th>
<th>STP-GEN (NCS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low noise</td>
<td>MSE ($10^{-3}$)</td>
<td>51.6</td>
<td>56.1</td>
<td>178.8</td>
<td>45.6</td>
<td>54.9</td>
</tr>
<tr>
<td></td>
<td>PCIS (%)</td>
<td>94.5</td>
<td>93.7</td>
<td>33.2</td>
<td>94.3</td>
<td>92.6</td>
</tr>
<tr>
<td>High noise</td>
<td>MSE ($10^{-3}$)</td>
<td>58.2</td>
<td>60.9</td>
<td>196.6</td>
<td>82.9</td>
<td>67.7</td>
</tr>
<tr>
<td></td>
<td>PCIS (%)</td>
<td>91.2</td>
<td>88.9</td>
<td>19.8</td>
<td>67.5</td>
<td>84.1</td>
</tr>
</tbody>
</table>

$\text{Figure 2. The true source time courses at two different locations and the corresponding brain maps in Simulation II.}$

and the estimated solution, which is calculated by

$$MSE = E\left(\frac{1}{p}||\beta - \hat{\beta}||^2_F\right),$$

where $\|\cdot\|_F^2$ is the Frobenius norm.

Since the data set in this simulation is fairly small, computation may not be a serious issue for the GEN method. Therefore, we implement the GEN algorithm within our STP framework (STP-GEN). Here we only examine the NCS basis because it has been considered superior to the SVD basis in this particular example. Table 1 lists MSE and PCIS for the three methods. In Scenario 1, MCE performed the best due to the smallest MSE and its high PCIS. STP with NCS is the second best method. Its MSE was small and its PCIS was as high as MCE if not higher. MNE was the least satisfactory method because its MSE was large and it only correctly identified 33% of the active/inactive sources. STP dominated over other methods in Scenario 2, where the noise level was high. The accuracy of MCE dropped in this scenario. It only correctly identified about 67.5% active/inactive sources, while its MSE increased approximately 90% compared to that in Scenario 1. STP still produced accurate, stable estimations. The MSE and PCIS did not change much. STP with SVD performed slightly inferior to STP with NCS basis. This followed our expectations because SVD basis tends to capture too many noise characteristics when the noise level is high.

Interestingly, IGEN outperformed the standard GEN method in both scenarios, especially when the noise level is high. This indicates that the proposed ensemble method is more robust to noise than the standard GEN method.

5.2 Simulation II

In this section we simulate the data based on a real-world MEG study. The forward operator $X$ is obtained from the head model of a human subject and is normalized. It is computed using a boundary element model based on the subject’s head geometric measurements. It is a 248-by-15,360 matrix. We simulate 100 active signals in two compact sub-areas in left and right hemispheres, respectively. Among these 100 sources, 50 are in a subarea in the left hemisphere and have the same function form. The other 50 are in a sub-area in the right hemisphere and have the same function form with a slightly longer offset in time than that with subarea 1 [see Figure 2 (a)]. These 100 $\beta_j(t)$’s are simulated from a sine-exponential function:

$$\beta_j(t) = 0.6 \sin(5\pi t) \exp(-5t).$$

As we can see, the two peaks for the two source locations are at time points 15 and 30, respectively. Figure 2 (b) and (c) show the overlook of the true brain maps at these two peaks. The source functions are recorded by 50 discrete time points from $(0, 1)$. As with Simulation I, elements from the white noise, $e_i(t)$, are generated from $\mathcal{N}(0, 0.1)$ and $\mathcal{N}(0, 1)$.
in Scenarios 1 and 2, respectively. The response, $y(t)$, is generated by Equation (1).

Figure 3 (a) and (d) show $y_i(t)$’s from 248 hypothetical sensors in Scenarios 1 and 2, respectively. Since $X$ is computed from a real head model, the measured $y_i(t)$’s are not as smooth as those in Simulation I even though they are at the same noise level. Scenario 2 is a more realistic situation because EEG/MEG data usually contain a high level of noise. This is because the brain’s magnetic field ($10^{-10}$ femtotesla) is considerably smaller than the ambient magnetic noise in an urban environment ($10^5$ femtotesla).

Figure 3 (b) and (c) show the estimated time courses in the two active subareas in Scenario 1. Different from Simulation I, the recovered sources listed in these figures are from one arbitrarily chosen source in the active subareas in order to make the graphs clearer. As we can see, STP and MNE recovered the entire time courses well. MCE was able to select one peak/valley in each source location, but it missed the valley for the first source location and the peak for the second source location. In addition, its solution was too “spiky”. Figures 4 (a) and (b) show the recovered brain maps at time points 15 and 30 using STP with SVD basis in Scenario 1. As can be seen, both subareas are identified correctly.

Figure 3 (d) and (e) show the estimated time courses in the two active subareas in Scenario 2. STP performed reasonably well in this Scenario too. Sources from the two active subareas were clearly recovered. MNE was able to identify the peaks for the two source locations with strong artifacts due to the time offset. MCE missed the valley of the first source location, and the spiky effect was obvious. Figures 4 (c) and (d) show the recovered brain maps at time points 15 and 30 using STP with SVD basis in Scenario 2. The two subareas can be identified even though they are not as clean as those in Scenario 1.

Table 2 lists the MSE and PCIS from the three methods. As can be seen, STP outperformed the other two methods due to small MSE and high PCIS in both scenarios. MNE and MCR produced poor MSE and PCIS. MCE is the least satisfactory among all the methods.

Figure 3. The measured signal, $y_i(t)$’s, from hypothetical 248 sensors, and the true, $\beta(t)$’s, from two arbitrary dipoles and their corresponding estimates, $\hat{\beta}_j(t)$’s, by different methods at two locations, respectively, in Simulation II.
In this section we perform sensitivity analyses to examine two of the most important tuning parameters. They are the number of iterations, \( L \), and the maximum sub-sample size, \( p_{\text{max}} \). Presumably, \( L \) should be as large as possible. However, it is impossible to assign an ultra-large \( L \) in practice due to the cost-effectiveness consideration. Therefore, we examine the relationship between \( L \) and \( p_{\text{max}} \) in terms of the MSE. The experiments in this section are all based on Simulation II Scenario 2.

We examine four combinations of \( L \) and \( p_{\text{max}} \): \( L = \frac{2p}{p_{\text{max}}} \), \( \frac{3p}{p_{\text{max}}} \), \( \frac{4p}{p_{\text{max}}} \), \( \frac{5p}{p_{\text{max}}} \). In each combination, three values of \( p_{\text{max}} \) are examined: \( p_{\text{max}} = 10, 20, 50 \). In this simulation \( p = 15,372 \). We plot the values of \( L \) as functions of the MSE [see Figure 5 (a)]. As we can see, when \( L = \frac{p}{p_{\text{max}}} \), the MSE for all three \( p_{\text{max}} \)’s are relatively large. But they all decrease as \( L \) increases. When \( L \geq \frac{2p}{p_{\text{max}}} \), the MSE starts to level off. This indicates that it is reasonable to set values of \( L \) to be approximately equal to \( \frac{2p}{p_{\text{max}}} \), because larger \( L \) will not significantly increase the accuracy but will dramatically increase the computational cost.

We also investigate the choice of \( p_{\text{max}} \) related to the computational cost. Letting \( L = \frac{2p}{p_{\text{max}}} \), we examine four values of \( p_{\text{max}} \): \( p_{\text{max}} = 10, 20, 50, 100 \). For each case, we record the CPU time in seconds on a 64-bit Dell Precision Workstation (24GB RAM, 4G quad CPU). We plot the \( p_{\text{max}} \) as a function of CPU time [see Figure 5 (b)]. As we can see, the CPU time roughly reaches the minimum when \( p_{\text{max}} = 20 \). When \( p_{\text{max}} \) is too small (e.g., 10), the computational cost is high because \( L \) is large. When \( p_{\text{max}} \) is too large (e.g., 50 or 100), \( L \) is small but GEN in each iteration takes a long time due to the large number of variables involved. Therefore, \( p_{\text{max}} \) should be chosen to balance the computational cost and accuracy. Based on our empirical study, we suggest to use \( L = \frac{2p}{p_{\text{max}}} \) and \( p_{\text{max}} = 20 \) for a data set with about 250 observed functions.

### 5.3 Tuning parameters

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Criterion</th>
<th>STP (NCS)</th>
<th>STP (SVD)</th>
<th>MNE</th>
<th>MCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low noise</td>
<td>MSE ((10^{-3}))</td>
<td>63.7</td>
<td>77.1</td>
<td>801.2</td>
<td>992.4</td>
</tr>
<tr>
<td></td>
<td>PCIS (%)</td>
<td>95.4</td>
<td>94.1</td>
<td>60.2</td>
<td>86.7</td>
</tr>
<tr>
<td>High noise</td>
<td>MSE ((10^{-3}))</td>
<td>105.1</td>
<td>117.5</td>
<td>1011.2</td>
<td>1501.0</td>
</tr>
<tr>
<td></td>
<td>PCIS (%)</td>
<td>85.6</td>
<td>80.3</td>
<td>31.0</td>
<td>67.2</td>
</tr>
</tbody>
</table>

### 6. AN MEG EXAMPLE

In this section we provide a real-world MEG example of a 44-year-old female patient with grade three left frontal astrocytoma who underwent the MEG test as part of the presurgical evaluation. The MEG test discussed here is a somatosensory task that is designed to non-invasively identify the somatosensory areas of the patient under evaluation, with the expectation that the primary somatosensory area contralateral to the site of activation will normally be activated. Data collection was done with a whole-head neuromagnetometer containing 248 first-order axial gradiometers. During the MEG somatosensory session, 558 repeated stimulations were delivered to her right lower lip through a pneumatically driven soft plastic diaphragm (diameter: 1 cm) at the pressure setting of 25 psi. Each stimulation lasted 40 ms with 450 ms epoch duration (including prestimulus baseline of 100 ms) and an interstimulus interval randomized between 0.5 s and 0.6 s. The data were collected with an analog high-pass filter of 0.1 Hz, a sampling rate of 508.63 Hz and a bandwidth of 200 Hz. During data processing a bandpass filter of 2 to 40 Hz was utilized and epochs were inspected to identify artifacts, then those 558 epochs without artifacts were averaged to obtain the final ERF (event-related magnetic field) response. Finally, a bad channel (sensor A67) was removed. The MEG device recorded three trials with 228 time points in each epoch. We choose this study because it is an important task to clinical research in presurgical mapping.
Figure 5. Investigating tuning parameters using Simulation II Scenario 1.

(a) MSE as functions of $L$ for three $p_{\text{max}}$ values

(b) CPU time as a function of $p_{\text{max}}$ (fix $L = \frac{2p_{\text{max}}}{P}$)

Figure 6. Human MEG recordings, $y_i(t)$’s, from 247 sensors and reconstructed time courses, $\hat{\beta}_j(t)$’s, by different methods for one arbitrary source in the somatosensory area.

(a) MEG recordings from 247 valid channels

(b) Reconstructed source time courses for one arbitrary source

After preprocessing, the measured data matrix becomes 247-by-228, where 247 is the number of good MEG channels and 228 is the number of recorded data points per epoch. That is, each row of this data matrix represents the time series of an MEG channel. From the first row to the last row, there are time series from sensor A1 to sensor A248 (except A67). The forward operator is arranged accordingly. Figure 6 (a) shows the time series from 247 valid channels.
As we can see, significant activation peaks corresponding to activations of somatosensory cortices are visible in this averaged event-related field response, and the peaks are achieved at around time points 85 and 99. We expect that the reconstructed signals should have significant activations in the somatosensory area at the corresponding two time points.

We apply STP with NCS and SVD bases \((q = 9)\), MNE and MCE to the data. Figure 6 (b) shows the reconstructed time courses for one arbitrary source at the somatosensory cortex. Based on the nature of the experiment, we expect two peaks at around time points 85 and 99, respectively. As we can see, the peak at time point 99 was identified by all methods. However, MCE missed the earlier peak at time point 85 (or the magnitude of the activation was not large enough), and it also introduced discontinuity between the two peaks. MNE and the two versions of STP produced smooth time courses. Note that the MEG recordings we used in this study were preprocessed, i.e., filtered and averaged. Even with STP (SVD) [see Figure 6 (a)] the reconstructed time courses were fairly smooth.

Figure 7 shows side views of the 3-D brain maps at time point 99 by three methods because all methods seem to work well at time point 99 based on the time course study. STP with both bases performed the best because they correctly identified the somatosensory area (located at the left post-central gyrus) and the active area was focal too. MCE also correctly identified the somatosensory area at time point 99, but its entire time courses are unsatisfactory [see Figure 6 (b)]. MNE was inferior to other methods, because the active areas from MNE were too broad. It picked out the somatosensory area but also incorrectly picked out some inactive areas.

### 7. CONCLUSION

We have presented a spatio-temporal penalization procedure for tackling the ill-posed EEG/MEG inverse problem. Given the measured EEG/MEG time courses at the sensor space, the goal of localizing and recovering the potential signal source time courses in the cortex is achieved by solving a nonparametric linear model.

The proposed approach utilizes the temporal bases to expand the sources and projects the temporal and spatial domains onto the same hyperplane. Then the nonparametric linear model is transformed to a high-dimensional linear regression problem. As a result, one source is related to multiple predictors. An iterative group elastic net algorithm is implemented to achieve temporal smoothness and spatial focality of the sources. The implementation of group shrinkage method allows us to select temporal information associated with an active source all together, while shrinking temporal information associated with an inactive source all towards zero. In particular, GEN results in a solution between \(L_1\) and \(L_2\). Hence, the reconstructed sources can smoothly transit from active areas to inactive areas. We demonstrate that STP performed considerably well over an \(L_2\) based MNE method and an \(L_1\) based MCE method in our simulation and MEG studies.

While STP showed advantages in terms of accuracy, a drawback is that it is more computationally expensive than MNE and MCE due to its iterative nature. However, it is claimed that this might not be a severe problem for the inverse problem, which does not require real-time solutions in most of the situations. Hence, a slightly time-consuming method is still desired as long as it can produce good accuracy. Another drawback of STP or probably many other methods based on basis representation is that using bases to represent the data will potentially lose some information. In addition, how to choose an appropriate basis may be another issue. Further investigation on fast solutions without doing basis expansion has been planned.

### ACKNOWLEDGEMENT

The authors thank the MEG Lab at the University of Texas Health Science Center Houston for providing the data. In particular, thanks to Professors Andrew Papanicolaou and Eduardo Castillo for their suggestions and comments.

*Received 17 January 2011*
REFERENCES


Spatio-temporal inverse solver