SPATIOTEMPORAL FILTERING METHODOLOGY FOR SINGLE-TRIAL
EVENT-RELATED POTENTIAL COMPONENT ESTIMATION

By

RUIJIANG LI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2008

© 2008 Ruijiang Li

To all scientists and researchers, who have lived in pursuit of knowledge, and have dedicated themselves to the advancement of science

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Dr. Jose C. Principe, for his great inspiration and encouragement throughout the course of my research. Not just that. He has really become a mentor and guide during pivotal times of my life, which I would have to say regretfully that I did not take full advantage of. One could ask for no more from such an advisor.

I wish to thank the members of my committee, Dr. John Harris, Dr. Jianbo Gao, and Dr. Mingzhou Ding, for their valuable time and interest in serving on my supervisory committee, as well as their comments, which helped improve the quality of this dissertation. I am grateful for Dr. Andreas Keil's expertise on psychology as well as his support, which made our collaboration fruitful.

I would like to thank my friends and colleagues at the Computational NeuroEngineering Laboratory. They have made my stay in Florida during the past four years an enjoyable experience. Last but not least, I wish to thank my parents, who raised me up. Without them, all is in vain.

TABLE OF CONTENTS

	<u>page</u>
ACKNOWLEDGMENTS	4
LIST OF TABLES	8
LIST OF FIGURES	9
ABSTRACT	10
CHAPTER	
1 INTRODUCTION	12
1.1 Basic Concepts of the ERP	12
1.1.1 Generation of the ERP	12
1.1.2 The ERP Components	13
1.2 Estimation of the ERP	16
2 SINGLE-TRIAL ERP ESTIMATION	19
2.1 Single-Trial ERP Estimation Using Single-Channel Recording	19
2.1.1 Time-Invariant Digital Filtering	20
2.1.2 Time-Varying Wiener Filtering	20
2.1.3 Adaptive Filtering	21
2.1.4 Kalman Filtering	22
2.1.5 Subspace Projection and Regularization	22
2.1.6 Parametric Modeling	23
2.1.7 Other Methods Using Single-Channel Recording	24
2.2 Single-Trial ERP Estimation Using Multi-Channel Recording	25
2.2.1 Generative EEG Model	25
2.2.2 What Is a Spatial Filter and What Can It Do?	27
2.3 Review of Spatiotemporal Filtering Methods	28
2.3.1 Principal Component Analysis (PCA)	28
2.3.2 Independent Component Analysis (ICA)	32
2.3.3 Spatiotemporal Filtering Methods for the Classification Problem	36
3 NEW SPATIOTEMPORAL FILTERING METHODOLOGY: BASICS	40
3.1 Spatial Filter as a Noise Canceller in the Spatial Domain	40
3.2 Deterministic Approach	42
3.2.1 Finding Peak Latency	42
3.2.2 Finding Scalp Topography and Peak Amplitude	44
3.3 Stochastic Approach	46
3.4 Simulation Study	48
3.4.1 Gamma Function as a Template for ERP Component	49

3.4.2.	Generation of Simulated ERP Data	49
3.4.3	Case Study I: Comparison with Other Methods	50
3.4.4	Case Study II: Effects of Mismatch	53
4	ENHANCEMENTS TO THE BASIC METHOD	59
4.1	Iteratively Refined Template	59
4.2	Regularization	61
4.2.1	Constrained Optimization	61
4.2.2	Unconstrained Optimization	64
4.3	Robust Estimation: the CIM Metric	65
4.4	Bayesian Formulations of the Topography Estimation	68
4.4.1	Model 1: Additive Noise Model	69
4.4.2	Model 2: Normalized Additive Noise Model	70
4.4.3	Model 3: Original Model	71
4.4.4	Comparison among the Three Models	71
4.4.5	Online Estimation	72
4.5	Explicit Compensation for Temporal Overlap of Components	72
5	APPLICATIONS TO COGNITIVE ERP DATA	90
5.1	Oddball Target Detection	90
5.1.1	Materials and Methods	90
5.1.2	Estimation Results	92
5.1.3	Discussions	98
5.2	Habituation Study	99
5.2.1	Materials and Methods	99
5.2.2	Estimation Results	99
6	CONCLUSIONS AND FUTURE RESEARCH	110
6.1	Conclusions	110
6.2	Future Research	112
APPENDIX		
A	PROOF OF VALIDITY OF THE PEAK LATENCY ESTIMATION IN (3-10)	114
B	GAMMA FUNCTION AS AN APPROXIMATION FOR MACROSCOPIC ELECTRIC FIELD	115
C	DERIVATION OF THE UPDATE RULE FOR THE CONSTRAINED OPTIMIZATION PROBLEM	116
D	DERIVATION OF THE UPDATE RULE FOR THE UNCONSTRAINED OPTIMIZATION PROBLEM	117
E	MAP SOLUTION FOR THE ADDITIVE MODEL	118

F	NORMALIZED ADDITIVE NOISE MODEL 1: MAP SOLUTION	120
G	NORMALIZED ADDITIVE NOISE MODEL 2: MAP SOLUTION	122
	LIST OF REFERENCES	123
	BIOGRAPHICAL SKETCH	135

LIST OF TABLES

<u>Table</u>	<u>page</u>
3-1 Latency estimation: mean and standard deviation	55
3-2 Amplitude estimation: mean and standard deviation	55
3-3 Scalp topography estimation: correlation coefficient	55
3-4 Effects of mismatch I: SNR = -20dB	56
3-5 Effects of mismatch II: SNR = -10dB	56
4-1 Latency estimation: mean and standard deviation	75
4-2 Amplitude estimation: mean and standard deviation	75
4-3 Scalp topography estimation: correlation coefficient	75
4-4 Estimation results for the iteratively refined template method	76
4-5 Estimation with MCC for the mismatch case at SNR = 0dB	76
4-6 Estimation with MCC for the mismatch case at SNR = -20dB	76
4-7 Amplitude estimation for three Bayesian models	77
4-8 Scalp topography estimation for three Bayesian models	77
5-1 Correlation statistics for the 4 subjects: Scalp topography estimation	102
5-2 Regression statistics for response time and estimated amplitude	102

LIST OF FIGURES

<u>Figure</u>	<u>page</u>
3-1 Gamma functions with different shapes and scales.	57
3-2 Waveforms of synthetic and presumed ERP component.....	58
4-1 Mean and standard deviation of the estimated amplitude under different SNR conditions.....	78
4-2 The waveforms of the synthetic component, presumed template and refined template under 4 SNR conditions.	79
4-3 Waveforms of two overlapped components used in regularization.....	80
4-4 Scalp topography of two overlapped ERP components used in regularization.	81
4-5 Amplitude and scalp topography estimation I with regularization (constrained optimization) under 3 SNR conditions..	82
4-6 Amplitude and scalp topography estimation II with regularization (constrained optimization) under 3 SNR conditions	84
4-7 Amplitude and scalp topography estimation I with regularization (unconstrained optimization) under 3 SNR conditions.	86
4-8 Amplitude and scalp topography estimation II with regularization (unconstrained optimization) under 3 SNR conditions.	88
5-1 Pictures used in the experiment as stimuli.....	103
5-2 Cost function in (3.9) versus time lag for different regularization parameters for subject #2..	104
5-3 Estimated pdf of time lags corresponding to local minima of the cost function in (3-9) using the Parzen windowing pdf estimator with a Gaussian kernel size of 4.2..	105
5-4 Scalp topographies for the four subjects.....	106
5-5 Scatter plot of the response time versus the estimated amplitude for each single trial for the four subjects.	107
5-6 Estimated scalp topography for mixed and habituation phase.....	108
5-7 Estimated amplitude for mixed and habituation phase.....	109

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

SPATIOTEMPORAL FILTERING METHODOLOGY FOR SINGLE-TRIAL
EVENT-RELATED POTENTIAL COMPONENT ESTIMATION

By

Ruijiang Li

December 2008

Chair: Jose Principe

Major: Electrical and Computer Engineering

Event-related potential (ERP) is an important technique for the study of human cognitive function. In analyzing ERP, the fundamental problem is to extract the waveform specifically related to the brain's response to the stimulus from electroencephalograph (EEG) measurements that also contain the spontaneous EEG, which may be contaminated by artifacts. A major difficulty for this problem is the low (typically negative) signal-to-noise ratio (SNR) in EEG data. The most widely used tool analyzing ERP has been to average EEG measurements over an ensemble of trials. Ensemble averaging is optimal in the least square sense provided that the ERP is a deterministic signal. However, over four decades of research have shown that the nature of ERP is a stochastic process. In particular, the latencies and the amplitudes of the ERP components can have random variation between repetitions of the stimulus. Under these circumstances, estimation of the ERP on a single-trial basis is desirable. Traditional single-trial estimation methods only consider the time course in a single channel of the EEG. With the advent of dense electrode EEG, a number of spatiotemporal filtering methods have been proposed for the single-trial estimation of ERP using multiple channels.

In this work, we introduce a new spatiotemporal filtering method for the problem of single-trial ERP component estimation. The method relies on modeling of the ERP component local descriptors (latency and amplitude) and thus is tailored to extract faint signals in EEG. The model allows for both amplitude and latency variability in the actual ERP component. The extracted ERP component is constrained through a spatial filter to have minimal distance (with respect to some metric) in the temporal domain from a template ERP component. The spatial filter may be interpreted as a noise canceller in the spatial domain. Study with simulated data shows the effectiveness of the proposed method to signal to noise ratios down to -10 dB. The method is also tested in real ERP data from cognitive experiments where the ERP are known to change, and corroborate experimentally the expected behavior.

CHAPTER 1 INTRODUCTION

1.1 Basic Concepts of the ERP

Event-related potential (ERP) is an important and well-established technique for neuroscientists and psychologists to study human cognitive function. In this section, we briefly review some of the basic facts and concepts related to ERP generation and analysis (Coles et al. 1995).

1.1.1 Generation of the ERP

When a pair of electrodes are attached to the surface of the human scalp and connected to a differential amplifier, the output of the amplifier reveals a pattern of voltage variation over time. This voltage variation is known as the electroencephalograph (EEG). The amplitude of the normal EEG varies between approximately $\pm 100 \mu V$, and most of the EEG frequency contents range between 0.5Hz and 40Hz. Here, we do not review the recording techniques of EEG (Ruchkin 1987).

If we present a stimulus to a human subject while recording the EEG, we can define a period of time (an epoch or a trial) where some of the EEG components are time-locked to the stimulus. Within this epoch, there may be voltage changes that are specifically related to the brain's response to the stimulus. These voltage changes constitute the event-related potential, or ERP.

Although it is not completely understood how the measurements at the scalp relate to the underlying brain activity, the following points appear to be clear and are generally accepted (Scherg and Picton 1991, Wood 1987). First, ERP recorded from the scalp represents net electrical fields associated with the activity of sizeable neuron populations. These neuron populations act as current sources whose electrical fields propagate to the entire scalp through

volume conduction. Second, the individual neurons that compromise such a population must be synchronously active and have a certain geometric configuration to produce measurable potentials at the scalp. In particular, these neurons must be configured in such a way (usually in a parallel orientation) that their individual fields summate to yield a dipolar field.

Therefore, the ERP recorded at the scalp is selective of the totality of the brain activity. This is advantageous in that the resultant measurements would otherwise be so complex as to be difficult or impossible to analyze. On the other hand, we should also be aware that there are certainly numerous functionally important neural processes that cannot be detected by the ERP technique.

1.1.2 The ERP Components

The issue of ERP components has aroused much controversy among the ERP research community, particularly the question of the definition of an ERP component. Suppose for the moment that we have obtained the ERP using some method. A simple way to define a component is to focus on some feature of the resulting waveform (for instance, a peak or trough), and this feature becomes the component of interest. Some common features include the amplitude and latency parameters of a particular peak or trough.

A major problem with the simple approach mentioned above is component overlap, both spatially and temporally. Since the brain is a conducting medium, activity generated in one spatial location may be propagated through the brain tissue and appear at other locations. Thus, the waveform we observe by measuring the voltage at the scalp may well be attributed to a variety of different sources in different spatial locations of the brain. One consequence of volume conduction is that there need be no direct correspondence between the timing of the distinctive features of an ERP waveform (peaks and troughs) and the temporal characteristics of the underlying neural systems. For instance, an ERP peak with a latency of 300 ms, might reflect the

activity of a single neural generator maximally active at that time, or the combined activity of two (or more) neural generators, maximally active before and after 300 ms, but with fields summing to a maximum at that time.

Due to these ambiguities surrounding the interpretation of peaks and troughs in ERP waveforms, other definitions for ERP components have been proposed. Naatanen and Picton (1987) adopted what might be called the physiological approach to component definition. They proposed that a defining characteristic of an ERP component is its anatomical source within the brain. According to this definition, to measure a particular ERP component, we must have a method of identifying the contributing sources. Dochin (1979, 1981) adopted what might be called the functional approach to ERP component definition, which is concerned more with the information processing operations with which a particular component is correlated. According to this definition, it is entirely possible for a component to be identified with a particular feature of the waveform that reflects the activity of multiple generators within the brain, so long as these generators constitute a functionally homogeneous system (Coles, et al.1995).

Although the above physiological and psychological approaches to component definition seem to be counteractive, for many investigators it is more appropriate to combine both approaches. A classical approach to component definition, was proposed by Dochin et al. (1978). They argued that an ERP component should be defined by a combination of its polarity, its characteristic latency, its distribution across the scalp and its sensitivity to characteristic experimental manipulations. Notice that polarity and scalp distribution imply a consistency in physiological source, while latency and sensitivity imply a consistency in psychological function.

ERP components can be broadly classified into two types: exogenous and endogenous components. Characteristics of the exogenous components (amplitude, latency and scalp projection) largely depend on the physical properties of sensory stimuli, such as their modality and intensity. On the other hand, endogenous components largely depend on the nature of the subject's interaction with the stimulus. These components vary as a function of such factors as attention, task relevance and the nature of the information processing required by the stimulus. The dichotomy of the exogenous-endogenous distinction turned out to be an oversimplified version of the reality. Many early 'sensory' components have been shown to be modifiable by cognitive manipulations (e.g., attention) and many of the later 'cognitive' components have been shown to be influenced by the physical attributes of stimulus (e.g., modality).

In what follows, we briefly discuss one particular well-known ERP component, the P300. For a comprehensive review on other well-known components, we refer to Coles, et al. 1995.

The P300 is probably the most important and the most studied component of the ERP. It was first described in the 1960s by Sutton et al. (1965). The P300 is evoked by a task known as the odd-ball paradigm. During this task a series of one type of frequent stimuli is presented to the experimental subject. A different type of non-frequent (target) stimulus is also presented. The task of the subject is to react to the presence of target stimulus by a given motor response, typically by pressing a button, or just by mental counting to the target stimuli. Virtually any sensory modality (auditory, visual, somatosensory, olfactory) can be used to elicit the P300 response. (Polich 1999). The shape and latency of the P300 differs with each modality. This indicates that the sources generating the P300 differ and depend on the stimulus modality (Johnson 1989).

There are several theories on the neural processes underlying the origin of the P300. The most cited and most criticized theory was proposed by Donchin and Coles (1988a, b). According to their theory, the P300 reflects a process of context or memory updating by which the current model of the environment is modified as a function of incoming information. Several investigators (e.g., Johnson 1986) have pointed out that the P300 does not appear to be a unitary component, and instead may represent the activity of a widely distributed system which may be more or less coupled depending on the situation. More information about the underlying neural systems is required before a consensus is attained about the functional significance of this component. For a more recent review on the research of P300, we refer to Linden, 2005.

1.2 Estimation of the ERP

The fundamental problem in the analysis of ERP is to extract the signals that are brain's specific response to the stimulus from the EEG measurements that also contain 'noise'. By noise, or the background EEG, we mean the electrical activities from heart, muscles and eye movements as well as the spontaneous brain activities that are not related to brain's response to the stimulus.

A major difficulty with the extraction of ERP is that, in most cases, ERP signals are small (on the order of microvolts) relative to the background EEG (on the order of tens of microvolts) in which they are embedded. For this reason, it is necessary to employ signal processing techniques to estimate the ERP signals in the presence of noise.

By far the most commonly used technique has been the averaging of the EEG measurements over an ensemble of time-locked epochs. This is optimal in the mean square sense, given the assumption that the ERP is a deterministic signal time-locked to the stimulus and the additive background EEG is zero-mean and uncorrelated with the ERP.

However, for over four decades it has been evident that the nature of ERP is more or less random. In particular, the amplitudes and latencies of the peaks in the ERP can have random variations between repetitions of the stimuli (Brazier 1964). In addition, the variations may be trend-like and the mean of the amplitudes and the latencies can change across the trials. Under these circumstances, the information regarding to these variations in ERP is lost through averaging. Furthermore, the average waveform may not, in fact, resemble the actual ERP waveform that is recorded in an individual trial. The resulting estimates for the ERP, therefore, may not correspond to the underlying neural processes and inference about the cognitive function may be misleading.

Estimation of the ERP on a single-trial basis is desired for the situations when the peak amplitude and latency of a particular component change significantly across trials. A major difficulty with single-trial ERP estimation is again the very low signal-to-noise-ratio (SNR) in the single-trial EEG, typically lower than -10dB.

Statistically speaking, the average ERP, or the sample mean, is an example of the use of the first order statistics, where only the first order moment of the population parameters is estimated. The next obvious improvement is to use the second order statistics, i.e., covariance analysis. The most common approach is to form an estimator (filter) with which the unwanted contribution of the background EEG can be filtered out. To find such an estimator, some models or assumptions are imposed on the ERP and background EEG concerning their respective second order statistics. The estimator that satisfies the minimum mean square criterion can then be derived. The performance of the estimator then largely depends on how realistic these assumptions are.

For historical reasons, these traditional single-trial estimation methods only consider the time course of a single recording channel in the EEG. In some cases, simple ERP components, e.g., the brainstem auditory evoked potentials, can be adequately examined using a single channel. However, for most ERPs, simultaneous recording from multiple electrode locations is necessary to disentangle overlapping ERP components on the basis of their topographies, to recognize the contribution of artifactual potentials to the ERP waveform, and to measure different components in the ERP that may be optimally recorded at different scalp sites (Picton et al. 2000).

Today, high-density EEG can simultaneously record scalp potentials in up to 256 electrodes. This increased number of sensors and thus increased spatial resolution has created a need for signal processing methods that can simultaneously analyze the time series of multiple channels. Recently, various methods have been proposed for single-trial analysis that linearly combine the time series in multiple channels to generate a representation of the observed data that is easier to interpret (Chapman and McCrary 1995, Makeig et al. 1996, Parra et al. 2002). This linear projection combines the information from all the available sensors into a single channel with reduced interference from other neural sources and may provide a better estimate of the underlying neural activity than the EEG measurements in a single channel. The linear projection, in this sense, may be called a ‘spatial filter’ and these methods can be generally called spatiotemporal filtering methods.

We will review these and other single-trial estimation methods in detail in Chapter 2.

CHAPTER 2

SINGLE-TRIAL ERP ESTIMATION

In this chapter, we review the existing methods for single-trial ERP estimation. The methods are broadly categorized into two classes: those based on single-channel EEG recording and those based on multi-channel EEG recording. Methods based on single-channel recording rely solely on the modeling of temporal characteristics of the ERP and EEG, while methods based on multi-channel recording investigate both the spatial and temporal characteristics of the ERP and EEG, and are termed with the general notion of spatiotemporal filtering methods.

2.1 Single-Trial ERP Estimation Using Single-Channel Recording

For the estimation of ERP using a single channel, all the available information is contained in a single time series $x(t)$: the EEG recording at a certain electrode. We assume that the measurements consist of two parts: the signal of interest (ERP) and additive noise (background EEG), denoted by $s(t)$ and $n(t)$, respectively. The observation model for the EEG can be written as:

$$x(t) = s(t) + n(t) \quad (2-1)$$

In this review, the EEG measurements for a single trial $\mathbf{x}$ is a finite-length vector with elements sampled from the original continuous-time waveform. When the time series in (2-1) are interpreted as stochastic processes, $\mathbf{x}$ becomes a random vector. Its length equals the number of samples in one trial. In vector form, the observation model is:

$$\mathbf{x} = \mathbf{s} + \mathbf{n} \quad (2-2)$$

Here we do not attempt to give an exhaustive review on the topic of single-trial ERP estimation with a single channel. For other reviews on the general ERP estimation problem, we refer to Aunon *et al.* (1981), Ruchkin (1987), McGillem and Aunon (1987), Silva (1993), Karjalainen (1997).

2.1.1 Time-Invariant Digital Filtering

Digital filtering is a good place to start for time series analysis. The simplest approach is to design digital filters that have a desired frequency response. Ruchkin and Glaser (1978) used simple moving average FIR filters to estimate ERP on a single trial basis. More complicated ones may be designed to estimate some particular component such as P300 (Farwell, *et al.* 1993).

Wiener filtering may also be used to estimate single-trial ERP, provided that the power spectra of the ERP and background EEG can be estimated appropriately (Aunon and McGillem 1975, Cerutti, *et al.* 1987).

The major problem with linear time-invariant filtering is the fact that the ERP is typically a transient and smooth waveform with no periodicity. The spectrum for this kind of signal is not defined properly. Consequently, the spectra of the ERP and background EEG (if they are estimated) were usually found to overlap significantly (Krieger *et al.* 1995, Spreckelsen and Bromm 1988, Steeger *et al.* 1983). Thus the application of digital filters with constant frequency response is not expected to give desirable results in most cases. The effects of digital filtering are studied in (Ruchkin and Glaser 1978, Maccabee, *et al.* 1983, Nishida *et al.* 1993).

2.1.2 Time-Varying Wiener Filtering

When the optimal Wiener solution is computed for single-trial EEG data, the filtering becomes time-varying with respect to each trial. In general, these methods necessitate some analytic model for the ERP. Yu and McGillem (1983) introduced what was called the time-varying minimum mean square error filter. A crucial task for their method is to obtain a good estimate for the cross-covariance between the ERP and measurements. Under the assumption that the ERP and background EEG are uncorrelated, the cross-covariance becomes the auto-covariance of the ERP signal itself. The ERP is parametrically modeled by a superposition of the

components with random location and amplitude. The parameters for the ERP are then calculated from the Wiener solution on a single trial basis.

2.1.3 Adaptive Filtering

The use of adaptive filtering for the analysis of single-trial ERP, particularly the use of the least mean square (LMS) algorithm (Widrow 1985), was extensively studied during the 1980s (Madhavan *et al.* 1984, 1986; Vila *et al.* 1986; Thakor 1987; Doncarli 1988). For these methods, the measurement $x(t)$ is selected as the desired signal, and several choices are proposed for the input signal.

Thakor (1987) is probably the one of the most cited works among the adaptive filtering methods for ERP estimation. From the principles of adaptive noise cancellation, Thakor proposed a novel way of choosing reference and primary inputs. Two sets of single-trial measurements $x_i(t), x_j(t)$ (i, j being the trial index) serve as the reference and primary inputs, respectively. The idea is to estimate the primary input with a set of delayed version of the reference input on which some form of ensemble averaging is performed. The criticism of Thakor's work is summarized in (Madhavan 1988), where the author asserted that if the ERP signal is assumed to be identical across trials, the above approach does not provide any signal-to-noise ratio improvement and distorts the signal at frequencies where signal and noise power spectra overlap.

Madhavan (1992) proposed a modified adaptive line enhancement method. In this method the pre-stimulus EEG data are adaptively modeled with an autoregressive (AR) model, which is then used to filter the post-stimulus EEG data. The notion of 'modified' means that a non-adaptive filter is used to process the post-stimulus data.

2.1.4 Kalman Filtering

Al-Nashi (1986) adopted the Kalman filtering approach for the ERP estimation problem. It is assumed that the ERP can be modeled as a deterministic signal with additive random noise. The additive noise is assumed to be an autoregressive moving average (ARMA) process and another ARMA model is used for the background EEG. The scalar Kalman filter is then used to predict the single-trial ERP. The basic assumption for Al-Nashi's approach is that the difference between the single-trial ERP and the ensemble average is a stationary process. This is not consistent with (Ciganek 1969), which found that the differences are usually larger in late components than in early components.

Liberati *et al.* (1991) model the single-trial ERP as a time-varying AR process using the ensemble average data and model the background EEG as a stationary AR process using the pre-stimulus data. The AR parameters are then used to create the state and observation equations with the ERP as the unknown states. The single-trial ERP is then estimated using the Kalman filter equations (Kalman 1960).

2.1.5 Subspace Projection and Regularization

The subspace projection approach starts with the linear observation model:

$$\mathbf{x} = \mathbf{s} + \mathbf{n} = H\boldsymbol{\theta} + \mathbf{n} \quad (2-3)$$

where, $\boldsymbol{\theta}$ contains the parameters to be estimated and the ERP signal is constrained to lie in the subspace spanned by some basis vectors, namely the columns of the matrix H .

If the ERP is assumed to consist of positive and negative humps, sampled Gaussian functions may be a good choice for the basis vectors instead of a generic basis (e.g., polynomials). An alternative is to choose the eigenvectors of the EEG data autocorrelation matrix that correspond to the first few largest eigenvalues. This is motivated by the fact that the eigenvectors constitute a basis set with the minimum number of basis vectors that are required to

model the ERP, assuming that the ERP spans nearly the same subspace with the EEG measurements. The least square solution with this basis set is equivalent to the principal component regression approach (Lange 1996, Karjalainen 1997).

The above two basis sets may be combined into a single criterion, with the Gaussian basis vectors modeling the ERP, and the subspace spanned by the eigenvectors representing the prior information about the problem. This leads to the subspace regularization method, which is closely related to the Bayesian mean square estimation (Karjalainen *et al.* 1999). The ERP may also be estimated recursively using Kalman filtering (Karjalainen *et al.* 1996).

2.1.6 Parametric Modeling

There is one type of parametric models, which uses damped sinusoids as basis function for the modeling of single-trial ERP. The model for the ERP with additive noise can be written as:

$$x(t) = \sum_{i=1}^p A_i \rho_i^t \sin(\omega_i t) + n(t) \quad (2-4)$$

Estimation of the parameters A_i, ρ_i, ω_i is a nonlinear problem, which can be solved with an approximation method called Prony's method (Marple 1987). Its use with generalized singular value decomposition was proposed by Hansson and Cedholt (1990) and Gansler and Hansson (1991). The Prony's method was utilized by Hansson *et al.* (1996) for the estimation of single-trial ERP and robust performance was achieved for EEG data with SNR>10dB. A more recent improvement, called piecewise Prony's method was proposed by Garoosi and Jansen (2000) to deal with nonstationary characteristics of the sinusoids.

Another well studied tool that can be used for the analysis of single-trial ERP is the wavelet transform (Daubechies 1992). Wavelets provide a tiling of time-frequency space that gives a balance between time and frequency resolution and they can represent both smooth signals and singularities. This makes them suitable models for the analysis of transient and

nonstationary signals like the ERP (Thakor 1993; Schiff *et al.* 1994; Samar 1995; Coifman 1996; Basar *et al.* 1999; Effern *et al.* 2000; Quian and Garcia 2003).

The idea is based on a technique called ‘wavelet shrinkage’ or ‘wavelet denoising’, which can automatically select an appropriate subset of basis functions and the corresponding wavelet coefficients. This relies on the property that natural signals, such as images, neural activity, can be represented by a sparse code comprising only a few large wavelet coefficients. Gaussian noise, on the other hand, comprises a full set of wave coefficients whose size depends on the noise variance. By shrinking these noise coefficients to zero using a thresholding procedure (Donoho and Johnstone 1994), one can denoise data.

However, the application of the wavelet method to single-trial ERP analysis requires some form of ensemble averaging in order to derive an ‘optimal’ wavelet basis set that is tuned to the ERP signal. Sometimes the appropriate selection of the ERP ensemble may be a difficult task due to the effects of internal and external experimental parameters (Effern *et al.* 2000).

2.1.7 Other Methods Using Single-Channel Recording

Some methods exist that try to explicitly estimate the latencies of the single-trial ERP components. A simple approach is to use cross-correlation of the signal with a template waveform and find the maximum point of the correlation (Gratton *et al.* 1989). Pham *et al.* (1987) applied a maximum likelihood (ML) method to estimate the latencies of ERP assuming a constant shape and amplitude. The ML method was extended in (Jaskowski and Verleger 1999) incorporating variable amplitude into consideration.

Truccolo *et al.* (2003) developed a Bayesian inference framework for estimation of single-trial multicomponent ERP termed differentially variable component analysis(dVCA). Each component is assumed to have a trial-invariant waveform with trial-dependent amplitude scaling factors and latency shifts. A Maximum a Posteriori solution of this model is implemented via an

iterative algorithm from which the component's waveform, single-trial amplitude scaling factors and latency shifts are estimated. The method works well for relatively low-frequency and large-amplitude event-related components.

2.2 Single-Trial ERP Estimation Using Multi-Channel Recording

The use of multi-channel recording for the estimation of single-trial ERP gave rise to a number of spatiotemporal filtering methods. These methods assume, either explicitly or implicitly, a generative EEG model, which we will introduce in the next section. We then explain what is meant by a 'spatial' filter and illustrate what it can do for us in estimating ERP on a single-trial basis. A review on existing spatiotemporal filtering methods is furnished in the following section, where we concentrate on methods that are based on well-known statistical principles.

2.2.1 Generative EEG Model

We start with the 'neural generator' assumption of EEG data, i.e., neuron populations in cortical and subcortical brain tissues act as current sources (Caspers *et al.* 1980, Sams 1984). Within the EEG frequency range (below 100Hz), brain tissues can be assumed to be primarily a resistive medium (Reilly 1992). Thus, according to Ohm's law, the electrical potentials collected at each sensor (channel) as a result of volume conduction, is basically a linear combination of neural current sources (and non-neural artifacts). The linear generative model for EEG data can be written in matrix form:

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \quad (2-5)$$

or:

$$\mathbf{X} = \sum_{i=1}^N \mathbf{a}_i \mathbf{s}_i^T \quad (2-6)$$

where, N is the number of current sources.

We denote the single-trial EEG data with a $D \times T$ matrix $\mathbf{X}$, with D channels and T samples; $\mathbf{S}$ is a $N \times T$ matrix with each row $\mathbf{s}_i^T$ representing the time course of the current density of the i -th current source; $\mathbf{A}$ is an unknown $D \times N$ matrix. Strictly speaking, the number of the neural current sources N is necessarily much larger than the number of channels D . It is usually assumed that the numbers of sources and sensors are equal for the purpose of convenience.

The column vector $\mathbf{a}_i$ of the matrix $\mathbf{A}$ represents the projection of the i -th current source to each sensor at the scalp and is called the *forward model* associated with the source. This scalp projection is generally unknown and depends on the location and orientation of the dipolar current source as well as the conductivity distribution of the underlying brain tissues, skull, skin and electrodes (Parra *et al.* 2005). Thus, if the scalp projection can be estimated, it may provide us some further evidence to the neurophysiological significance of the corresponding estimated source.

An equivalent way to write the generative EEG model (2-5) is to use the notation of time series:

$$\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{s}(t) \quad (2-7)$$

where, $\mathbf{x}(t) = [x_1(t), \dots, x_D(t)]^T$ is a column vector representing the EEG recordings in D channels; $\mathbf{s}(t) = [s_1(t), \dots, s_N(t)]^T$ is a column vector representing the time course of the current density of the sources. $\mathbf{A}$ is the same matrix defined as before.

There is some degeneracy in the model, i.e., the scaling factor of the current source $\mathbf{s}_i^T$ and its corresponding scalp projection $\mathbf{a}_i$. In this case, either the current sources or the scalp projections are constrained to have unit power to avoid ambiguity. We wish to point out that

there is no ambiguity whatsoever if we want to extract or eliminate from the EEG data the contribution of the i -th current source, i.e., $\mathbf{X}^i \propto \mathbf{a}_i \mathbf{s}_i^T$.

2.2.2 What Is a Spatial Filter and What Can It Do?

To illustrate our point, we begin with a simple example. Suppose we measure the EEG from 3 electrodes where only 3 sources are present. Using the time series notation (the numbers are selected for illustration purposes):

$$\begin{cases} x_1(t) = s_1(t) + 2 \cdot s_2(t) + s_3(t) \\ x_2(t) = 2 \cdot s_1(t) + s_2(t) + s_3(t) \\ x_3(t) = s_1(t) + s_2(t) + 2 \cdot s_3(t) \end{cases} \quad (2-8)$$

We would like to recover each of the three sources using the EEG measurements from all the available sensors. Since the measurements are linearly related to the current sources, we speculate that this could be done by linearly combining all the EEG measurements through a weight vector $\mathbf{w}$:

$$y(t) = \mathbf{w}^T \mathbf{x}(t) = \sum_{i=1}^D w_i x_i(t) \quad (2-9)$$

In fact, if we select the weight vector in (2-9) as: $\mathbf{w}^T = [-1, 3, -1]^T$, we will get,

$$y(t) = -x_1(t) + 3 \cdot x_2(t) - x_3(t) = 4s_1(t) \quad (2-10)$$

which is exactly the first current source with a scaling factor. The other two sources can be recovered by using the weight vectors: $[3, -1, -1]^T, [-1, -1, 3]^T$ respectively.

Of course the above example is simplistic, because in reality, there are certainly numerous current sources simultaneously active in the brain and the number of sensors is usually up to a few hundred. We also do not know any of the elements of the matrix $\mathbf{A}$ in general. However, the following point should be clear: by combining the EEG measurements from multiple channels with a simple weight vector, we are able to recover (or estimate) many sources of interest that

could not be recovered using a single channel. In principle, the rejection of the interferences can be perfect, as shown in (2-10) if the coefficients are known.

The weight vector $\mathbf{w}$ in (2-9), which operates on the EEG measurements in the sensor space, is called a *spatial filter*. Just like a filter operating in the time domain, a spatial filter can have either low-pass or high-pass characteristics in the spatial frequency domain. For instance, a spatial filter summing the measurements from a group of neighboring sensors have a low-pass characteristic; the use of a single channel recording corresponds to a high-pass spatial filter with an ‘impulse’ response attenuating the data from all the other channels to zero.

The selection of the spatial filter $\mathbf{w}$ is usually based on some constraints or desired characteristics of the output $y(t)$. Different constraints will generally lead to different methods of extracting the outputs. Loosely speaking, maximum power of the outputs leads to principal component analysis (PCA); statistical independence among the outputs leads to independent component analysis (ICA); and maximum difference between the outputs leads to linear discriminant analysis (LDA). We will review these and other spatiotemporal filtering methods in the following section.

2.3 Review of Spatiotemporal Filtering Methods

In this section, we provide a review on existing spatiotemporal filtering methods for single-trial ERP analysis. Particularly, we focus on two popular and well established methods, namely, principal component analysis (PCA), independent component analysis (ICA).

2.3.1 Principal Component Analysis (PCA)

From the early days of cognitive ERP research, principal component analysis (PCA) was already proposed as a linear, multivariate data-reduction approach (Donchin, 1966). Since then, PCA has been one of most widely used tools among psychologists for ERP analysis (Glaser and

Ruchkin, 1976; Donchin and Heffley, 1978; Mocks and Verleger 1991; Chapman and McCrary 1995; van Boxtel 1998). By identifying unique variance patterns in a given set of ERP data, PCA decomposes the variance structure of the observed data into a set of latent variables that ideally correspond to the individual ERP components. In the ERP research community, these latent variables are usually called ‘factors’ instead of ‘components’ to avoid confusion with the ERP components.

Among the vast literatures on PCA applied to ERP analysis, one classical method using the PCA-Varimax strategy, is particularly popular and is the primary analytic tool for many ERP researchers (Gaillard and Ritter 1983). The method treats the recorded potential at a given time of the EEG epoch as variables. The domain of the observations is taken to be the Cartesian product of the recording channels, experimental conditions, participants. Suppose we have T samples in a given EEG epoch, D recording channels, C experimental conditions and P participants. The data matrix for this method has a dimension of $T \times (D \times C \times P)$. This particular arrangement of the data matrix leads to the so-called temporal PCA approach, which gives orthogonal factors (eigenvectors of the covariance matrix). The PCA solution is then followed by the Varimax rotation (Kaiser, 1955). The Varimax rotation is an orthogonal rotation that aims to maximize the values that are large for a factor and minimize the values that are small (by maximizing the fourth power the factor). This corresponds to the ‘maximum compactness’ criterion, which will make the new factors have a small number of large values and a large number of zero (or small) values. This is reasonable for ERP estimation because for the most part, ERP components appear to be monophasic and compact in time.

It is easy to see that the above PCA-Varimax approach is a spatiotemporal filtering method. We denote $\mathbf{X}_{c,p}$ as the single-trial EEG data defined in (2-5) from c -th experimental condition and p -th participant. Then the covariance matrix is:

$$\mathbf{C} = \frac{1}{D^2} \mathbf{X}^T \cdot \mathbf{X} \quad (2-11)$$

where,

$$\mathbf{X} = \frac{1}{PC} \sum_{p=1}^P \sum_{c=1}^C \mathbf{X}_{c,p} \quad (2-12)$$

Formally, PCA is equivalent to the singular value decomposition (SVD) of the data matrix defined in (2-12), which is the average EEG data matrix for all experimental conditions and all participants. Suppose we have the SVD of $\mathbf{X}$ as follows:

$$\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \quad (2-13)$$

where, $\mathbf{U}, \mathbf{V}$ are orthogonal matrices of dimension $D \times D$ and $T \times T$, and contain the left-singular and right-singular vectors, respectively. $\mathbf{\Sigma}$ contains the singular values of the data matrix. Equation (2-13) can be equivalently written as:

$$\mathbf{U}^T \mathbf{X} = \mathbf{\Sigma} \mathbf{V}^T \quad (2-14)$$

The Varimax rotation procedure simply adds another $D \times D$ orthogonal matrix $\mathbf{R}$ multiplied on the both sides of (2-14):

$$\mathbf{R} \mathbf{U}^T \mathbf{X} = \mathbf{R} \mathbf{\Sigma} \mathbf{V}^T \quad (2-15)$$

The right side of (2-15) is a $D \times T$ matrix, whose rows can be seen as the factors extracted by the PCA-Varimax approach.

We define a $D \times D$ matrix:

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_D^T \end{bmatrix} = R\mathbf{U}^T \quad (2-16)$$

We further denote:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_D^T \end{bmatrix} = R\mathbf{\Sigma}\mathbf{V}^T \quad (2-17)$$

Thus, we have:

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{X} \quad (2-18)$$

or:

$$\mathbf{y}_i^T = \mathbf{w}_i^T \cdot \mathbf{X}, \quad \text{for } i = 1, \dots, D \quad (2-19)$$

This is the familiar form for the spatial filter defined in (2-9), which is now written in matrix form. Clearly, the matrix $\mathbf{W}$ is an orthogonal matrix. So PCA finds a number of (D) outputs that are uncorrelated with the constraint that the spatial filters are orthogonal. On the other hand, the PCA-Varimax method searches for outputs that are maximally compact in time while still constraining the spatial filters to be orthogonal.

In the context of the generative EEG model, PCA basically assumes that there are equal number of sources and channels. If we multiply $\mathbf{W}^{-1}$ on both sides of (2-18), we get:

$$\mathbf{X} = \mathbf{W}^{-1}\mathbf{Y} \quad (2-20)$$

Thus, the rows of the output matrix $\mathbf{Y}$ contain the time course of the current sources, while the columns of the matrix $\mathbf{W}^{-1}$ constitute the scalp projections of the corresponding sources. This means that the scalp projections for the underlying current sources are orthogonal to each other, which is a highly dubious assumption.

Due to the above problem, an oblique rotation like Promax (Hendrickson and White 1964) has been proposed as a post-processing stage after Varimax, to relax the orthogonality constraint on the scalp projections. Studies with both simulated and real dataset have shown that temporal PCA with Promax extracted markedly more accurate ERP components (Dien 1998).

An alternative approach to the popular temporal PCA is the spatial PCA (Duffy *et al.* 1990, Donchin 1997, Spencer, *et al.* 1999), which treats the recorded potential at a given channel as variables. The EEG data matrix is formed with channel as one dimension, and time by experimental condition by participant as the other dimension. The same rotation procedures follow as in temporal PCA. However, spatial PCA still assumes orthogonality of the scalp projections.

Two other well-documented problems for the PCA approach are the misallocation of variance (Wood and McCarthy 1984) and the issue of latency jitter. Dien (1998) using extensive simulations, has argued that spatial PCA as a complement to temporal PCA, together with parallel analysis (Horn, 1965) to identify noise factors, and oblique rotation to allow for correlated factors, can address these and other shortcomings of PCA.

More recent developments include a combined spatial and temporal PCA approach that is successfully applied to real ERP data extracting known ERP components (Spencer *et al.* 1999, 2001). Dien *et al.* (2005) have presented a standard protocol to optimize the performance of PCA when it is applied to ERP datasets, recommending the use of covariance matrix over correlation matrix, and Promax rotation over Varimax rotation, etc.

2.3.2 Independent Component Analysis (ICA)

ICA was originally proposed to solve the blind source separation (BSS) problem, to recover a number of source signals after they are linearly mixed and observed in a number of sensors, while assuming as little as possible about the mixing process and the individual sources

(Comon 1994). The most basic ICA model assumes linear and instantaneous mixing, which means that the source signals arrive at the sensors without time delay and are mixed in the sensors linearly with other source signals. This basic ICA model naturally fits into the generative EEG model in (2-5), which we repeat here, assuming that the sources and measurements are random vectors:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \quad (2-21)$$

In the ICA literature, the observation $\mathbf{x}$ is called ‘mixtures’, and the unknown matrix $\mathbf{A}$ is called ‘mixing matrix’. Usually it is assumed that the number of sources is equal to the number of sensors. In this case, the mixing matrix becomes a square matrix.

The key assumption used in ICA to solve the BSS problem is that the time courses of the sources are as statistically independent as possible. Statistical independence means that the joint probability density function (pdf) of the outputs factorizes. For the linear instantaneous BSS problem, the solution is in the form of a square ‘demixing’ matrix $\mathbf{W}$, specifying spatial filters that linearly invert the mixing process. If the mixing matrix is invertible, the outputs should be identical to the original source signals, except for scaling and permutation indeterminacies (Comon 1994):

$$\mathbf{y} = \mathbf{W}\mathbf{x} \quad (2-22)$$

There are a multitude of algorithms that have been proposed to solve the basic ICA problem, among which, Infomax (Bell and Sejnowski 1995, Lee *et al.* 1999), FastICA (Hyvarinen and Oja, 1997), JADE (Cardoso, 1999), SOBI (Belouchrani *et al.* 1997) are probably the most widely used. Some algorithms are based on the canonical information-theoretic contrast function for ICA, i.e., mutual information, or its approximations (Infomax, FastICA, etc.); others utilize higher-order statistics of the data (e.g., forth-order cumulant) to perform source separation

(JADE); still others make use of the difference in the temporal spectra of the source signals (SOBI). For a more detailed review on ICA and its applications to BSS problems, we refer to the following: Cardoso (1998); Hyvarinen *et al.* (2001); Roberts and Everson (2001); Cichocki and Amari (2002); James and Hesse (2004); Choi *et al.* (2005). Review papers comparing different ICA algorithms and their relationships are also available: Hyvarinen (1999); Lee *et al.* (2000).

We notice that while PCA requires the spatial filters to be orthogonal, here in the case of ICA, there is no more constraint on the spatial filters (or the demixing matrix $\mathbf{W}$). On the other hand, while PCA only uses second-order statistics (the covariance matrix), to decorrelate outputs, ICA imposes a much stronger condition, i.e., independence on the outputs. The fact that ICA tries to factorize the joint pdf of the outputs implies that all the higher-order statistics (HOS) are taken into consideration by ICA. This means that for non-Gaussian data, the structures contained in HOS (e.g., kurtosis), while totally ignored by PCA, may be captured by ICA. Since many natural signals are non-Gaussian distributed (e.g., speech signals usually follow a Laplacian distribution), ICA may be more suitable for this and other applications than PCA.

Since Makeig *et al.* (1996) published their seminal paper on the application of ICA to EEG data, there have been numerous studies during the last decade dedicated to this research topic (Makeig *et al.* 1997, 1999, 2002, 2004; Vigario *et al.* 1998, 2000; Jung *et al.* 1999, 2000, 2001; Delorme *et al.* 2002, 2003, 2007; Debener *et al.* 2005). Until now, ICA and its variants still remains a powerful tool for the analysis of EEG and ERP data.

The application of ICA to the study of EEG data requires that the following conditions be satisfied: (1), statistical independence of all the underlying neural source signals; (2), their linear instantaneous mixing at the sensors; (3), the stationarity of the mixing process.

Since most of the energy in EEG data lies below 100Hz, the quasistatic approximation of Maxwell equations holds. So there is (virtually) no propagation delay of the electrical potentials from the neuronal sources to the sensors through volume conduction. Thus the assumption of instantaneous mixing is valid. The linearity of the mixing follows from the Maxwell equations as well. The stationarity of the mixing process corresponds to a constant mixing matrix. For the dipole source model, this means that the dipolar neuronal sources should have fixed locations and orientations. Although there is no reason to believe that these neuronal sources are spatially fixed over time, for those that are involved in a specific information processing task and therefore are of interest to ERP researchers, they should at least have a relatively stable configuration or a stable scalp projection, which is congruent with the definition of ERP components as proposed by Fabiani *et al.* (1987).

We have seen that conditions (2) and (3) are approximately valid for EEG data. The most debatable and perhaps perplexing condition is the first one: statistical independence of all the underlying neural source signals. The independence criterion applies solely to the amplitudes of the source signals, and does not correspond to any consideration of the morphology or physiology of the neural structures. However, the different nature of the sources originated from completely different mechanisms often yields signals that appear to be statistically independent. Particularly, analysis of the distributions of artifacts such as the cardiac cycle, ocular activity has shown the statistical independence assumption approximately holds (Vigario 2000).

Although ICA continues to be a useful tool for EEG and ERP analysis, there are also some limitations to it. First, ICA can decompose up to (or at most) D sources from data recorded at D scalp electrodes (D may be ranged from several dozen to a few hundred). On one hand, the researcher has to analyze the extracted D components one by one (including the time course and

scalp projection), which is laborious when D is large and the results are subject to interpretation. If he/she chooses to analyze only a part of all the components, the subsequent analysis is correlated with the retention criterion. (Note that PCA also has this problem). On the other hand, the effective number of statistically independent signals contributing to scalp EEG is almost certainly much larger than the number of electrodes D . Using simulated EEG data, Makeig *et al.* (2000) has found that given a large number of sources with a limited number of available channels, ICA algorithm can accurately identify a few relatively large sources but fails to reliably extract smaller and briefly active sources. This suggests that ICA decomposition in high dimensional space is an ill-posed problem.

Second, the assumption of statistical independence used by ICA is violated when the training dataset is too small or separate topographically distinguishable phenomena nearly always co-occur in the data (Li and Principe, 2006). In the latter case, simulations show that ICA may derive a single component accounting for the co-occurring phenomena, along with additional components accounting for their separate activities (Makeig *et al.* 2000).

These limitations imply that the results obtained with ICA must be validated by researchers using behavioral and/or physiological evidence before their functional significance can be correctly interpreted. Current research on applications of ICA is focused on incorporating domain-specific knowledge into the ICA framework. Recently there has been work on combining ICA with the Bayesian approach (Tsai, *et al.* 2006) or with the regularization technique (Hesse and James, 2006).

2.3.3 Spatiotemporal Filtering Methods for the Classification Problem

It is worthwhile to point out a related but different approach, which is the (supervised) single-trial EEG classification problem. It is generally less difficult than the (unsupervised) single-trial estimation problem in the sense that the availability of label information for

classification facilitates learning. Many spatiotemporal filtering methods have been proposed for the single-trial EEG classification problem, which include, but not limited to, common spatial patterns (CSP) (Ramoser et al., 2000), common spatio-spectral patterns (CSSP) (Lemm et al., 2005), linear discriminant analysis (Parra et al., 2002), bilinear discriminant component analysis (Dyrholm et al., 2007).

The common spatial patterns method was initially proposed by Koles *et al.* (1990) to classify normal versus abnormal EEG (Koles *et al.* 1994). The method has been used for single-trial EEG classification in brain-computer interface (BCI) systems (Müller-Gerking *et al.* 1999; Ramoser *et al.* 2000).

Given the single-trial EEG data for two different experimental conditions, the CSP method decomposes the EEG data into spatial patterns, which maximize the difference between the two conditions. The spatial filters are designed such that the variances of the outputs are optimal (in the least-square sense) for the discrimination of the two conditions. This is realized by simultaneously diagonalizing the two covariance matrices of the EEG associated with the two experimental conditions. The two resulting diagonal matrices (containing the eigenvalues for the two covariance matrices) add up to the identity matrix. Thus, the spatial filters that give the (n , an integer) largest variance in their outputs (associated with the largest eigenvalues) for one condition, will accordingly give the (n) smallest variance in their outputs for the other condition; and vice versa. It is along these directions that the largest differences between the two conditions lie. In (Müller-Gerking *et al.* 1999), the CSPs are called the source distribution matrix (equivalent to the mixing matrix in ICA), and the spatially filtered outputs are claimed to be the source signals, although the EEG data were temporally band-pass filtered between 8-30Hz prior to analysis. Para *et al.* (2005) showed that the simultaneous diagonalization of the covariance

matrices is equivalent to the generalized eigenvalue decomposition, and according to Parra and Sajda (2003), the CSP method is in fact estimating the independent components of the temporally filtered EEG data.

The original CSP method does not take into account the temporal information of the filtered EEG data. In light of this, Lemm *et al.* (2005) proposed an algorithm called common spatio-spectral pattern (CSSP), which utilized the method of delay embedding and extended the CSP algorithm to the state space (with only one tap-delay). Dornhege *et al.* (2006) further improved the CSSP algorithm by optimizing an arbitrary finite-impulse response (FIR) filter within the CSP framework. The overfitting of the spectral filter is controlled by a regularizing sparsity constraint.

The CSP method and its variants all use the relevant oscillatory brain activity for EEG classification. Sometimes it is more appropriate to use coherent evoked potentials (of low-pass nature) instead. Para *et al.* (2002) proposed a spatiotemporal filtering method using conventional linear discrimination to compute the optimal spatial filters for single-trial detection in EEG. Specifically, the search for the optimal spatial filter given the single-trial EEG data as in (2-7), is based on constraining the output $y(t)$ to be maximally discriminating between two different experimental conditions. The optimality criterion is restricted to a pre-specified time interval, i.e., the time corresponding to a number of samples prior to an explicit button push. After finding the optimal spatial filter using conventional logistic regression, the output is averaged within that period of time to obtain a more robust feature. The detection performance is then evaluated using receiver operating characteristic (ROC) analysis on a single-trial basis.

Unlike other conventional methods such as ICA, where the scalp projections are given directly by inverting the demixing matrix containing all the spatial filters, here since there is only

one spatial filter and one output, other techniques have to be sought in order to estimate the scalp projection associated with that output. Parra *et al.* did this by projecting the EEG data to the discriminating output $y(t)$ assuming that the output is uncorrelated with all other brain sources, and found that the discrimination model captured information directly related to the underlying cortical activity. The method was improved in (Luo and Sajda, 2006), where the pre-specified time interval is allowed to be different and optimized for each EEG channel. This effectively defines a discrimination trajectory in the EEG sensor space.

CHAPTER 3

NEW SPATIOTEMPORAL FILTERING METHODOLOGY: BASICS

In this chapter, we propose a new spatiotemporal filtering method for single-trial ERP estimation. The method relies on modeling of the ERP component descriptors and thus is tailored to extract small signals in EEG. The model allows for both amplitude and latency variability in the actual ERP component. We constrain through a spatial filter $\mathbf{w}$ the extracted ERP component to have minimal distance (with respect to some metric) in the temporal domain from a presumed ERP component. Note that we do not constrain the entire ERPs, but instead a single ERP component. We maintain the point in the next section that the spatial filter may be interpreted as a noise canceller in the spatial domain. We then introduce two approaches for the proposed method: the deterministic approach and the stochastic approach.

3.1 Spatial Filter as a Noise Canceller in the Spatial Domain

Since the method deals with one ERP component at a time, we wish to distinguish between ‘signal’ and ‘noise’ instead of using the general term ‘sources’. To accommodate this distinction, we rewrite the generative EEG model in (2-6) as follows:

$$\mathbf{X} = \mathbf{a} \cdot \mathbf{s}^T + \sum_{i=1}^N \mathbf{b}_i \mathbf{n}_i^T \quad (3-1)$$

where, $\mathbf{s}$ is the time course of the ERP component to be extracted, $\mathbf{n}_i$ denotes noise in general.

The distinction between ‘signal’ and ‘noise’ is somewhat arbitrary, e.g., when P300 is the signal of interest, N100 will become noise in the model. Note that for notational convenience, we have assumed the effective number of sources to be $N + 1$.

The EEG model in (3-1) can in turn be rewritten as:

$$\mathbf{X} = \sigma_s \mathbf{a}_o \cdot \mathbf{s}_o^T + \sum_{i=1}^N \sigma_i \mathbf{b}_{oi} \cdot \mathbf{n}_{oi}^T \quad (3-2)$$

where, $\mathbf{a}_o$, $\mathbf{s}_o$, $\mathbf{b}_{oi}$, $\mathbf{n}_{oi}$ are the normalized versions (with respect to some norm, e.g., l_2) of their counterparts in (3-1) and $\sigma_i > \sigma_{i+1}$. The scalars σ_s, σ_i may be seen respectively as the overall contributions of the signal and noise to the single-trial EEG data. In the case of independent noise, we may define the SNR for the single-trial EEG data (note that it is different from SNR in a single channel.) as:

$$SNR = 20 \log \left(\sigma_s / \sqrt{\sum_{i=1}^N \sigma_i^2} \right) \quad (3-3)$$

The vectors $\mathbf{a}_o$, $\mathbf{b}_i'$, represent the scalp topography of the corresponding signal and noise. For a meaningful ERP component, it must have a stable scalp topography $\mathbf{a}_o$. Thus, we may assume that $\mathbf{a}_o$ is fixed for all trials. We also assume that the waveform of a particular ERP component $\mathbf{s}_o$ (dimensionless) remains the same for all trials, although its amplitude σ_s may change across trials.

Next, we claim the following lemma, which is basically a direct consequence of the linear generative EEG model in (3-1).

Lemma: There exists a spatial filter $\mathbf{w}$, that will completely reject the interference from the first $D-1$ largest noise in the output when it is applied to the single-trial EEG data, if,

$$\det[\mathbf{a} \quad \mathbf{b}_1 \quad \dots \quad \mathbf{b}_{D-1}] \neq 0 \quad (3-4)$$

Further, the extracted ERP component will approach the actual ERP component if,

$$\sigma_s \gg \sigma_D \quad (3-5)$$

Such a spatial filter $\mathbf{w}$ can be found by taking the first row of the inverse of the matrix in (3-4). Note that (3-4) implies that,

$$\text{angle}(\mathbf{a}, \mathbf{b}_i) \rightarrow 0, i = 1, \dots, D-1 \text{ and, } \text{angle}(\mathbf{b}_i, \mathbf{b}_j) \rightarrow 0, 1 \leq i < j \leq D-1 \quad (3-6)$$

which means that at least the scalp topography of the source and the first $D - 1$ noise should not be the same or very similar to each other from a computational perspective.

Here, we wish to stress the point that the spatial filter specified in the above lemma may be interpreted as a noise canceller in the spatial domain. It may or may not be the optimal spatial filter for enhancing the SNR in the extracted component. In addition, the SNR enhancement due to the spatial filter increases monotonically with the number of channels (electrodes) if the EEG data were measured in those channels. This means that the more channels we use to record the EEG, the higher SNR we will get (theoretically) in the extracted component. Note that the lemma is an existence statement, it does not tell us how to find such a ‘optimal’ spatial filter. This will be the subject of the next two sections.

3.2 Deterministic Approach

Most ERP components are monophasic waveforms with compact support in time. The morphology of the waveform can be considered relatively fixed due to the common cytoarchitecture of the neocortex and similar neuron populations, but may vary in both its peak latency and amplitude from trial to trial. Based on this, we assume that a particular ERP component can be modeled by a fixed *dimensionless* template (e.g. no physical unit), in the temporal domain, denoted by $s_o(l)$ (where l is the unknown peak latency), multiplied by an unknown and possibly variable amplitude σ_s across trials. We attempt to estimate the variable peak latency and amplitude on a single-trial basis.

3.2.1 Finding Peak Latency

Since we do not know the peak latency in a single trial, we denote the template as $s_o(\tau)$ with a variable time lag parameter τ and slide it one lag a time to search for the peak latency. The search for the optimal filter $\mathbf{w}$ could be realized by minimizing some distance measure

between the spatially filtered output $\mathbf{w}^T \cdot \mathbf{X}$ and the assumed waveform $\mathbf{s}_o(\tau)$ for the particular ERP component. We propose the following cost function based on second-order statistics (SOS):

$$\min_{\mathbf{w}} \left\| \mathbf{w}^T \cdot \mathbf{X} - \mathbf{s}_o(\tau) \right\|_2^2 \quad (3-7)$$

Note that the above optimization is with respect to $\mathbf{w}$ only, with τ fixed. The optimal solution for $\mathbf{w}$ is given by:

$$\mathbf{w}(\tau) = (\mathbf{X}\mathbf{X}^T)^{-1} \cdot \mathbf{X} \cdot \mathbf{s}_o(\tau) \quad (3-8)$$

Obviously, the optimal spatial filter $\mathbf{w}$ depends on which ERP component is to be extracted, and also is a function of the variable time lag τ . From (3-7) and (3-8), we obtain the cost solely as a function of the time lag τ :

$$J(\tau) = \left\| \mathbf{s}_o(\tau)^T \cdot \left[\mathbf{X}^T (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X} - \mathbf{I} \right] \right\|_2^2 \quad (3-9)$$

The peak latency of the ERP component can be set as the time lag where the local minimum of $J(\tau)$ occurs within the meaningful range of peak latencies (T_s) for that particular component (provided that its waveform is monophasic) i.e.,

$$l = \arg \min_{\tau \in T_s} J(\tau) \quad (3-10)$$

The estimated ERP component is then (this need not be normalized):

$$\mathbf{y}_s(l) = \mathbf{X}^T \cdot \mathbf{w}(l) \quad (3-11)$$

It can be shown that under certain conditions, the solution in (3-10) is identical to the true peak latency of the ERP component (Appendix A). Exact match between the modeled and actual ERP component is not a necessary condition for the solution in (3-10) to be correct. For instance, it is easy to show that when both components are symmetric waveforms, then (3-10) also gives the correct latency.

3.2.2. Finding Scalp Topography and Peak Amplitude

In the following, we make the index for trial number k explicit. Denote the estimated ERP component for k -th trial by (the peak latency depends on the trial number):

$$\mathbf{y}_k(l_k) = \sigma_k \mathbf{y}_{sk}(l_k) \quad (3-12)$$

We can absorb the scalar σ_k into a variable scalp topography:

$$\mathbf{a}_k = \sigma_k \mathbf{a}_o \quad (3-13)$$

In order to estimate the unknown scalp topography and amplitude of the ERP component, we assume that the ERP component is uncorrelated with all the noise sources. Replacing the dimensionless ERP component in (3-2) by its estimate in (3-11) and multiplying $\mathbf{y}_{sk}(l_k)$ on both sides of (3-2), we will get an estimate for the single-trial scalp topography (the cross terms $\mathbf{n}_{oi}^T \cdot \mathbf{y}_{sk}(l_k)$ vanish because of the uncorrelatedness assumption):

$$\hat{\mathbf{a}}_k = \frac{\mathbf{X}_k \cdot \mathbf{y}_{sk}(l_k)}{\mathbf{y}_{sk}^T(l_k) \mathbf{y}_{sk}(l_k)} \quad (3-14)$$

Taking the normalized version (note that $\hat{\mathbf{a}}_k$ is in *Volts*) we have,

$$\hat{\mathbf{a}}_{ok} = \frac{\hat{\mathbf{a}}_k}{\sqrt{\hat{\mathbf{a}}_k^T \hat{\mathbf{a}}_k}} \quad (3-15)$$

Ideally, the normalized $\hat{\mathbf{a}}_{ok}$ should be the same as the dimensionless scalp topography $\mathbf{a}_o$.

However, in low-SNR EEG data, the above estimation in (3-14) is very poor, due to the interference from background activity in the finite-sample data. To estimate the scalp topography for a stable ERP component, we propose the following cost function:

$$\min_{\mathbf{a}_o} \sum_{k=1}^K \|\mathbf{a}_o - \hat{\mathbf{a}}_{ok}\|_2^2 \quad (3-16)$$

This corresponds to a maximum likelihood (ML) estimator for $\mathbf{a}_o$ under the assumption that each entry of the normalized single-trial scalp topography is an independent identically distributed (*i.i.d.*) Gaussian random variable. The optimal solution for (3-16) is a simple average of the estimated single-trial scalp topographies for all K trials. Taking the normalized version, the following estimate for $\mathbf{a}_o$ is obtained:

$$\hat{\mathbf{a}}_o = \frac{1}{K} \sum_{k=1}^K \hat{\mathbf{a}}_{ok} \bigg/ \left\| \frac{1}{K} \sum_{k=1}^K \hat{\mathbf{a}}_{ok} \right\|_2 \quad (3-17)$$

Notice that (3-17) is in fact a weighted average of the estimates in (3-14). We also point out that (3-17) is different from summing up directly (3-14) for all trials since the peak latency parameter is involved and it changes from trial to trial.

In the ideal case, the two vectors $\mathbf{a}_o$ and $\mathbf{a}_k$ are identical except for a scaling factor, which is exactly the unknown amplitude σ_k associated with the ERP component in the k -th trial. Replacing their respective estimates in (3-14), we can find σ_k using again a SOS criterion:

$$\min_{\sigma_k} \left\| \hat{\mathbf{a}}_k - \sigma_k \hat{\mathbf{a}}_o \right\|_2^2 \quad (3-18)$$

Simple calculation leads to the following estimate for the amplitude:

$$\hat{\sigma}_k = \hat{\mathbf{a}}_o^T \cdot \hat{\mathbf{a}}_k \quad (3-19)$$

This estimate involves information from all the available channels. In order to eliminate the indeterminacies of the linear generative EEG model, we set the peak amplitude as the maximum of the estimated ERP component in (3-12), i.e., $\hat{\sigma}_k \cdot \mathbf{y}_{sk}(l_k)$. (All the amplitudes in the rest of the paper refer to this quantity). Accordingly, the contribution of the ERP component to the EEG data may be computed by:

$$\hat{\mathbf{X}}_s = \hat{\sigma}_s \cdot \hat{\mathbf{a}}_o \cdot \mathbf{y}_{sk}(l_k) \quad (3-20)$$

These estimates for the scalp topography, peak latencies and amplitudes of ERP component can be used to analyze its psychological significance on a single-trial basis.

Note that we do not directly compute the amplitude from the estimated component, nor do we measure it in any single channel. Instead, the amplitude is computed in (3-19) indirectly through an inner product of two scalp topographies, which involves information from all available channels. These estimates for the scalp topography, peak latencies and amplitudes of ERP component can then be used to analyze their psychological significance on a single-trial basis.

We wish to point out that in contrast to all the spatiotemporal methods mentioned before, where only one, representative spatial filter (or matrix consisted of spatial filters), is computed given all the EEG data, here the optimal spatial filter is computed on a single-trial basis, i.e., given a single-trial EEG data matrix, we can get a spatial filter, as in (3-8). The reason we did it in this way is that, we believe that in theory the optimal spatial filter that is designed to extract small signals should change from trial to trial. Note in (3-2) the ‘noise’ sources are sorted in decreasing order of their power. It is likely that the noisy sources that have relatively large power change drastically across trials. In effect, this will change the configuration in (3-2) and accordingly, the optimal spatial filter will also change.

3.3 Stochastic Approach

In the deterministic approach, the ERP component is considered to be a deterministic signal except for a random latency and amplitude across trials. It does not take into account the intrinsic error in the modeling of the ERP component itself. Here we propose a stochastic approach for the spatiotemporal filtering method. The idea is to constrain the extracted ERP component to be close to the presumed component with respect to some statistic.

We still use the generative EEG model as in (3-1), but here both the signal and the noise are interpreted as *stochastic processes*. The approach utilizes the following observation model to search for the optimal spatial filter:

$$\mathbf{X}^T \mathbf{w} = \mathbf{s} + \mathbf{v} \quad (3-21)$$

where, $\mathbf{s}$ is the actual ERP component and $\mathbf{v}$ is the observation noise appearing in the spatially filtered EEG data. They are both random vectors with each entry as a sample within a certain time interval from the corresponding stochastic processes.

We assume that the ERP component is generated from the following additive model:

$$\mathbf{s} = \mathbf{g} + \mathbf{u} \quad (3-22)$$

where, $\mathbf{g}$ is a fixed signal with a certain morphology serving as the template for the ERP component, and $\mathbf{u}$ represents the model uncertainties of the ERP component and is assumed to be independent of the observation noise $\mathbf{v}$.

Given the above model in (3-21) and (3-22), the optimal spatial filter may be found by maximizing the log-likelihood of the filtered EEG data $\mathbf{y} = \mathbf{X}^T \mathbf{w}$:

$$\max L(\mathbf{w}) = \log p(\mathbf{y} | \mathbf{g}) \quad (3-23)$$

or,

$$\max L(\mathbf{w}) = \log p(\mathbf{u} + \mathbf{v}) \quad (3-24)$$

The log-likelihood function has a simple form under the assumption that $\mathbf{u}$ and $\mathbf{v}$ are zero-mean Gaussian stochastic processes with covariance matrices $\mathbf{C}_M, \mathbf{C}_D$ respectively, and they are independent of each other, since $\mathbf{u} + \mathbf{v}$ is nothing but another Gaussian stochastic process with covariance matrix $\mathbf{C}_M + \mathbf{C}_D$. In this case, maximizing the log likelihood of the (transformed) observed data given the observation model and the template of the ERP component yields the following solution for the optimal spatial filter:

$$\mathbf{w}_o = \arg \min_{\mathbf{w}} \left(\mathbf{X}^T \mathbf{w} - \mathbf{g} \right)^T (\mathbf{C}_M + \mathbf{C}_D)^{-1} (\mathbf{X}^T \mathbf{w} - \mathbf{g}) \quad (3-25)$$

This is a quadratic form of $\mathbf{w}$, so it has a closed-form solution:

$$\mathbf{w}_o = \left(\mathbf{X} \mathbf{\Sigma}^{-1} \mathbf{X}^T \right)^{-1} \cdot \mathbf{X} \mathbf{\Sigma}^{-1} \mathbf{g} \quad (3-26)$$

where,

$$\mathbf{\Sigma} = \mathbf{C}_M + \mathbf{C}_D \quad (3-27)$$

The estimate for the ERP component is then:

$$\mathbf{y}_o = \mathbf{X}^T \mathbf{w}_o \quad (3-28)$$

We point out that (3-25) suggests that in the Gaussian assumption, observational uncertainties and model uncertainties simply combine by addition of their respective covariance matrices. We also note that if $\mathbf{C}_M$ and $\mathbf{C}_D$ are both chosen as identity matrix (i.e., incorporating the least amount of *a priori* information into the model), then the solution in (3-26) essentially reduces to (3-8) in the deterministic approach. The optimality of the spatial filter will depend on how we choose the two covariance matrices.

The estimation of the scalp topography and amplitude follow the same procedures as described in (3-15) – (3-20).

3.4 Simulation Study

We present in this section a simulation study with synthetic data and real EEG data recorded from subjects during a passive picture-viewing experiment. The goal of the simulation study is to evaluate the latency and amplitude precision of synthetically generated transients immersed in real EEG background with different SNRs and waveform mismatching conditions.

3.4.1. Gamma Function as a Template for ERP Component

Lange *et al.* (1996) have used a Gaussian function as the template for an ERP component. Here, we prefer the Gamma function for the shape of the synthetic ERP component because this is a very flexible function for waveform modeling and has been used extensively in neurophysiological modeling (Koch et al., 1983; Patterson et al., 1992). Freeman(1975) argued that the macroscopic EEG electrical field is created from spike trains by a nonlinear generator with a second-order linear component with real poles. According to this model, the impulse response of the system is a monophasic waveform with a single mode, where the rising time depends on the relative magnitude of the two real poles (Appendix B). This may be approximated by a Gamma function, which is expressed by:

$$g(t) = c \cdot t^{k-1} \exp(-t / \theta), \quad t > 0 \quad (3-29)$$

where, $k > 0$ is a shape parameter, $\theta > 0$ is a scale parameter and c is a normalizing constant. The Gamma function is a monophasic waveform with the mode at $t = (k - 1)\theta, (k > 1)$. It has a short rise time and a longer tail for small k , and approximates a symmetric waveform for large k . This makes it a good candidate for modeling both early and late ERP components that tend to be symmetric in the early components and have longer tails in the late components. Figure 3-1 shows four Gamma functions with different shape and scale parameters.

3.4.2. Generation of Simulated ERP Data

EEG data were recorded from subjects during a passive picture-viewing experiment, consisting of 12 alternating phases: the habituation phase and mixed phase. Each phase has 30 trials. During the 30 trials of the habituation phase, the same picture was repeatedly presented 30 times. During the mixed phase, the 30 pictures are all different. Each trial lasts 1600 ms, and there is 600 ms pre-stimulus, and 1000 ms post-stimulus.

The scalp electrodes were placed according to the 128-channel Geodesic Sensor Nets standards. All 128 channels were referred to channel Cz and were digitally sampled for analysis at 250Hz. A bandpass filter between 0.01Hz and 40Hz was applied to all channels, which were then converted to average reference. Ocular artifacts (eye movement) were corrected with EOG recordings.

The scalp topography of the synthetic ERP component is chosen as the normalized P300 scalp topography from another study (Li, et al., 2006). The simulation data were created by taking the superposition of the 600ms (150 samples) prestimulus data from 120 trials as the background EEG data and the scalp projected Gamma waveform as a proxy for the ERP component.

The SNR levels given the background EEG data can be easily adjusted by modifying the normalizing constant c in (17). We define the SNR given the single-trial EEG data as:

$$SNR = 20 \log \left(\frac{\sigma_s}{\sqrt{Tr(\mathbf{X}\mathbf{X}^T)/T}} \right) \quad (3-30)$$

Note that since the ‘actual’ ERP component is a nonstationary signal, the magnitude σ_s is taken be its *peak* amplitude. This is different from the conventional definition of SNR. The SNR levels given the background EEG data can be easily adjusted by modifying the normalizing constant c in (3-29).

3.4.3 Case Study I: Comparison with Other Methods

We will test the performance of the proposed spatiotemporal filtering method at varying SNR levels (from -20dB to 12dB), where we have access to the ‘actual’ (synthetic) ERP component. Two scenarios will be investigated: one where there is an exact match between the synthetic and the ERP component template and the other where there is a mismatch between the

two components. For the case of exact match, we use the parameters $k = 3$, $\theta = 13$ for both ERP components. For the mismatch case, the synthetic ERP component remains the same, but the template has a different waveform with parameters $k = 5$, $\theta = 6$. Figure 3-2(a) shows the waveforms of the two components for the mismatch case. We fixed the peak latency of the synthetic component at 200ms for all the 120 trials.

For Woody's filter, we selected channel Pz for analysis, and use the initial ensemble average as the template, avoiding the iterative update on the template (since the true latency is fixed, this is the best-case scenario for Woody's filter). We search around the true latency within 100ms for maximum correlation. We estimate the peak amplitude by taking the average of the peak value and its two adjacent values (corresponding to a noncausal low pass filter with cut-off frequency of 12Hz). For spatial PCA, we select the eigenvector which has the maximal correlation with the P300 scalp topography (Note this is an ideal case for PCA, since in reality we do not know exactly the true scalp topography, nor the exact time course).

The simulation results are summarized in Table 3-1, 3-2 and 3-3, which show the estimation mean and standard deviation for the estimated latency and the ratio between estimated and true peak amplitude, as well as the correlation coefficient between estimated and true scalp topography. Since PCA does not give an explicit estimate for latency and amplitude, we will omit the latency estimation and only compute the amplitude ratio between the estimated ERP component and the synthetic ERP component.

First, we note that the single-trial estimation of the peak latency is very stable in the case of exact match. Notably, for EEG data with SNR higher than 4dB, the method estimates the latency correctly for all the trials. Second, the amplitude estimate for the exact case is also stable but is more variable than the latency estimation. The mean amplitude approaches to one

and the standard deviation decreases to zero as the SNR increases. We may say that in the case of exact match between the model and the component, the estimator for the amplitude is *asymptotically unbiased* and *asymptotically consistent* with increasing SNR.

The mismatch between the model and the generated component effectively introduces a bias in the estimation of the latency for realistic SNRs. From Table 3-1, we can see that the mean latency approaches to 188ms, yielding a bias of around 3 samples and the standard deviation is around 9ms (around 2 samples). However, the mean does converge to its true value (200ms) and the variance does approach to zero with increasing SNR, although very slowly. For instance, for SNR as high as 40dB and 60dB, the estimated latency has a mean-std statistic of $193 \pm 4\text{ms}$ and $200 \pm 0.7\text{ms}$, respectively. Therefore, empirically we can see that the estimation for latency under the mismatch case is also *asymptotically unbiased* and *asymptotically consistent* with increasing SNR. The mismatch of components introduced a bias in the estimation of the amplitude for realistic SNRs, which is partly due to the difference in the waveforms of the synthetic component and template. The estimated amplitude has a statistic of 1.0458 ± 0.0075 and 1.0015 ± 0.0002 at a SNR of 40dB and 60dB, respectively. Thus in the case of mismatch, the estimator for the amplitude is also *asymptotically unbiased* and *asymptotically consistent* with respect to SNR, although the convergence is much slower than the exact match case. Of course, the estimation variance does increase notably as SNR decreases. But as evident from Table 4-2, our method at -12dB still gives a estimation variance smaller than Woody's method at 0dB.

Table 3-1 and 3-2 clearly show the advantage of using spatial information, in contrast to the Woody filter based on single-channel analysis. Specifically, the estimation variance of Woody filter for both latency and amplitude are much larger for realistic (negative) SNR conditions.

PCA overestimates the amplitude of the ERP component for low SNR data. In contrast to our method, PCA gives a statistically significant bias even at 0dB from the baseline at 12dB (p -value less than 0.0001). This means that varying SNR (below 0dB) imposes a serious problem on the application of PCA in low SNR conditions.

Finally, the simulated mismatch of components affects the estimation of the scalp topography negligibly for the proposed method when SNR is higher than -10dB. In fact, the estimation for scalp topography with mismatch with our method at -20 dB is comparable to PCA at -4 dB.

3.4.4 Case Study II: Effects of Mismatch

The second simulation concerns the effects of the mismatch on the estimation, specifically mismatch in the spread parameter which is the most important. We use the same synthetic component as before and vary the spread parameter with K fixed. Fig. 3-2(b) shows the waveforms of the synthetic and 3 of the templates, including a Gaussian with a spread of 20. The results are summarized in Table 3-4 and 3-5, for SNR = -20dB and -10dB, respectively. We have included a new quantification for the amplitude estimation: coefficient of variation (CV), which is defined as the ratio of the standard deviation to the mean of a random variable. It is used as a measure of dispersion of the estimated amplitude (since its true mean is fixed at 1).

From Table 3-4 and 3-5, we can see that at the same SNR level, the mean and variance of the estimated amplitude systematically change with respect to the spread parameter, i.e., larger spread parameter gives smaller amplitude. The degree of variability in the amplitude estimation (measured by coefficients of variation) for mismatch cases exceeds the exact match case by less than 10% except at -10dB for a spread parameter of 7. In some cases, CV is even smaller than the exact match case, which gives a better estimate, but *only* in terms of the amplitude. This is important because although the estimated amplitudes differ in the mismatch case, as long as we

use the same template, these amplitudes on average will always be magnified or attenuated by a constant factor at a certain SNR level. Intuitively, this means that given a fixed template and varying SNR ($>-20\text{dB}$), the dominant source of variability in amplitude estimation mainly comes from the estimation variance (not bias) and this variability (measured by CV) is well-bounded from above for a range of spread parameters. Therefore, these amplitude estimates may still be effectively compared across experimental conditions as long as the same (meaningful) template is used and the SNR of the ERP data does not fall below -20dB . However, the mismatch clearly introduces a bias in the latency estimation, which may be as large as 50ms in absolute terms. This may or may not be significant depending on the applications. We can also see from the two tables that choosing a higher spread parameter will lead to a slightly better estimation for the scalp topography. But of course, this comes at the cost of much worse latency estimation.

Table 3-1. Latency estimation: mean and standard deviation

SNR (dB)	Woody	Exact match	Mismatch
-20	191 $\pm$ 60	202 $\pm$ 10	190 $\pm$ 16
-16	190 $\pm$ 60	201 $\pm$ 7	189 $\pm$ 14
-12	190 $\pm$ 61	201 $\pm$ 5	188 $\pm$ 11
-8	191 $\pm$ 63	200 $\pm$ 3	188 $\pm$ 10
-4	195 $\pm$ 60	200 $\pm$ 2	188 $\pm$ 10
0	195 $\pm$ 54	200 $\pm$ 1	187 $\pm$ 10
4	199 $\pm$ 45	200 $\pm$ 0	187 $\pm$ 10
8	199 $\pm$ 32	200 $\pm$ 0	188 $\pm$ 10
12	200 $\pm$ 15	200 $\pm$ 0	188 $\pm$ 9

True latency: 200ms.

Table 3-2. Amplitude estimation: mean and standard deviation

SNR (dB)	Woody	PCA	Exact match	Mismatch
-20	-0.22 $\pm$ 9.32	10.7 $\pm$ 2.12	1.68 $\pm$ 2.20	2.22 $\pm$ 2.93
-16	0.07 $\pm$ 5.95	6.80 $\pm$ 1.35	1.34 $\pm$ 1.29	1.71 $\pm$ 1.75
-12	0.40 $\pm$ 3.84	4.33 $\pm$ 0.86	1.18 $\pm$ 0.77	1.49 $\pm$ 1.05
-8	0.36 $\pm$ 2.60	2.78 $\pm$ 0.55	1.10 $\pm$ 0.47	1.34 $\pm$ 0.64
-4	0.53 $\pm$ 1.75	1.86 $\pm$ 0.36	1.06 $\pm$ 0.29	1.26 $\pm$ 0.40
0	0.64 $\pm$ 1.14	1.35 $\pm$ 0.23	1.04 $\pm$ 0.18	1.23 $\pm$ 0.25
4	0.80 $\pm$ 0.74	1.13 $\pm$ 0.13	1.02 $\pm$ 0.12	1.20 $\pm$ 0.16
8	0.90 $\pm$ 0.45	1.05 $\pm$ 0.08	1.01 $\pm$ 0.07	1.19 $\pm$ 0.10
12	0.98 $\pm$ 0.24	1.02 $\pm$ 0.05	1.01 $\pm$ 0.05	1.18 $\pm$ 0.07

True amplitude: 1

Table 3-3. Scalp topography estimation: correlation coefficient

SNR (dB)	PCA	Exact match	Mismatch
-20	0.568	0.829	0.759
-16	0.579	0.910	0.854
-12	0.601	0.959	0.926
-8	0.649	0.982	0.966
-4	0.759	0.993	0.986
0	0.906	0.997	0.994
4	0.979	0.999	0.998
8	0.996	1.000	0.999
12	1.000	1.000	1.000

Table 3-4. Effects of mismatch I: SNR = -20dB

Spread parameter	Latency(ms)	Amplitude	Coefficient of Variation	CC. of scalp topography
7	150 ± 20	2.32 ± 3.02	1.30	0.68
9	175 ± 16	2.18 ± 2.74	1.25	0.73
11	191 ± 12	1.93 ± 2.51	1.30	0.79
13	202 ± 10	1.68 ± 2.20	1.31	0.83
15	215 ± 11	1.43 ± 1.87	1.31	0.87
17	228 ± 15	1.24 ± 1.60	1.29	0.88
19	246 ± 18	1.07 ± 1.41	1.31	0.87
20 (Gaussian)	210 ± 22	2.14 ± 2.89	1.35	0.76

Table 3-5. Effects of mismatch II: SNR = -10dB

Spread parameter	Latency(ms)	Amplitude	Coefficient of Variation	CC. of scalp topography
7	151 ± 21	1.30 ± 0.92	0.71	0.93
9	175 ± 16	1.34 ± 0.78	0.58	0.95
11	190 ± 8	1.24 ± 0.69	0.55	0.96
13	200 ± 4	1.14 ± 0.60	0.53	0.97
15	211 ± 5	1.03 ± 0.53	0.51	0.98
17	220 ± 7	0.94 ± 0.46	0.49	0.98
19	229 ± 9	0.86 ± 0.41	0.47	0.99
20 (Gaussian)	208 ± 11	1.30 ± 0.78	0.60	0.95

Figure 3-1 Gamma functions with different shapes and scales.

A

B

Figure 3-2 Waveforms of synthetic and presumed ERP component. A) Synthetic component Gamma: $K = 3$, $\theta = 13$; presumed component: Gamma: $K = 5$, $\theta = 6$. B) Synthetic component Gamma: $K = 3$, $\theta = 13$; presumed components: Gamma: $K = 3$, $\theta = 7$, and $\theta = 19$, Gaussian with a spread of 20.

CHAPTER 4 ENHANCEMENTS TO THE BASIC METHOD

In chapter 3, we developed the basic spatiotemporal filtering method for single-trial ERP component estimation. In this chapter, we will consider some modifications to the basic method. Some serves as a heuristic post-processing technique (Section 4.1: iteratively refined template); some aims to deal with large salient EEG artifacts (Section 4.2: robust estimation); some utilizes the *a priori* knowledge on the scalp topography of the ERP component (Section 4.3: regularization); some provides alternative formulations for our previous results and also derives new ones (Section 4.4: Bayesian formulations of the topography estimation); still others try to deal with the interference from other overlapping ERP components (Section 4.5: explicit compensation for temporal overlap). The details are presented below.

4.1 Iteratively Refined Template

The deterministic approach of our estimation in Chapter 3 uses a fixed waveform for the template, regardless of the SNR. The stochastic approach incorporates some degree of variability in the template, but it is still implicit. We would like to explicitly utilize the posterior information from the data to update or refine our *a priori* assumed template. Intuitively, this should improve our estimation at least for high SNR conditions.

We use the estimated scalp topography $\mathbf{a}_o$ as a spatial filter. The output is optimal in the sense that it has the largest correlation coefficient with the actual component with the uncorrelated noise assumption. (Of course, we use the estimate as a proxy for the true topography. Note that it is different from $\mathbf{w}$). The refined template is the ensemble average of spatially filtered data, with $\mathbf{a}_o$ as the filter. The results are shown in Table 4-1, 4-2 and 4-3 (with one iteration of refining).

We can see that for the latency estimation, the refined template method has a larger bias below -12dB than the original template, but improves quickly and approaches to the exact match case for positive SNR conditions. For the amplitude estimation, the refined template consistently beats the original template for all SNR conditions in terms of both bias and variance. It also approaches to the exact match case for positive SNR conditions. Also see Figure 4-1. For the scalp projection estimation, the refined template is worse below -8dB but is very close to the exact match case above -8dB.

Figure 4-2 shows the waveforms of the synthetic component, original template and refined template at 4 SNR conditions: -20dB, -12dB, 0dB and 12dB. The spatially filtered ensemble average is still quite erratic below -10dB. That possibly accounts for the worse performance of refined template for very low SNR conditions. So, there exists a critical SNR below which, using the spatially filtered ensemble average will probably worsen performance (in terms of scalp topography, but not amplitude) compared with the original template. For this particular data set, it would be safe to use a refined template as long as the SNR is above -10dB.

With 1 iteration, there comes 2, 3 and infinity. The natural question then is: Will it converge? If so, what does it converge to? Theoretically, these are difficult questions. Aside from the variable latency parameter, the scalp topography is still nonlinearly related to the data. However, we can experimentally determine the limiting results. These are shown in Table 4-4 for 3 SNR conditions. The algorithm converges within 10 (sometimes 2) iterations. Compared with the first iteration, the estimation for latency and scalp projection barely changes, but there is a reduced bias and variance in terms of amplitude for negative SNR conditions.

Of course, all these results are a function of the mismatch in the waveforms and the number of trials we use to compute the ensemble average (if we had 1000 trials instead of 120, it would

be a different story). At this point, it seems that, what matters the most to deal with negative SNR is to accumulate more data, either in space or in time.

4.2 Regularization

Sometimes, we have some *a priori* knowledge of the scalp topography of the ERP component. For instance, the P300 component usually has a large positive projection around Pz area. In these cases, we should utilize that information and incorporate it into our model.

4.2.1 Constrained Optimization

Assuming that the ERP component latency has been estimated, we attempt to minimize the following cost function with respect to the amplitude σ and scalp topography $\mathbf{a}$ (which is constrained to have a unit norm in l_2):

$$\arg \min_{\sigma, \mathbf{a}} \left\| \mathbf{X} - \sigma \cdot \mathbf{a} \cdot \mathbf{s}^T \right\|_F^2 + \lambda \cdot \left\| \mathbf{a} - \mathbf{a}_0 \right\|_2^2 \quad (4-1)$$

$$S.t., \left\| \mathbf{a} \right\|_2 = 1$$

where, λ is a regularization parameter and $\mathbf{a}_0$ is a normalized vector representing the *a priori* knowledge of the scalp topography of the ERP component. $\left\| \cdot \right\|_F$ denotes the Frobenius norm of a matrix. The reason that we chose this norm will be evident later: the minimization of this norm gave the same solution as (3.14), which was derived with the uncorrelatedness assumption.

$$\text{In Appendix C we derive a fix-point update equation: } \begin{cases} \sigma = \frac{\mathbf{a}^T \mathbf{X} \mathbf{s}}{\mathbf{s}^T \mathbf{s}} \\ \mathbf{a} = \frac{\lambda \mathbf{a}_0 + \sigma \mathbf{X} \mathbf{s}}{\left\| \lambda \mathbf{a}_0 + \sigma \mathbf{X} \mathbf{s} \right\|_2} \end{cases} \quad (4-2)$$

Particularly, when $\lambda = 0$ (no regularization), the optimal solution becomes:

$$\begin{cases} \sigma = \frac{\|\mathbf{Xs}\|_2}{\mathbf{s}^T \mathbf{s}} \\ \mathbf{a} = \frac{\mathbf{Xs}}{\|\mathbf{Xs}\|_2} \end{cases} \quad (4-3)$$

When $\lambda \rightarrow \infty$, the optimal solution is:

$$\begin{cases} \sigma = \frac{\mathbf{a}_0^T \cdot \mathbf{Xs}}{\mathbf{s}^T \mathbf{s}} \\ \mathbf{a} = \mathbf{a}_0 \end{cases} \quad (4-4)$$

When λ takes intermediate values, the scalp topography is a weighted average of the two extreme case solutions.

This is a “real” single-trial estimation scheme in that the amplitude is computed from one single-trial data matrix. For the case of no regularization, it is different from our original formulation, where the amplitude is the inner product between $\mathbf{Xs}$ and the normalized *average* scalp topographies from all the trials. The original formulation makes the reasonable assumption that the ERP component has a fixed scalp topography and utilizes that information.

We demonstrate the effectiveness of regularization to deal with the interference from other overlapping (possibly unknown, if they are all known, we can explicitly compensate for that- See Section 4.5) ERP components. Specifically, we will investigate the effect of regularization on the estimation of amplitude and scalp topography under well controlled conditions.

We assume that there are 2 overlapping ERP components and their latencies are fixed and known. Their waveforms are shown in Figure 4-3. They are both Gamma functions with the same parameters $K = 3, \theta = 10$, with a peak interval of 80ms. The two ERP components have a correlation coefficient of 0.36. We use templates that are exactly matched with the synthetic components. These components are projected to spontaneous EEG data to generate simulated

ERP data. Their scalp topographies are shown in the Figure 4-4. We use that for $\mathbf{a}_0$ (exact *a priori* knowledge).

We then find the optimal solution of σ for $\mathbf{a}$ given λ . The fixed point update always converges within 2 steps. We summarize the results for 3 SNR conditions (12dB, -12dB, -20dB), shown in Figure 4-5. For high SNR (12dB) data, regularization brings little difference. Since the overlapping ERP components have the exact opposite scalp topography, it is expected that the estimated amplitude is smaller than 1 for high SNR data. It converges to around 0.42 and 0.41 for component 1 and 2 for large λ , respectively. Note the huge bias in the estimated amplitude for low SNR (especially -20dB) without regularization. But it converges to as small as 0.69 and 0.27 for component 1 and 2 for large λ , respectively. This demonstrates the necessity of regularization for the constrained optimization problem. A reasonable regularization parameter for *all* the SNR conditions is between 10^4 and 10^5 , where an unbiased estimation for amplitude could be achieved. Also notice that there is a hump for the standard deviation of the estimated amplitude. Interestingly, this is near the reflection point of the mean amplitude. In practice, this can give us some hint for finding a reasonable regularization parameter. Note that for these choices of λ , the correlation coefficient of scalp topography is already very close to 1.

This is an example when the overlapping components have negative correlation on the scalp. What about positive correlated components? Figure 4-6 shows the results for the same components, except that now they have the same scalp topography. The estimated amplitude is expected to be larger than 1 for high SNR data. It converges to 1.6 for both components for large λ . Note the huge bias for low SNR (especially -20dB). It converges to 1.8 and 1.7 for component 1 and 2 for large λ , respectively. This also demonstrates the benefits of regularization for the constrained optimization problem: using a sufficiently large regularization

parameter in this case can reduce the bias of the amplitude estimation, while the variance are not affected very much. As expected, for large λ , the correlation coefficient of scalp topography converges to 1.

4.2.2 Unconstrained Optimization

Parallel with the above constrained optimization, we can also frame the problem into an unconstrained optimization one.

$$\arg \min_{\mathbf{a}} \left\| \mathbf{X} - \mathbf{a} \cdot \mathbf{s}^T \right\|_F^2 + \lambda \cdot \left\| \mathbf{a} / \|\mathbf{a}\|_2 - \mathbf{a}_0 \right\|_2^2 \quad (4-5)$$

Note that the optimization variable $\mathbf{a}$ contains the amplitude parameter as well as the topography information. As shown in Appendix D, a fixed-pointed update can be obtained:

$$\mathbf{a} = \frac{\mathbf{X}\mathbf{s} + \lambda \mathbf{a}_0 / \|\mathbf{a}\|_2}{\mathbf{s}^T \mathbf{s} + \lambda \mathbf{a}_0^T \mathbf{a} / \|\mathbf{a}\|_2^3} \quad (4-6)$$

Particularly, when $\lambda = 0$ (no regularization), the optimal solution is the same as our original solution in (3-14). When $\lambda \rightarrow \infty$, the optimal solution is not unique (any scaled version of $\mathbf{a}_0$ can be a solution). The problem becomes ill-posed.

We have obtained the regularized solution for *single-trial* scalp topography. The same procedures for estimating the amplitude follow: take the average of normalized single-trial scalp topography as our estimate for the overall scalp topography, then the amplitude for a particular trial is just the inner product between this vector and the corresponding scalp topography.

We test the performance of regularization under the same conditions as in the constrained optimization. We find the optimal solution for $\mathbf{a}$ given λ . Fixed point update usually converges within 10 steps.

We summarize the results for 3 SNR conditions (12dB, -12dB, -20dB) shown in Figure 4-7 and 4-8. As in constrained optimization, for high SNR (12dB) data, regularization has little effect

on the estimation. The amplitude converges to around the same values as before (0.4 for opposite topography and 1.6 for the same topography). The variance is not affected much for all the SNR conditions, either. As expected, the scalp topography gets monotonically better as λ increases.

The difference is that the estimated amplitude mean increases monotonically (except for a small interval) with increasing regularization parameter λ . This translates to a larger bias (particularly for low SNR data) when the overlapping components have positively correlated topographies. The estimated amplitude becomes unstable for large λ . There is no evidence at this point that regularization can benefit the estimation for general scalp topography configurations (both positive and negative topography correlations). So the unconstrained optimization formulation need not be regularized, at least for overlapping components with positively correlated topographies. This lends support to our original solution in (3.14), which is exactly the unregularized solution to the unconstrained optimization problem here.

We also point out that, unlike the constrained optimization problem, here we utilize the reasonable assumption that an ERP component has a fixed scalp topography. So it does not suffer from the huge bias problem in the constrained optimization. For instance, at -20dB without regularization, the estimated mean amplitude is around 2.3 and 1.9 for the two components with positively correlated topographies, a modest increase from 1.6 at 12dB.

4.3 Robust Estimation: the CIM Metric

We have seen in Section 4.2 that the estimation for single-trial scalp topography in (3.14) can be found equivalently from the minimization the following criterion:

$$\arg \min_{\mathbf{a}_k} \left\| \mathbf{X}_k - \mathbf{a}_k \cdot \mathbf{y}_k^T \right\|_F^2 \quad (4-7)$$

where, $\|\cdot\|_F$ denotes the Frobenius norm.

It is evident from (3-14) that the estimate for single-trial scalp topography bears a linear relationship with the EEG data. Because of the noisy nature of EEG (particularly large salient artifacts), this gives a noisy estimate for the single-trial scalp topography (with large variance) and in turn translates into a noisy estimate for the single-trial amplitude in (3-18).

We would like to derive a robust estimator in order to reduce the effects of large EEG artifacts. We can replace the Frobenius norm in (4-7) with other norms (e.g., l_1 norm) or metrics. Here we will consider a special metric: correntropy induced metric (CIM) proposed by Liu *et al* (2007).

First we introduce what is called correntropy. Given two scalar random variables X and Y , correntropy is defined as:

$$V_h(X, Y) = E[k_h(X - Y)] \quad (4-8)$$

where, $k_h(X - Y)$ is the Gaussian kernel (h is the kernel size),

$$k_h(X - Y) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{(X - Y)^2}{2h^2}\right) \quad (4-9)$$

The correntropy function is a localized similarity measure in the joint probability space, which is controlled by the bandwidth parameter h (also called kernel size in kernel methods). It induces a metric (CIM) in the sample space which behaves like the l_2 norm when the sample point is close to the origin (relative to the kernel size); when the sample point gets further apart from the origin, the metric is similar to the l_1 norm and eventually saturates and approaches to the l_0 norm (Liu, 2007). As such, CIM practically incorporates the l_2 norm as a special case (if h is chosen to be sufficiently large).

Minimization of CIM is equivalent to the maximum correntropy criterion (MCC). It has been shown that MCC has a close relation to M-estimation (Huber, 1981) and since correntropy

is inherently insensitive to outliers, MCC is especially suitable for rejecting impulsive noise (Liu, 2007).

Now, treating each entry of the matrix in (4-7) as a realization of a random variable, we can write our new cost function as:

$$\arg \min_{\mathbf{a}_k} \left\| \mathbf{X}_k - \mathbf{a}_k \cdot \mathbf{y}_k^T \right\|_{CIM}^2 \quad (4-10)$$

The nuisance parameter h (kernel size) should be tuned to the data (most notably to the standard deviation). Here, we use the Silverman's rule as a baseline to quantify different values of kernel sizes that we use in the simulation. It is given by $h = 1.06dN^{-0.2}$, where N is the number of samples, and d the standard deviation of the data (Silverman, 1986).

While minimization of the Forbenius norm has a closed-form solution, minimization of CIM does not. So we have to search for a local minimum, using the standard gradient descent method. The convergence to a certain local minimum is guaranteed by adopting a stopping criterion that the change in the correlations of the estimate and the MSE solution between the previous and current iteration is less than 10^{-6} .

Table 4-5 and 4-6 summarize the estimation results for 2 SNR conditions (0dB and -20dB) for the mismatch case. We also include the MSE solution for comparison. The results are mixed: MCC gives a slightly higher variance than MSE for the estimation of amplitude, but the bias is marginally reduced; it also gives a more accurate estimate for the scalp topography. We also notice that at 0dB, the results of MCC barely change from MSE. Intuitively, when the SNR becomes sufficiently large, the optimal MCC solution should converge to the MSE solution (both agree with the true values of the parameters).

These differences in the results are by no means statistically significant. We venture 2 reasons why the MCC results do not change very much from MSE.

First, the EEG data are already preprocessed and relatively clean. Large artifacts have already been removed. The resulting distribution is not far from Gaussian. So MSE should give a solution already close to optimal. MCC has its edge when there is large noise, especially impulse noise. Strictly speaking, it is only optimal (in the sense of maximum likelihood) for one particular type of distribution, just as MSE is strictly optimal for Gaussian distribution. It is not clear that how these two criteria compare when the data distribution changes in a *neighborhood* of their optimal ones. In reality (when EEG data are usually preprocessed and artifacts are removed), there is no reason to believe that MCC will outperform MSE uniformly.

Another less compelling reason concerns the optimization process associated with MCC. The initial condition of MCC is set to be the MSE solution (starting a random initial condition will seldom beat MSE). When the kernel size is small, the performance surface is highly irregular, so the optimization will never go far from MSE solution (it is stuck around the local minimum near the MSE solution). When kernel size is sufficiently large, it is easy to see that MCC approaches to MSE. Only intermediate values of kernel size will produce somewhat different results from MSE. This is seen in both SNR levels, though less evident for 0dB.

There are two other cost functions in the estimation of amplitude that use the MSE criterion, i.e., (3.14) and (3.16), which can also be replaced by MCC. Li *et al* (2007) have tested its performance and the improvement was shown to be marginal.

In practice, one has to weigh the small improvement in the performance of MCC against its high computational cost (and no guarantee of convergence to global optima).

4.4 Bayesian Formulations of the Topography Estimation

The amplitude estimation consists of 3 steps. First we estimate the single-trial scalp topography $\mathbf{a}_k$ (either by the uncorrelated assumption or equivalently through the minimization

of the Frobenius norm of $\mathbf{X}_k - \mathbf{a}_k \mathbf{y}_k$). Then we compute the normalized scalp topography $\mathbf{a}_o$ as the normalized version of a weighted average of the single-trial scalp topography $\mathbf{a}_k$ (the weights being their respective l_2 norm). The third step is the minimization of the l_2 norm $\|\mathbf{a}_k - \sigma_k \mathbf{a}_o\|_2$, which gives the optimal single-trial amplitude as the inner product between $\mathbf{a}_k$ and $\mathbf{a}_o$.

We have seen that the first step gives a near optimal solution (as opposed to CIM) if the EEG data have been cleaned. We have also shown that in the third step, using other metrics (e.g., CIM) gives a marginally better results than using the l_2 norm. Here we will investigate other alternatives to the estimation of the normalized scalp topography $\mathbf{a}_o$ in the second step in (3.17).

We maintain that after estimating the single-trial scalp topography in the first step, we treat $\mathbf{a}_k$ as known and given. Then we ask the question: what is the best estimate for the normalized scalp topography $\mathbf{a}_o$ given $\mathbf{a}_k$ for K trials. This is a divide-and-conquer approach and simplifies matters.

Naturally the problem is best formulated in a Bayesian framework. Of course, the formulation will depend on the model we assume for the data. Next, we will present three different models and compare their performance.

4.4.1 Model 1: Additive Noise Model

$$\mathbf{a}_k = \sigma_k \mathbf{a}_o + \mathbf{u}_k \quad (4-11)$$

where, σ_k is the single-trial amplitude, and $\mathbf{u}_k$ is the error (model uncertainty).

This model is consistent with our linear generative EEG model. We shall assume conditional independence between $\mathbf{a}_i | \sigma_i, \mathbf{a}_o$ and $\mathbf{a}_j | \sigma_j, \mathbf{a}_o$, for any $i \neq j$. Of course, we also assume σ_i and

σ_j are independent for any $i \neq j$, and they are all independent of the normalized scalp topography $\mathbf{a}_o$.

We wish to maximize *a posteriori* probability (MAP):

$$\arg \max_{\mathbf{a}_o, \sigma_1 \dots \sigma_K} p(\mathbf{a}_o, \sigma_1 \dots \sigma_K \mid \mathbf{a}_1 \dots \mathbf{a}_K) \quad (4-12)$$

It can be shown (Appendix E) that the MAP solution occurs when $\mathbf{a}_o$ is the normalized eigenvector of the matrix $\mathbf{A}$ corresponding to the largest eigenvalue, where,

$$\mathbf{A} = \sum_{k=1}^K \mathbf{A}_k = \sum_{k=1}^K \mathbf{a}_k \cdot \mathbf{a}_k^T. \quad (4-13)$$

Now we have solved the MAP problem under model 1. But, there is a weakness: the model error has a constant covariance across all trials. This is a dubious assumption. Intuitively, when the data are noisy, the variance in $\mathbf{u}_k$ will increase accordingly. We should somehow “normalize” the data in our model. This leads to model 2.

4.4.2 Model 2: Normalized Additive Noise Model

$$\frac{\mathbf{a}_k}{\sigma_k} = \mathbf{a}_o + \mathbf{u}_k \quad (4-14)$$

The left hand side is in fact a proxy for $\mathbf{a}_{ok}$ in our linear generative EEG model, except that it may not be normalized. $\mathbf{u}_k$ is again zero-mean *i.i.d.* Gaussian noise, but now with unit norm.

The model can also be written as: $\mathbf{a}_k = \sigma_k \cdot \mathbf{a}_{ok} + \sigma_k \cdot \mathbf{u}_k$.

Again, it can be shown (Appendix F) that the MAP solution occurs when $\mathbf{a}_o$ is the normalized eigenvector of the matrix $\mathbf{B}$ corresponding to the largest eigenvalue, where,

$$\mathbf{B} = \sum_{k=1}^K \mathbf{B}_k = \sum_{k=1}^K \frac{\mathbf{a}_k \cdot \mathbf{a}_k^T}{\mathbf{a}_k^T \cdot \mathbf{a}_k}. \quad (4-15)$$

Note that the components $\mathbf{B}_k$ is a “normalized” version of $\mathbf{A}_k$ in that the trace of $\mathbf{B}_k$ is always 1. We also note that, we can treat σ_k only as a normalizing factor, not necessarily as the single-trial “amplitude”. In this case, we separate the estimation of σ_k and $\mathbf{a}_o$. As before, we can compute the amplitude by: $\sigma_k = \mathbf{a}_o^T \cdot \mathbf{a}_k$.

4.4.3 Model 3: Original Model

Our last model is actually simpler. We do not include the unknown amplitude in this stage and only consider the estimation of $\mathbf{a}_o$. The posteriori probability is simply

$$\arg \max_{\mathbf{a}_o} p(\mathbf{a}_o | \mathbf{a}_1 \dots \mathbf{a}_K).$$

Given the model: $\frac{\mathbf{a}_k}{\|\mathbf{a}_k\|_2} = \mathbf{a}_o + \mathbf{u}_k$, Appendix G shows that the MAP solution actually

coincides with our original solution in (3.17). The single-trial amplitude estimation is the same as before.

4.4.4 Comparison among the Three Models

We compare the amplitude and scalp topography estimation for the three models. The results are summarized in Table 4-7 and 4-8. We can see that model 3 (original solution) consistently gave the best results among the three under different SNR conditions. The topography estimation with model 1 is poor at negative SNR conditions. Model 2 is an obvious improvement. Note that its amplitude estimation has a huge bias and variance for low SNR data. This may be due to the improper prior we assign to σ_k (it assign large probability to large values) in the derivation of the MAP estimator. We also included the amplitude estimation with the traditional inner product. There is a significant improvement in both the bias and variance, especially in low SNR conditions.

4.4.5 Online Estimation

Sometimes we wish to know the single-trial parameters after recording each trial. It is then necessary to obtain an on-line estimation method. We assume that the stimulus onset time is known to us. Again, the problem is best formulated in a Bayesian framework in order to utilize all the information in previous trials. In fact, it is trivial given the above analysis. Here, we adopt model 3.

After the first trial, set: $\mathbf{a}_o^{(1)} = \mathbf{a}_1 / \|\mathbf{a}_1\|_2$, $\sigma_1 = \|\mathbf{a}_1\|_2$

At each trial, we store the running average: $\mathbf{c}_k = \sum_{k=1}^K \mathbf{a}_k / \|\mathbf{a}_k\|_2$.

When finishing recording trial $K+1$, we update the topography estimate:

$$\mathbf{a}_o^{(K+1)} = \frac{\mathbf{c}_K + \mathbf{a}_{K+1} / \|\mathbf{a}_{K+1}\|_2}{\|\mathbf{c}_K + \mathbf{a}_{K+1} / \|\mathbf{a}_{K+1}\|_2\|_2},$$

and the amplitude for the newly recorded trial: $\sigma_{K+1} = \mathbf{a}_{K+1}^T \cdot \mathbf{a}_o^{(K+1)}$

4.5 Explicit Compensation for Temporal Overlap of Components

In developing our basic method in Chapter 3, we assumed that ERP components are uncorrelated with each other. In reality, this is seldom satisfied. Because ERP components have relatively stable waveforms and latency, when they overlap in time, there will generally be a nonzero correlation among them. Here we are mainly concerned with the situation where ERP components overlap in time, but latency jitter across trials is relatively small.

Consider two ERP components overlapping in time (it is easy to generalize to multiple components). We assume that the time courses of the components $\mathbf{s}_1$ and $\mathbf{s}_2$ are known from physiological knowledge. We also assume that the latency is given or can be estimated, e.g., from ensemble average, and it is relatively fixed. For the purpose of amplitude estimation,

latency jitter is considered here as a minor issue compared with the possibly heavy overlap of components. Note that, in the case of two overlapping (correlated) components, we lose the ability to estimate the latency simply from the cost function in (3-9).

If we assume that all other components are uncorrelated with these two components (or have negligible temporal overlap with them), then we can compensate for the correlation (due to temporal overlap) to get an unbiased estimate for both components' amplitudes. We write the linear generative EEG model in this case:

$$\mathbf{X} = \mathbf{a}_1 \cdot \mathbf{s}_1^T + \mathbf{a}_2 \cdot \mathbf{s}_2^T + \sum_{i=1}^N \mathbf{b}_i \mathbf{n}_i^T \quad (4-16)$$

We wish to estimate the scalp topographies for the two overlapping components. With the uncorrelated assumption, we can get two set of equations:

$$\begin{cases} \mathbf{a}_1 \cdot \mathbf{s}_1^T \cdot \mathbf{s}_1 + \mathbf{a}_2 \cdot \mathbf{s}_2^T \cdot \mathbf{s}_1 = \mathbf{X} \cdot \mathbf{s}_1 \\ \mathbf{a}_1 \cdot \mathbf{s}_1^T \cdot \mathbf{s}_2 + \mathbf{a}_2 \cdot \mathbf{s}_2^T \cdot \mathbf{s}_2 = \mathbf{X} \cdot \mathbf{s}_2 \end{cases} \quad (4-17)$$

Solve for $\mathbf{a}_1$ and $\mathbf{a}_2$:

$$\begin{cases} \mathbf{a}_1 = \frac{D_2 \mathbf{X} \cdot \mathbf{s}_1 - C \mathbf{X} \cdot \mathbf{s}_2}{D_1 D_2 - C^2} \\ \mathbf{a}_2 = \frac{D_1 \mathbf{X} \cdot \mathbf{s}_2 - C \mathbf{X} \cdot \mathbf{s}_1}{D_1 D_2 - C^2} \end{cases} \quad (4-18)$$

where, $D_i = \mathbf{s}_i^T \cdot \mathbf{s}_i, (i=1,2)$, and $C = \mathbf{s}_1^T \cdot \mathbf{s}_2 = \mathbf{s}_2^T \cdot \mathbf{s}_1$.

If $C = 0$, we have the same solution as before:

$$\mathbf{a}_i = \frac{\mathbf{X} \cdot \mathbf{s}_i}{\mathbf{s}_i^T \cdot \mathbf{s}_i}, i=1,2 \quad (4-19)$$

The procedures for estimating the amplitude are the same as before. The above technique assumes that accurate estimates of the waveforms of the components are available, since the cross-correlation in (4-18) depends on the tails of the overlapping components. This restricts its

applicability in practice. But if the researchers believe that the ERP components are heavily overlapped and are fairly certain of their waveforms, this technique should serve as a first attempt to reduce the bias in the estimation.

Table 4-1. Latency estimation: mean and standard deviation

SNR (dB)	Exact match	Mismatch	Refined template
-20	202 $\pm$ 10	190 $\pm$ 16	182 $\pm$ 12
-16	201 $\pm$ 7	189 $\pm$ 14	183 $\pm$ 13
-12	201 $\pm$ 5	188 $\pm$ 11	187 $\pm$ 12
-8	200 $\pm$ 3	188 $\pm$ 10	191 $\pm$ 12
-4	200 $\pm$ 2	188 $\pm$ 10	197 $\pm$ 6
0	200 $\pm$ 1	187 $\pm$ 10	199 $\pm$ 4
4	200 $\pm$ 0	187 $\pm$ 10	200 $\pm$ 1
8	200 $\pm$ 0	188 $\pm$ 10	200 $\pm$ 1
12	200 $\pm$ 0	188 $\pm$ 9	200 $\pm$ 1

True latency: 200ms.

Table 4-2. Amplitude estimation: mean and standard deviation

SNR (dB)	Exact match	Mismatch	Refined template
-20	1.68 $\pm$ 2.20	2.22 $\pm$ 2.93	2.17 $\pm$ 2.73
-16	1.34 $\pm$ 1.29	1.71 $\pm$ 1.75	1.61 $\pm$ 1.76
-12	1.18 $\pm$ 0.77	1.49 $\pm$ 1.05	1.33 $\pm$ 1.05
-8	1.10 $\pm$ 0.47	1.34 $\pm$ 0.64	1.18 $\pm$ 0.59
-4	1.06 $\pm$ 0.29	1.26 $\pm$ 0.40	1.11 $\pm$ 0.34
0	1.04 $\pm$ 0.18	1.23 $\pm$ 0.25	1.06 $\pm$ 0.20
4	1.02 $\pm$ 0.12	1.20 $\pm$ 0.16	1.04 $\pm$ 0.12
8	1.01 $\pm$ 0.07	1.19 $\pm$ 0.10	1.02 $\pm$ 0.07
12	1.01 $\pm$ 0.05	1.18 $\pm$ 0.07	1.01 $\pm$ 0.05

True amplitude: 1

Table 4-3. Scalp topography estimation: correlation coefficient

SNR (dB)	Exact match	Mismatch	Refined template
-20	0.829	0.759	0.644
-16	0.910	0.854	0.772
-12	0.959	0.926	0.891
-8	0.982	0.966	0.963
-4	0.993	0.986	0.987
0	0.997	0.994	0.995
4	0.999	0.998	0.998
8	1.000	0.999	0.999
12	1.000	1.000	1.000

Table 4-4. Estimation results for the iteratively refined template method

SNR (dB)	Latency estimation		Amplitude estimation		Scalp topography	
	Refined template	Iterative Refined	Refined template	Iterative Refined	Refined template	Iterative Refined
-20	182 ± 12	183 ± 12	2.17 ± 2.73	1.83 ± 2.36	0.644	0.644
-12	187 ± 12	187 ± 12	1.33 ± 1.05	1.23 ± 1.01	0.891	0.879
0	199 ± 4	199 ± 2	1.06 ± 0.20	1.06 ± 0.20	0.995	0.995

Table 4-5. Estimation with MCC for the mismatch case at SNR = 0dB

Kernel size (multiples of h)	Amplitude	Correlation coefficient of Scalp projection
0.5	1.22 ± 0.25	0.994
1	1.22 ± 0.26	0.994
2	1.22 ± 0.27	0.995
5	1.22 ± 0.26	0.995
10	1.22 ± 0.25	0.994
20	1.23 ± 0.25	0.994
MSE	1.23 ± 0.25	0.994

Table 4-6. Estimation with MCC for the mismatch case at SNR = -20dB

Kernel size (multiples of h)	Amplitude	Correlation coefficient of Scalp projection
0.5	2.22 ± 2.95	0.758
1	2.21 ± 3.05	0.754
2	2.17 ± 3.18	0.786
5	2.18 ± 3.14	0.789
10	2.21 ± 3.00	0.770
20	2.22 ± 2.93	0.759
MSE	2.22 ± 2.94	0.759

Table 4-7. Amplitude estimation for three Bayesian models

SNR (dB)	Model 1	Model 2	Model 2 (2)	Model 3 (original)
-20	0.38 ± 1.30	12.2 ± 378	1.95 ± 3.79	2.22 ± 2.93
-12	0.86 ± 0.85	9.96 ± 60	1.42 ± 1.34	1.49 ± 1.05
0	1.20 ± 0.26	1.66 ± 0.26	1.23 ± 0.25	1.23 ± 0.25
12	1.18 ± 0.07	1.21 ± 0.07	1.18 ± 0.07	1.18 ± 0.07

Table 4-8. Scalp topography estimation for three Bayesian models

SNR (dB)	Model 1	Model 2	Model 3 (original)
-20	0.188	0.642	0.759
-12	0.558	0.841	0.926
0	0.971	0.994	0.994
12	0.998	1.000	1.000

Figure 4-1 Mean and standard deviation of the estimated amplitude under different SNR conditions. The refined template method approaches to the exact match case for high SNR conditions.

Figure 4-2 The waveforms of the synthetic component, presumed template and refined template under 4 SNR conditions. A) -20dB. B) -12dB. C) 0dB. D) 12dB. The refined template appears erratic for low SNR and approaches to the synthetic component for high SNR.

Figure 4-3 Waveforms of two overlapped components used in regularization. The correlation coefficient between the two waveforms is around 0.36.

Figure 4-4 Scalp topography of two overlapped ERP components used in regularization.

A

Figure 4-5 Amplitude and scalp topography estimation I with regularization (constrained optimization) under 3 SNR conditions. A) 12dB. B) -12dB. C) -20dB. The overlapping ERP components have the exact opposite scalp topography. So the estimated amplitude is smaller than 1 for high SNR data. It converges to 0.42 and 0.41 for component 1 and 2 for large λ , respectively. Notice the huge bias for low SNR (especially -20dB) without regularization. But it converges to as small as 0.69 and 0.27 for component 1 and 2 for large λ , respectively.

B

C

Figure 4-5 Continued

A

Figure 4-6 Amplitude and scalp topography estimation II with regularization (constrained optimization) under 3 SNR conditions. A) 12dB. B) -12dB. C) -20dB. The overlapping ERP components have the same scalp topography. So the estimated amplitude is larger than 1 for high SNR data. It converges to 1.6 for both components for large λ . Notice the huge bias for low SNR (especially -20dB) without regularization. It converges to 1.8 and 1.7 for component 1 and 2 for large λ , respectively.

B

C

Figure 4-6 Continued

A

Figure 4-7 Amplitude and scalp topography estimation I with regularization (unconstrained optimization) under 3 SNR conditions. A) 12dB. B) -12dB. C) -20dB. The overlapping ERP components have the exact opposite scalp topography. So the estimated amplitude is smaller than 1 for high SNR data. The estimated amplitude mean generally increases with increasing λ while the variance is not affected much. It becomes unstable for large λ .

B

C

Figure 4-7 Continued

A

Figure 4-8 Amplitude and scalp topography estimation II with regularization (unconstrained optimization) under 3 SNR conditions. A) 12dB. B) -12dB. C) -20dB. The overlapping ERP components have the same scalp topography. So the estimated amplitude is larger than 1 for high SNR data. The estimated amplitude mean generally increases with increasing λ while the variance is not affected much. It becomes unstable for large λ .

B

C

Figure 4-8 Continued

CHAPTER 5

APPLICATIONS TO COGNITIVE ERP DATA

In this chapter, we apply the spatiotemporal filtering method proposed in Chapter 3 to the single-trial ERP estimation problem in two different experiments. The first application is an oddball target detection task with different pictures as stimuli, where the difficulty of the task or saliency of the stimuli leads to decreased P300 amplitude. The second one is the habituation study where the subjects were repeatedly presented identical pictures and the amplitude of certain ERP components is expected to decrease rapidly with respect to trials.

5.1 Oddball Target Detection

5.1.1 Materials and Methods

Because we were interested in single-trial, single-subject analyses of amplitude and latency, we selected 4 participants that met a minimum signal-to-noise ratios based on their averaged ERPs, from a pilot study ($n=8$) on implicit content processing during feature selection. They were right-handed according to the Edinburgh Handedness Questionnaire and all had normal or corrected vision.

Stimuli consisted of pictures from the International Affective Picture System, depicting adventure scenes, emotionally neutral social interactions, erotica, attack scenes, and mutilations. Their color content was manipulated such that they contained only shades of green or shades of red, and for each, color brightness was systematically manipulated to yield one bright and one dim version (Figure 5-1). All pictures were presented for 200 ms on the center of a 21-inch monitor, situated 1.5 m in front of the subjects. From this viewing distance the checkerboards subtended 4.0 deg. x 4.0 deg. of visual angle. A fixation cross was always present, even when no picture was presented on the screen. Target stimuli ($p = 0.25$) were defined for each experimental block (see below) by a combination of color and brightness.. All pictures were presented in

randomized order, with an inter-stimulus-interval varying randomly between 1000 to 1500 ms in 4 blocks of 120 trials each. One block lasted 7 min. on average. At the beginning of each block subjects were instructed to attend either to the bright/dark green or red pictures and to press the space bar of the computer keyboard when they detected a target. The target color and brightness were designated in counter-balanced order. Furthermore, the responding hand was changed half way through the experiment, and the sequence of hand usage was counterbalanced across subjects. Subjects were also instructed to avoid blinks and eye-movements and to maintain gaze onto the central fixation cross. Practice trials were provided for each subject for each condition to make sure that every subject had fully understood the task.

EEG was recorded continuously from 257 electrodes using an Electrical Geodesics™ (EGI) EEG system and digitized at a rate of 250 Hz, using Cz as a recording reference. Impedances were kept below 50 k Ω , as recommended for the Electrical Geodesics high input-impedance amplifiers. A subset of EGI net electrodes located at the outer canthi as well as above and below the right eye was used to determine horizontal and vertical Electrooculogram (EOG). All channels were preprocessed on-line by means of 0.1 Hz high-pass and 100 Hz low-pass filtering.

Epochs of 1000 ms (280 ms pre-, 720 ms post-stimulus) were obtained for each picture from the continuously recorded EEG, relative to picture onset. The mean voltage of a 120-msec segment preceding startle probe onset was subtracted as the baseline. In a first step, data were low-pass filtered at a frequency of 40 Hz (24 dB / octave) and then submitted to the procedure proposed by (Junghöfer *et al.*, 2000), which uses statistical parameters to exclude channels and trials that are contaminated with artifacts. This procedure resulted in rejection of trials that were contaminated with artifacts (including ocular artifacts). Artifacts were also evaluated by visual

inspection and respective trials were rejected. Recording artifacts were first detected using the recording reference (i.e. Cz). Subsequently, global artifacts were detected using the average reference and distinct sensors from particular trials were removed interactively, based on the distribution of their mean amplitude, standard deviation and maximum slope. Data at eliminated electrodes were replaced with a statistically weighted spherical spline interpolation from the full channel set. The mean number of approximated channels across conditions and subjects was 20. With respect to the spatial arrangement of the approximated sensors, it was ensured that the rejected sensors were not located within one region of the scalp, as this would make interpolation for this area invalid. Spherical spline interpolation was used throughout both for approximation of sensors and illustration of voltage maps (Junghöfer *et al.*, 1997).

Single epochs with excessive eye-movements and blinks or more than 30 channels containing artifacts in the time interval of interest were discarded. The validity of this procedure was further tested by visually inspecting the vertical and horizontal EOG as computed from a subset of the electrodes that were part of the electrode net. Subsequently, data were arithmetically transformed to the average reference, which was used for all analyses. After artifact correction an average of 69 % of the trials were retained in the analyses. The present analysis highlighted the most reliable signal available in this feature-based target identification task, which is the P300 component in response to a target stimulus (defined by a combination of color and brightness, irrespective of picture content). Thus, all subsequent analyses focused on amplitude and latency estimates for single trials belonging to the target condition.

5.1.2 Estimation Results

The present study illustrates the application of the method for a single late potential component. In reality, we do not know a priori how many ERP components there are in a single-

trial recording, nor do we know exactly when they occur. However, we may be able to estimate these values from single-trial EEG data in the data analysis session.

This is a good time to mention one technical requirement of our latency estimation. The single-trial latency is estimated from the cost function in (3.9), which involves the inversion of the matrix $\mathbf{XX}^T$. In reality, this matrix is usually ill-conditioned for dense-array EEG data (it will certainly be rank-deficient if there are any bad channels which were linearly interpolated from other channels.). This poses a computational problem in practice. Thus, the solution in (3.9) somehow has to be regularized. Here, we adopt a simple approach and add a regularization term $\lambda \mathbf{I}$ ($\lambda > 0$) to the matrix $\mathbf{XX}^T$ before taking the matrix inversion operation. The regularization parameter λ acted as a smoother to the cost function in (3.9). Generally, the solution is rather irregular without regularization, leading to too many local minima and spurious candidates for single-trial latencies due to large noise. With increasing λ , the cost function becomes smoother. This is clearly seen in Fig. 5-2, which shows the cost function in (3.9) for four different λ , for a particular trial from subject 2. With a smooth cost function, we can avoid the dilemma of choosing the right latency from too many candidates.

Now we have to select an appropriate value (or a meaningful range) for the regularization parameter λ . A good value for λ is one that achieves a balance between two extremes: too few and too many local minima. The idea is this: for a particular λ , we group all the candidates for single-trial latencies (time lags corresponding to local minima) together and perform 1D density estimation on these candidates. We count the number of modes (peaks) from the estimated probability density function (pdf). If this number is close to the average number of candidates for each trial, then the regularization parameter λ is at least *internally consistent*. Otherwise, it will contradict with itself and should not be used.

We illustrate our point using the results from one subject. Fig. 5-3 shows the estimated pdf of the candidates for single-trial latency from 200ms up to 600ms after stimulus onset when the regularization parameter λ equals 10^{-5} . We used the Parzen windowing pdf estimator (Parzen, 1962) with a Gaussian kernel size of 4.2. The kernel size was selected according to Silverman's rule (Silverman, 1986), which is given by $h = 1.06\sigma N^{-0.2}$, where N is the number of samples, and σ is the standard deviation of the data. The number of peaks depends on the kernel size, but we found that a kernel size between $0.5h$ and $2h$ will give the same number of peaks in the estimated pdf for this data. We can see that the pdf consists of 4 modes (peaks) after 200ms of the stimulus onset. There are 418 local minima and 102 trials in total, so the average number of local minima for each trial is about 4.1 (very close to the number of peaks in estimated pdf). This indicates that $\lambda = 10^{-5}$ gives an internally consistent estimate for latency.

We can repeat the above procedures for a wide range of regularization parameters and compute the ratio of the number of peaks in estimated pdf to the average number of local minima for each trial. For instance, the ratio was computed as around 4.75, 1.03, 0.96, 0.74 for the 4 regularization parameters in Fig. 5-2 respectively. Clearly, the first and last regularization parameter should not be used since they generate self-contradictory results. It is interesting to note that for a wide range of regularization parameters (from 10^{-5} to 10^0), the results are quite similar. This can also be seen from Fig.5-2, where both cost functions display 4 local minima and all time lags are near to their counterparts. For practical purposes, we can select any value from this range as a regularization parameter.

We were primarily interested in the P300 component, preferably the largest one. From the ensemble average, we know that the maximum ERP occurred around 380ms after stimulus onset. In Fig. 5-3, the estimated pdf displays a mode around 420ms. Thus, we searched around this

latency and set the single-trial peak latency as the one that was closest to it. The mode of latency is 360ms, 420ms and 400ms for the other three subjects, respectively.

We should point out that since there is about 1 local minimum per mode, the search need not be around the true mode for latency (we do not know this anyway). The results would be almost the same as long as the estimated mode is not skewed to its two neighboring true modes.

Figure 5-4 shows the scalp topographies for the four subjects plotted using EEGLAB (Delorme et al., 2004). As expected for a P300 topography, it has a large positive topography around the Pz area. To evaluate the single-trial estimation of the scalp topography, we compute the correlation between the single-trial scalp topography in (3.15) and the overall normalized scalp topography (3.17). For comparison, we also compute the correlation between the single-trial scalp topography in (3.15) and the scalp topography obtained from ensemble average for each subject. We name these two correlations r_1 and r_2 respectively. Statistical inference based directly on the correlation itself is difficult since its distribution is complicated. A popular approach is to first apply the Fisher Z transformation to correlation and then do inference on the transformed variable. The Fisher Z transform is given by (Fisher, 1915):

$$Z = 0.5 \ln \left(\frac{1+r}{1-r} \right) \quad (5-1)$$

Z has a simpler distribution and it converges more quickly to a normal distribution. We can calculate the mean and confidence interval of Z based on the correlation, if we assume that the estimation error in (3-15) is a normal distribution. The statistics of the correlation can be easily obtained from the inverse transform of (5-1).

The results are summarized in Table 5-1. We can see that there is a moderate amount of correlation between the single-trial and overall scalp topography (the average mean correlation for 4 subjects is around 0.40) although the mean correlation is lower for subject 3 at around 0.20.

There is a small degradation in mean correlation when the overall scalp topography is computed from the ensemble average. This is expected since the estimate in (3-17) is close to the ensemble averaged estimate. The correlation between these two estimates for the four subjects are: 0.80, 0.85, 0.89, 0.79 respectively.

To evaluate the effectiveness of the single-trial amplitude estimation, we related our estimates to a behavioral measure of target identification: response time in target trials. Response time was selected because task difficulty was relatively low, and therefore error rate did not show pronounced variability, with only limited numbers of misses (mean of 3.9 % across 4 participants) and false alarms (mean of 1.2 % across 4 participants). Thus, response time was used as a measure of target identification, with short response times indicating facilitated discrimination and long response times indicating difficulties with identification in a given trial. Using these measures, we were interested in the relationship between P300 amplitude and response time, expecting that trials in which participants found discrimination relatively easy (short RT trials) should be associated with greater P300 amplitude, which also indicates successful encoding of the target features and preparation for responding to a target that has been identified.

There seems to be little relationship between the response time and estimated single-trial peak latency. The correlation coefficients between these two for the four subjects are: 0.022, -0.248, 0.168 and 0.093 respectively. However, there were reliable negative correlations between the response time and estimated single-trial amplitude. Figure 5 shows the scatter plot of the response time versus the estimated amplitude for each single trial for the four subjects. To evaluate the statistical significance of the results, we performed linear regression on the response time and estimated single-trial amplitude for the four subjects. The results are summarized in

Table 5-2. the negative slope parameter estimated from linear regression is statistically significant under a significance level of 0.05 for all the four subjects, which supports our hypothesis that larger amplitude correspond to smaller response time, and vice versa.

To compare our results with conventional methods, we calculated the average P300 amplitude at channel 100 for subject #2. This was simply the average single-trial amplitude times the 100-th entry of the scalp topography in (3.17). It was found to be 22.1mV, compared with the 17.3mV from the ensemble average ERP. Taking into account of the possible latency jitter of P300, the true amplitude could be only larger than 17.3mV. Therefore, we obtained an upper bound of 28% on the positive bias of our average P300 estimate in channel 100. The coefficient of variation, which is defined as the ratio of the standard deviation to the mean of a positive random variable, is used as a measure of dispersion of the estimated amplitude and it was found to be around 0.60. This compares favorably with 0.79 obtained using the simple peak-picking method around its ensemble average peak at 400ms. Although the gain may seem small, we should keep in mind that this variation will incorporate the estimation error as well as that of the underlying change in P300 amplitude itself, because there are systematic changes in P300 amplitude as suggested above. So the estimation variance of our method is reduced by a factor of at least 1.7 from the peak-picking method. For instance, if one half of the total variance of our method came from the underlying P300 amplitude, this roughly means that our method reduced the estimation variance by a factor of 2.5 (assuming additive and uncorrelated estimation error). Of course, the comparison would be much more direct and informative if the P300 amplitude was expected to remain constant.

All the above results were obtained using a fixed Gamma template with $k = 11$, $\theta = 5$. If we change the template, specifically, the spread parameter θ , the estimated amplitude will also

change. However, we found that the amplitude estimation is only slightly affected by this change. For instance, the average estimated P300 amplitude in channel 100 for subject #2 was around 20.5mV when $\theta = 1$ (this is too small for P300, rise time 40ms) and was around 23.8mV when $\theta = 8$ (this is too large, rise time 320ms). There is less than 8% change from the result (22.1mV) obtained with the original template with $\theta = 5$. This agrees with our earlier findings using simulated ERP data (Li et al., 2008).

5.1.3 Discussions

As a straightforward test of the present method, we examined the relationship between target detection performance and features of the P300 component evoked by the targets in an oddball task with rare targets varying in terms of their salience on a trial-by-trial basis. In the present case, we replicated and extended a standard result in target detection studies in the visual domain: When target identification is made difficult or saliency is reduced (e.g., by presenting many targets in succession, Gonsalvez and Polich, 2002), P300 amplitude often decreases (Polich et al., 1997). This pattern has been interpreted as reflecting reduced resource allocation to a given target stimulus (Keil et al., 2007). Notably, previous work in this area has typically relied on averages across all trials of an experimental condition, or on block by condition averages across many trials (for a review, see Kok, 2001). The present results suggest that the relationship between response time and P300 amplitude in feature-based attention task is of a continuous nature, rather than a consequence of a bimodal function separating easy and hard trials. The sensitivity of the method was sufficient to demonstrate this linear relationship on a single-subject level, which is often desirable in clinical studies. In a similar manner, other research questions will benefit from the ability to examine hypotheses as to the time course and distribution of single brain responses, in terms of their magnitude and latency.

5.2 Habituation Study

5.2.1 Materials and Methods

EEG data were recorded from subjects during a passive picture-viewing experiment, consisting of 12 alternating phases: the habituation phase and mixed phase. Each phase has 30 trials. During the 30 trials of the habituation phase, the same picture was repeatedly presented 30 times. During the mixed phase, the 30 pictures are all different. Each trial lasts 1600 ms, and there is 600 ms pre-stimulus, and 1000 ms post-stimulus.

The scalp electrodes were placed according to the 128-channel Geodesic Sensor Nets standards. All 128 channels were referred to channel Cz and were digitally sampled for analysis at 250Hz. A bandpass filter between 0.01Hz and 40Hz was applied to all channels, which were then converted to average reference. To correct for vertical and horizontal ocular artifacts, an eye movement artifact movement correction procedure (Gratton et al., 1983) was applied to EEG recordings.

5.2.2 Estimation Results

We assume that the entire ERP may be decomposed into several monophasic components with compact support. We will estimate their parameters (amplitude and latency) one by one, using the Gamma template as in the simulation study. The present study illustrates the application of the method for a single late potential component, and the Gamma is not adapted. Its parameters are selected based on neurophysiology plausibility and are set as $k = 5$, $\theta = 6$, corresponding to a rise time of 96ms.

In reality, we do not know a priori exactly how many components there are in a single-trial, nor do we know when they occur. However, we may be able to estimate these values from single-trial EEG data in the data analysis session. Following the same procedures in Section 5.1.2, we identified 5 distinct peaks after 300ms of the stimulus onset. Assuming that the error in

the latency estimation is equally biased and independent from trial to trial and since there are also about 5 local minima for each trial, we conjecture that these peaks correspond to the latencies of 5 distinct components. These components, which may have different origins, are likely to compromise the Late Positive Potentials (LPP). According to Codispoti et al.(2006), the grand-average of LPP is maximal around 400ms to 500ms after stimulus. We will concentrate on the component with a latency of 500ms to exemplify the methodology. We search between 440ms and 560ms (which correspond to the two neighboring local minima) and set the component latency as the local minimum closest to 500ms.

To avoid the influence of EEG outliers from unexpected artifacts, we reject those trials with 3 times or larger amplitude of the minimum one. This will eliminate 14 trials from the total of 360 trials (rejection rate: 4%). The same rejection criterion was applied to the other two subjects, leading to the rejection of 8 (2%) and 33 (9%) trials, respectively.

Figure 5-6 shows the results of estimated scalp topography of the LPP component for 3 subjects. They are similar in the sense that all show large projections in the posterior area. The difference with subject 2 is that the scalp topography shifts its strength a bit to the occipital area. It may be that the pictures shown to the 3 subjects caused some emotional bias. It is also possible that the SNR of the ERP data is too low to allow for a stable estimate of the scalp projection across subjects (note that in habituation phase, the LPP amplitude decreases quickly with the trial index).

Figure 5-7 shows the results of estimated amplitude of the LPP component for 3 subjects. Each point in Fig. 5-7 stands for the average amplitude over 6 trials with the same index in the same phase (habituation or mixed). It is clear that for the habituation phase, the amplitude diminishes rapidly with the trial index, while for the mixed phase, the amplitude does not show

significant decay. To make the figure more intuitive, we also include the best fit (in the least square sense) to the estimated amplitude for both habituation and mixed phase. We fitted a straight line for the mixed phase, while an exponential curve was fitted to the estimated amplitude of the habituation phase. The fitted exponential curve for the habituation phase has a time constant of around 1.5 trials, which suggests that after 3 or 4 trials, the LPP amplitude decreases close to zero. We estimate the SNR for the mixed phase at around -4.1dB. Similar results were obtained with the ERP data from 2 other subjects as shown in Fig. 5-7 (B) and (C). The fitted exponential curves for the habituation phase for these 2 subjects has a time constant of around 1.5 and 2.0 trials, respectively. The SNR of the mixed phase for these 2 subjects are estimated to be around -7.6dB and -2.4dB respectively.

Table 5-1. Correlation statistics for the 4 subjects: Scalp topography estimation

Subject #	Sample size	r_1		r_2	
		Mean	Confidence interval (95%)	Mean	Confidence interval (95%)
1	98	0.520	[0.362, 0.649]	0.393	[0.215, 0.546]
2	85	0.367	[0.167, 0.538]	0.304	[0.097, 0.486]
3	74	0.198	[-0.027, 0.404]	0.200	[-0.025, 0.406]
4	65	0.446	[0.233, 0.619]	0.359	[0.133, 0.551]

Table 5-2. Regression statistics for response time and estimated amplitude

Subject #	Sample size	Correlation	R square	Slope estimate	t statistic	p value	Confidence interval (95%)
1	98	0.440	0.194	-0.712	-4.80	<0.0001	[-1.007, -0.418]
2	85	0.539	0.291	-0.310	-5.84	<0.0001	[-0.416, -0.205]
3	74	0.263	0.069	-0.105	-2.31	0.012	[-0.197, -0.014]
4	65	0.351	0.123	-0.108	-2.97	0.002	[-0.181, -0.036]

Figure 5-1 Pictures used in the experiment as stimuli

Figure 5-2 Cost function in (3.9) versus time lag for different regularization parameters for subject #2. A) $\lambda = 10^{-6}$. B) $\lambda = 10^{-5}$. C) $\lambda = 10^{-0}$. D) $\lambda = 10^2$. Regularization parameter that is too small led to ragged cost function and spurious latency estimates; Regularization parameter that is too large led to over-smoothed cost function and missed candidates for latency.

Figure 5-3 Estimated pdf of time lags corresponding to local minima of the cost function in (3-9) using the Parzen windowing pdf estimator with a Gaussian kernel size of 4.2. Regularization parameter $\lambda = 10^{-5}$.

Figure 5-4 Scalp topographies for the four subjects. A) Subject 1. B) Subject 2. C) Subject 3. D) Subject 4.

Figure 5-5 Scatter plot of the response time versus the estimated amplitude for each single trial for the four subjects. A) Subject 1. B) Subject 2. C) Subject 3. D) Subject 4. Note that the estimated amplitude is with respect to the EEG data in all the channels as a whole. There appears to be a negative relationship between the response time and the estimated amplitude.

A

B

C

Figure 5-6 Estimated scalp topography for mixed and habituation phase. A) Subject 1. B) Subject 2. C) Subject 3.

A

B

C

Figure 5-7 Estimated amplitude for mixed and habituation phase. A) Subject 1. B) Subject 2. C) Subject 3. Note that the LPP amplitude decreases with trials.

CHAPTER 6 CONCLUSIONS AND FUTURE RESEARCH

6.1 Conclusions

Traditional ERP analysis has relied on ensemble average over a large number of trials to deal with the typically low SNR environments in EEG data. To analyze ERP on a single event basis, we have introduced a new spatiotemporal filtering method for the problem of single-trial ERP estimation. Our method relies on explicit modeling of ERP components (not the full ERP waveform), and its output is limited to local descriptors (amplitude and latency) of these components. The reason that we model the ERP components instead of the full ERP waveform is to exploit the localization of scalp projection for each single ERP component, which is impossible to do for the entire ERP. Indeed, note that the ensemble ERP in different channels usually have different morphology because there are multiple neural sources originating from different locations of the brain that give rise to different scalp projections. Since one spatial filter can extract effectively only one scalp projection, in order to utilize the spatial information in a meaningful way, only a component based analysis is viable. Concentrating only on latency and amplitude of each component together with optimal spatial filtering presents an alternative to deal with the negative SNR. Moreover, since these are in fact the features of importance in cognitive studies, the methodology has the same descriptive power of traditional approaches.

The proposed methodology can be seen as a generalization of Woody's filter (Woody 1967) in the spatial domain for latency estimation. It also obtains an explicit expression for amplitude estimation on a single-trial basis. By design, the method is especially suitable to extract ERP features in the spontaneous EEG activity, in contrast to PCA and ICA which work best for reliable (large) signals. Another distinction is that, unlike most methods based on PCA and ICA, our method utilizes explicitly the timing information, as well as the spatial information. The

methodology as presented is based on least squares, but it can be further extended to robust estimation (Li et al. 2007) for better results.

Using simulated ERP data, we have shown that although the mismatch between the presumed and synthetic ERP components introduces a bias for both latency and amplitude estimation, the bias for the latency is relatively small and the estimated amplitudes are still comparable across experimental conditions for ERP data with a SNR higher than -20dB. Furthermore, the mismatch of components has minimal influence on the estimation of scalp projection. These all compare favorably with some of the popular methods (Li et al., 2008).

Despite its advantages over traditional methods, there are still some issues with our spatiotemporal filtering method. First it is based on the linear generative EEG model in (3.1). While this greatly simplifies the analysis, it may not be adequate to fully describe the complex information processing in the brain. One weak link of the method is that it requires an explicit template that is unknown *a priori*. Mismatch between the template and the true ERP component waveform brings both bias and larger variance to the estimation of the latency and amplitude that increases with decreasing SNR (Li, *et al.* 2008). It may be desirable to be able to adapt the template while estimating the model parameters. Another weakness of the method is the assumption of statistical uncorrelatedness among all the ERP components in deriving (3.14). With monophasic waveforms, this is equivalent to the condition that all the ERP components do not overlap in time (but overlap in space is allowed), which is seldom satisfied in practice. Temporal overlap will bring bias to the amplitude estimation and poses a serious problem for the latency estimation, since it works effectively only for monophasic waveforms that are well separated in time. When there is heavy overlap among multiple components (e.g., P300 and possibly other unknown late components), the peak latency estimation based on (3.10) may fail.

Therefore, care must be taken not to over-interpret the results of single-trial estimates. A crucial factor for amplitude estimation is a reasonably low SNR ($>-20\text{dB}$). This may not be satisfied for some ERP components under certain experimental conditions. Our ability to infer the template accurately, which are selected heuristically from real data, deteriorates with decreasing SNR. As a rule of thumb, we would recommend against the use of the present method for data with SNR less than -15dB .

6.2 Future Research

The use of a parametric template (Gamma function) provides the flexibility to change the shape and scale parameters continuously. However, this introduces undesirable bias when there is a mismatch between the template and true ERP component. Using a stochastic formulation, our method may be extended to a noisy template model and potentially the two nuisance Gamma parameters may be extracted from the data also for best fit.

It is almost certain that activations of different ERP components overlap in time. If this is the case, the temporal overlap will introduce a bias to the estimation of single-trial scalp projection, because the derivation in (3.15) assumes the uncorrelatedness between the ERP component and all the other sources (including the overlapping ERP components) and unlike the background EEG, these overlapping components are coherent in all the trials. This bias, together with the estimation variance due to finite-sample data, constitute the two main sources of error in the estimation of the scalp projection. Note that this in turn will influence the estimation of the amplitude. In Chapter 4, we have proposed an explicit procedure to compensate for the overlapping effects for the amplitude and topography estimation. This assumes that the latency is (relatively) fixed and a fairly accurate knowledge of the shape of all the overlapping ERP components. When this is not the case (particularly when we wish to find the latency change

from trial to trial in the presence of component overlap), we need to come up with new procedures to compensate for the overlapping issue.

The current method considers the single-trial amplitude of an ERP component as *i.i.d.* data. It may be advantageous to take into account the dynamics of certain properties of the component with respect to the trial index. For instance, we expect that during the habituation phase, the amplitude of LPP components diminishes rapidly with the number of trials. Using regularization techniques, this *a priori* information may be incorporated into the proposed single-trial estimation method to provide more stable estimate for the amplitude. In the end, the evolution of the single-trial amplitude with trial index may be inferred with more resolution and more confidence.

APPENDIX A
PROOF OF VALIDITY OF THE PEAK LATENCY ESTIMATION IN (3-10)

We justify the use of the time lag corresponding to the local minimum of $J(\tau)$ in (3-10) as the peak latency. Given the single-trial data matrix $\mathbf{X}$, the peak latency of an ERP component coincides with the local minimum of $J(\tau)$ if the following conditions are satisfied:

- (1), the presumed component $\mathbf{s}_o$ and the actual component $\mathbf{s}$ have the same morphology;
- (2), $\mathbf{n}_i^T \cdot \mathbf{s}_o(\tau) = 0$, for $i = 1, \dots, N$, and $\tau \in T_s$. (the signal and noise are uncorrelated);
- (3), $\mathbf{X}$ is full rank.

Proof: The optimal spatial filter is given by (3-8). We plug it into (3-7) and get:

$$J(\tau) = \|\mathbf{s}_o(\tau)^T \cdot (\mathbf{C} - \mathbf{I})\|_2^2 = \mathbf{s}_o(\tau)^T \cdot (\mathbf{C}^T \mathbf{C} - \mathbf{C}^T - \mathbf{C} + \mathbf{I}) \cdot \mathbf{s}_o(\tau)$$

where, $\mathbf{C} = \mathbf{X}^T (\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X}$. note that $\mathbf{C}^T \mathbf{C} = \mathbf{C}$, so,

$$J(\tau) = -\mathbf{s}_o(\tau)^T \cdot \mathbf{C} \cdot \mathbf{s}_o(\tau) + \mathbf{s}_o(\tau)^T \mathbf{s}_o(\tau) = -\left[\mathbf{s}_o(\tau)^T \mathbf{s}\right]^2 \cdot (\mathbf{a}^T \mathbf{R}^{-1} \mathbf{a}) + \mathbf{s}_o(\tau)^T \mathbf{s}_o(\tau)$$

where, $\mathbf{R} = \mathbf{X} \cdot \mathbf{X}^T$ is a positive definite matrix independent of the time lag τ .

With the constraint that $\mathbf{s}_o(\tau)^T \mathbf{s}_o(\tau) = \text{const}$, the minimum of the cost function $J(\tau)$ is achieved when $\mathbf{s}_o(\tau)^T \mathbf{s}$ achieves its maximum, since $\mathbf{a}^T \mathbf{R}^{-1} \mathbf{a}$ is positive. This happens when τ coincides with the peak latency l of the actual ERP component $\mathbf{s}$.

APPENDIX B

GAMMA FUNCTION AS AN APPROXIMATION FOR MACROSCOPIC ELECTRIC FIELD

The macroscopic electrical field is created from spike trains by a nonlinear generator with a second-order linear component with real poles (Freeman 1975). Suppose that the transfer function of the second-order system with stable real poles a, b is:

$$H(s) = \frac{1}{(s-a)(s-b)}$$

where, without loss of generality: $b < a < 0$.

Then the impulse response in the time domain is: $h(t) = \frac{1}{a-b}(e^{at} - e^{bt})$.

This is also a monophasic waveform with a single mode at $t_0 = \frac{1}{a-b} \ln(b/a)$. The rising time depends on the relative magnitude of the two real poles. The impulse response can be expanded:

$$h(t) = \frac{1}{a-b} e^{bt} \cdot (e^{(a-b)t} - 1) = \frac{1}{a-b} e^{bt} \cdot \sum_{n=1}^{\infty} \frac{[(a-b)t]^n}{n!}.$$

Thus we can see that the impulse response is a sum of infinite weighted Gamma functions. However, it is always possible to find a few dominant terms around the mode, where, $t_0(a-b) = \ln(b/a)$. If we knew the values for a, b , we can choose the shape parameter K of the Gamma function as the largest term, i.e., the integer part of $\ln(b/a)$.

A special case is when the system has two identical poles. Then, the impulsive response is exactly modeled by a single Gamma function with $K = 1$, $\theta = -1/a$. This is also approximately true when the magnitude of one pole is much larger than the other one.

APPENDIX C DERIVATION OF THE UPDATE RULE FOR THE CONSTRAINED OPTIMIZATION PROBLEM

Using one Lagrange multiplier, we convert the constrained optimization problem in (4-1) to an unconstrained optimization problem.

$$\arg \min_{\sigma, \mathbf{a}} \left\| \mathbf{X} - \sigma \cdot \mathbf{a} \cdot \mathbf{s}^T \right\|_F^2 + \lambda \cdot \left\| \mathbf{a} - \mathbf{a}_0 \right\|_2^2 + \mu \cdot \left(\left\| \mathbf{a} \right\|_2^2 - 1 \right)$$

Note that,

$$\begin{aligned} \left\| \mathbf{X} - \sigma \cdot \mathbf{a} \cdot \mathbf{s}^T \right\|_F^2 &= \text{Tr} \left[\left(\mathbf{X} - \sigma \cdot \mathbf{a} \cdot \mathbf{s}^T \right) \cdot \left(\mathbf{X} - \sigma \cdot \mathbf{a} \cdot \mathbf{s}^T \right)^T \right] \\ &= \text{Tr}(\mathbf{X}\mathbf{X}^T) - 2\sigma \text{Tr}(\mathbf{X}\mathbf{s}\mathbf{a}^T) + \sigma^2 \text{Tr}(\mathbf{a}\mathbf{s}^T\mathbf{s}\mathbf{a}^T) \\ &= \text{Tr}(\mathbf{X}\mathbf{X}^T) - 2\sigma \text{Tr}(\mathbf{a}^T\mathbf{X}\mathbf{s}) + \sigma^2 \mathbf{s}^T\mathbf{s} \text{Tr}(\mathbf{a}\mathbf{a}^T) \\ &= \text{Tr}(\mathbf{X}\mathbf{X}^T) - 2\sigma \mathbf{a}^T\mathbf{X}\mathbf{s} + \sigma^2 \mathbf{s}^T\mathbf{s} \cdot \mathbf{a}^T\mathbf{a} \end{aligned}$$

Setting the gradient of the Lagrangian function to 0 with respect to $\mathbf{a}, \sigma, \mu$ respectively, we find that the following set of equations holds:

$$\begin{cases} 2\sigma^2 \mathbf{s}^T\mathbf{s} \cdot \mathbf{a} - 2\sigma \mathbf{X}\mathbf{s} + 2\lambda(\mathbf{a} - \mathbf{a}_0) + 2\mu\mathbf{a} = 0 \\ -2\mathbf{a}^T\mathbf{X}\mathbf{s} + 2\sigma \mathbf{s}^T\mathbf{s} \cdot \mathbf{a}^T\mathbf{a} = 0 \\ \left\| \mathbf{a} \right\|_2^2 - 1 = 0 \end{cases}$$

Solving for $\mathbf{a}, \sigma$, we have:

$$\begin{cases} \sigma = \frac{\mathbf{a}^T\mathbf{X}\mathbf{s}}{\mathbf{s}^T\mathbf{s}} \\ \mathbf{a} = \frac{\lambda\mathbf{a}_0 + \sigma\mathbf{X}\mathbf{s}}{\left\| \lambda\mathbf{a}_0 + \sigma\mathbf{X}\mathbf{s} \right\|_2} \end{cases}$$

This is not a closed-form solution for the optimal values. However, it can be effectively used as a fixed point update to iteratively find the optimal values.

APPENDIX D
DERIVATION OF THE UPDATE RULE FOR THE UNCONSTRAINED OPTIMIZATION
PROBLEM

The unconstrained optimization problem is:

$$\arg \min_{\mathbf{a}} \left\| \mathbf{X} - \mathbf{a} \cdot \mathbf{s}^T \right\|_F^2 + \lambda \cdot \left\| \mathbf{a} / \|\mathbf{a}\|_2 - \mathbf{a}_0 \right\|_2^2$$

First we note that:

$$\begin{aligned} \left\| \mathbf{a} / \|\mathbf{a}\|_2 - \mathbf{a}_0 \right\|_2^2 &= \left(\mathbf{a} / \|\mathbf{a}\|_2 - \mathbf{a}_0 \right)^T \cdot \left(\mathbf{a} / \|\mathbf{a}\|_2 - \mathbf{a}_0 \right) \\ &= \mathbf{a}^T \mathbf{a} / \|\mathbf{a}\|_2^2 + \mathbf{a}_0^T \mathbf{a}_0 - 2 \mathbf{a}^T \mathbf{a}_0 / \|\mathbf{a}\|_2 \\ &= 2 - 2 \mathbf{a}^T \mathbf{a}_0 / \|\mathbf{a}\|_2 \end{aligned}$$

Taking the derivative to 0, we get,

$$2 \mathbf{s}^T \mathbf{s} \cdot \mathbf{a} - 2 \mathbf{X} \mathbf{s} + 2 \lambda \left(\mathbf{a}^T \mathbf{a}_0 \mathbf{a} / \|\mathbf{a}\|_2^3 - \mathbf{a}_0 / \|\mathbf{a}\|_2 \right) = 0$$

Or equivalently,

$$\mathbf{a} = \frac{\mathbf{X} \mathbf{s} + \lambda \mathbf{a}_0 / \|\mathbf{a}\|_2}{\mathbf{s}^T \mathbf{s} + \lambda \mathbf{a}_0^T \mathbf{a} / \|\mathbf{a}\|_2^3}$$

This is not a closed-form solution for $\mathbf{a}$. However, it can be effectively used as a fixed point update to iteratively find the optimal values.

APPENDIX E

MAP SOLUTION FOR THE ADDITIVE MODEL

With the assumptions indicated in Section 4.4, the posteriori probability can be rewritten as:

$$p(\mathbf{a}_o, \sigma_1 \dots \sigma_K \mid \mathbf{a}_1 \dots \mathbf{a}_K) = \prod_{k=1}^K p(\mathbf{a}_k \mid \sigma_k, \mathbf{a}_o) p(\sigma_k) p(\mathbf{a}_o) / p(\mathbf{a}_1 \dots \mathbf{a}_K)$$

Given the model, maximization of the posterior probability is equivalent to:

$$\arg \max_{\mathbf{a}_o, \sigma_1 \dots \sigma_K} \prod_{k=1}^K p(\mathbf{a}_k \mid \mathbf{a}_o, \sigma_k) p(\sigma_k) p(\mathbf{a}_o)$$

We assume a uniform (flat) *a priori* distribution for σ_k . Since $\mathbf{a}_o$ is constrained to have a unit norm, its *a priori* distribution is a Dirac delta function: $\delta(\|\mathbf{a}_o\|_2 - 1)$. If we assume that $\mathbf{u}_k$ is zero-mean *i.i.d.* Gaussian noise with the same covariance matrix $d^2 \mathbf{I}$ across all the trials, maximization of the posterior probability can be further simplified:

$$\arg \max_{\mathbf{a}_o, \sigma_1 \dots \sigma_K} \log \prod_{k=1}^K p(\mathbf{a}_k \mid \mathbf{a}_o, \sigma_k) p(\mathbf{a}_o) = \arg \max_{\mathbf{a}_o, \sigma_1 \dots \sigma_K} \sum_{k=1}^K \left(-\|\mathbf{a}_k - \sigma_k \mathbf{a}_o\|_2^2 \right) + \log \delta(\|\mathbf{a}_o\|_2 - 1)$$

This can be converted to a constrained optimization problem:

$$\min J = \sum_{k=1}^K \|\mathbf{a}_k - \sigma_k \mathbf{a}_o\|_2^2$$

S.t.

$$\|\mathbf{a}_o\|_2 = 1$$

Setting the derivative to zero, a necessary condition for minimum is: $\sigma_k = \mathbf{a}_o^T \cdot \mathbf{a}_k$.

Plug this into the above cost function. We have:

$$\begin{aligned}
J &= \sum_{k=1}^K \left\| \mathbf{a}_k - \mathbf{a}_o^T \cdot \mathbf{a}_k \cdot \mathbf{a}_o \right\|_2^2 = \sum_{k=1}^K \left\| \mathbf{a}_k - \mathbf{a}_o \cdot \mathbf{a}_o^T \cdot \mathbf{a}_k \right\|_2^2 \\
&= \sum_{k=1}^K \mathbf{a}_k^T \cdot \left(\mathbf{I} - \mathbf{a}_o \cdot \mathbf{a}_o^T \right)^T \cdot \left(\mathbf{I} - \mathbf{a}_o \cdot \mathbf{a}_o^T \right) \cdot \mathbf{a}_k \\
&= \sum_{k=1}^K \mathbf{a}_k^T \cdot \left(\mathbf{I} - \mathbf{a}_o \cdot \mathbf{a}_o^T \right) \cdot \mathbf{a}_k \\
&= \sum_{k=1}^K \mathbf{a}_k^T \cdot \mathbf{a}_k - \mathbf{a}_o^T \cdot \sum_{k=1}^K \mathbf{a}_k \cdot \mathbf{a}_k^T \cdot \mathbf{a}_o
\end{aligned}$$

The first term does not depend on $\mathbf{a}_o$, so MAP is equivalent to:

$$\arg \max_{\mathbf{a}_o} \mathbf{a}_o^T \cdot \mathbf{A} \cdot \mathbf{a}_o,$$

S.t.

$$\left\| \mathbf{a}_o \right\|_2 = 1$$

where $\mathbf{A} = \sum_{k=1}^K \mathbf{A}_k = \sum_{k=1}^K \mathbf{a}_k \cdot \mathbf{a}_k^T$. The matrix $\mathbf{A}$ is symmetric, so it can be diagonalized.

The maximum occurs when $\mathbf{a}_o$ is the normalized eigenvector of the matrix $\mathbf{A}$ corresponding to the largest eigenvalue.

APPENDIX F NORMALIZED ADDITIVE NOISE MODEL 1: MAP SOLUTION

Following the same rationale in model 1, we attempt to derive the MAP solution for the model 2.

The difference is in the covariance matrix of conditional probability: $p(\mathbf{a}_k | \mathbf{a}_o, \sigma_k) \sim N(0, \sigma_k^2 \mathbf{I})$

$$\text{So, } p(\mathbf{a}_k | \mathbf{a}_o, \sigma_k) = \frac{1}{(2\pi)^{D/2} |\sigma_k|^D} \exp\left(-\|\mathbf{a}_k / \sigma_k - \mathbf{a}_o\|_2^2 / 2\right)$$

The MAP becomes:

$$\begin{aligned} & \arg \max_{\mathbf{a}_o, \sigma_1 \dots \sigma_K} \prod_{k=1}^K p(\mathbf{a}_k | \mathbf{a}_o, \sigma_k) p(\sigma_k) p(\mathbf{a}_o) \\ &= \arg \max_{\mathbf{a}_o, \sigma_1 \dots \sigma_K} \sum_{k=1}^K \left(\|\mathbf{a}_k / \sigma_k - \mathbf{a}_o\|_2^2 \right) + \log \delta(\|\mathbf{a}_o\|_2 - 1) \end{aligned}$$

where, we have used the prior distribution for σ_k : $p(\sigma_k) = |\sigma_k|^D$

This is an *improper* prior (but still a prior for Bayesian inference). It is mainly motivated by analytical tractability. Similarly as in model 1, we can write the following constrained optimization problem:

$$\arg \min_{\mathbf{a}_o, \sigma_k} \sum_{k=1}^K \|\mathbf{a}_k / \sigma_k - \mathbf{a}_o\|_2^2$$

S.t.

$$\|\mathbf{a}_o\|_2 = 1$$

We first find a necessary condition for minimum: $\sigma_k = \mathbf{a}_k^T \cdot \mathbf{a}_k / \mathbf{a}_o^T \cdot \mathbf{a}_k$.

Let $\mathbf{B}_k = \frac{\mathbf{a}_k \cdot \mathbf{a}_k^T}{\mathbf{a}_k^T \cdot \mathbf{a}_k}$. Note that $\mathbf{B}_k$ is idempotent, i.e., $\mathbf{B}_k \mathbf{B}_k = \mathbf{B}_k$.

Plug this into the above cost function. We have:

$$\begin{aligned}
& \arg \min_{\mathbf{a}_o} \sum_{k=1}^K \left\| (\mathbf{I} - \mathbf{B}_k) \mathbf{a}_o \right\|_2^2 \\
&= \arg \min_{\mathbf{a}_o} \sum_{k=1}^K \mathbf{a}_o^T (\mathbf{I} - \mathbf{B}_k) \mathbf{a}_o \\
&= \arg \min_{\mathbf{a}_o} \left(K - \sum_{k=1}^K \mathbf{a}_o^T \mathbf{B}_k \mathbf{a}_o \right) \\
&= \arg \max_{\mathbf{a}_o} \mathbf{a}_o^T \mathbf{B} \mathbf{a}_o
\end{aligned}$$

$$\text{where, } \mathbf{B} = \sum_{k=1}^K \mathbf{B}_k = \sum_{k=1}^K \frac{\mathbf{a}_k \cdot \mathbf{a}_k^T}{\mathbf{a}_k^T \cdot \mathbf{a}_k} .$$

Again, the maximum occurs when $\mathbf{a}_o$ is the normalized eigenvector of the matrix $\mathbf{B}$ corresponding to the largest eigenvalue.

APPENDIX G

NORMALIZED ADDITIVE NOISE MODEL 2: MAP SOLUTION

Given the model: $\frac{\mathbf{a}_k}{\|\mathbf{a}_k\|_2} = \mathbf{a}_o + \mathbf{u}_k$, we attempt to find the MAP solution for the posteriori

probability .

$$\begin{aligned}
 & \arg \max_{\mathbf{a}_o} p(\mathbf{a}_o \mid \mathbf{a}_1 \dots \mathbf{a}_K) \\
 &= \arg \max_{\mathbf{a}_o} \log \prod_{k=1}^K p(\mathbf{a}_k \mid \mathbf{a}_o) p(\mathbf{a}_o) \\
 &= \arg \max_{\mathbf{a}_o} \sum_{k=1}^K \left(-\|\mathbf{a}_k / \|\mathbf{a}_k\|_2 - \mathbf{a}_o\|_2^2 \right) + \log \delta(\|\mathbf{a}_o\|_2 - 1)
 \end{aligned}$$

The equivalent constrained optimization problem is:

$$\arg \max_{\mathbf{a}_o} \sum_{k=1}^K \|\mathbf{a}_k / \|\mathbf{a}_k\|_2 - \mathbf{a}_o\|_2^2$$

S.t.

$$\|\mathbf{a}_o\|_2 = 1$$

This can be solved in a straightforward way:

$$\mathbf{a}_o = \frac{\sum_{k=1}^K \mathbf{a}_k / \|\mathbf{a}_k\|_2}{\left\| \sum_{k=1}^K \mathbf{a}_k / \|\mathbf{a}_k\|_2 \right\|_2}$$

LIST OF REFERENCES

- Al-Nashi H. (1986). A maximum likelihood method for estimating EEG evoked potentials. *IEEE Transactions on Biomedical Engineering*, 33:1087-1095.
- Aunon J.I. and McGillem C.D. (1975). Techniques for processing single evoked potentials. *Proceedings San Diego Biomedical Symposium*, 211-218
- Aunon J.I., McGillem C.D., and Childers D.G.. (1981). Signal processing in evoked potential research: averaging and modeling. *CRC Critical Reviews in Biomedical Engineering*, 4:323-367.
- Basar E, Demiralp T, Schurmann M, Basar-Erglu C, Ademoglu A. (1999). Oscillatory brain dynamics, wavelet analysis, and cognition. *Brain Lang*;66:146-183.
- Bell A., Sejnowski T., (1995). An Information Approach to Blind Separation and Blind Deconvolution, *Neural Computation*., vol. 7: 1129-1159.
- Belouchrani A., Abed-Meraim K., Cardoso J-F, Moulines E., (1997). A Blind Source Separation Technique Based on Second-Order Statistics, *IEEE Transaction on Signal Processing*, vol.45, 434-444.
- Brazier. M.A.B. (1964). Evoked responses recorded from the depths of the human brain. *Annals of New York Academy of Sciences*, 112: 33-59.
- Bruin K.J., Kenemans J.L., Verbaten M.N., Van der Heijden A.H., (2000) Habituation: an event-related potential and dipole source analysis study. *Intional Journal of Psychophysiol.* 36, 199-209.
- Cardoso J.-F. (1998). Blind signal separation: statistical principles, In *Proceedings of the IEEE, special issue on blind identification and estimation*, R.-W. Liu and L. Tong editors. Vol. 9(10), 2009-2025.
- Cardoso J.-F. (1999). High-Order Contrasts for Independent Component Analysis. *Neural Computation*, vol. 11, 157-192.
- Caspers, H., Speckmann, E.-J., & Lehmenkühler, A. (1980). Electrogenesis of cortical DC potentials. (eds. H. H. Kornhuber & L. Deecke), *Progress in brain research*: Vol. 54. Motivation, motor and sensory processes of the brain: Electrical potentials, behavior and clinical use 3-15. Amsterdam: Elsevier.
- Cerutti S., Bersani V., Carrara A., and Liberati D. (1987). Analysis of visual evoked potentials through Wiener filtering applied to a small number of sweeps. *Journal on Biomedical Engineering*, 9:3-12.

- Chapman, R., McCrary, J., (1995). EP component identification and measurement by principal components analysis. *Brain Cognition* 27 (3), 288–310. (Erratum in: *Brain Cogn.* 28 (3) 342, 1995.)
- Choi S., Cichocki A., Park H. -M., and Lee S. -Y. (2005). Blind source separation and independent component analysis: A review. *Neural Information Processing - Letters and Review*, vol. 6, no. 1, 1-57.
- Cichocki A. and Amari S. (2002). Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. New York: Wiley.
- Ciganek Z. (1969). Variability of the human visual evoked potential: normative data. *Electroencephalography and Clinical Neurophysiology*, 27:35-42.
- Codispoti M., Ferrari V., Bradley M., (2006). Repetitive picture processing: autonomic and cortical correlates. *Brain Research*. 1068: 213-220.
- Coifman R.R., Wickerhauser MY. Wavelets, (1996). adapted waveforms and denoising. *Clinical Neurophysiology Supplement*; 45:57-78.
- Coles MGH, Rugg MD. (1995). Event-related brain potentials: an introduction. *Electrophysiology of mind: event-related brain potentials and cognition*. (ed. Rugg MD, Coles MGH), Oxford University Press. Oxford, UK. 1–26.
- Comon P. (1994). Independent component analysis, a new concept? *Signal Processing*, vol. 36, 287–314.
- Cuthbert, B.N., Schupp, H. T., Bradley, M.M., and Lang, P.J. (1996). Affective picture viewing: Task and stimulus effects on startle P3 and blink. *Psychophysiology*, 33.
- Daubechies. I. (1992). Ten Lectures on Wavelets. *SIAM*.
- Debener S, Makeig S, Delorme A, Engel AK, (2005). What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis, *Experimental Brain Research: Cognitive Brain Research*. 22:309-321.
- Delorme A, Makeig S, Fabre-Thorpe M, Sejnowski TJ, (2002). From single-trial EEG to brain area dynamics, *Neurocomputing*, 44-46: 1057-1064.
- Delorme, A., Makeig, S. (2003) EEG changes accompanying learning regulation of the 12-Hz EEG activity. *IEEE Transactions on Rehabilitation Engineering*, 11(2), 133-136.
- Delorme, A., Sejnowski, T., Makeig, S. (2007) Improved rejection of artifacts from EEG data using high-order statistics and independent component analysis. *Neuroimage*, 34, 1443-1449

- Dien J. (1998). Addressing misallocation of variance in principal components analysis of event-related potentials. *Brain Topography*;11(1):43–55.
- Dien J, Beal DJ, Berg P. (2005). Optimizing principal components analysis of event-related potentials: Matrix type, factor loading weighting, extraction, and rotations. *Clinical Neurophysiology*;116:1808–25.
- Doncarli C. and Goerig L. (1988). Adaptive smoothing of evoked potentials: a new approach. *Proceedings 10th Annual Conference IEEE-EMBS*, New Orleans, Louisiana, 1152-1153
- Donchin E. (1966). A multivariate approach to the analysis of average evoked potentials. *IEEE Transaction on Biomedical Engineering*;13(3):131–9.
- Donchin E, Heffley EF. (1978). Multivariate analysis of event-related potential data: a tutorial review. In: Otto DA, editor. Multidisciplinary perspectives in event-related brain potential research. *Proceedings of the fourth international congress on event-related slow potentials of the brain* (EPIC IV), Hendersonville, NC, April 4–10, 1976. Washington, DC: The Office; 555–572.
- Donchin, E., Ritter, W., McCallum, C. (1978). Cognitive psychophysiology: The endogenous components of the ERP. *Brain event-related potentials in man* (ed. E. Callaway, P. Tueting, S. Koslow), Academic Press, New York. 349-441.
- Donchin, E. (1979). Event-related brain potentials: a tool in the study of human information processing. *Evoked potentials and behavior* (ed. H. Begleiter), Plenum, New York. 13-75.
- Donchin, E. (1981). Surprise! surprise? *Psychophysiology*, 18, 493-513.
- Donchin, E. and Coles, M. G. H. (1988a). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, 11, 355-72.
- Donchin, E. and Coles, M. G. H. (1988b). On the conceptual foundations of cognitive psychophysiology. *Behavioral and Brain Sciences*, 11, 406- 17.
- Donchin, E., Spencer, K.M. and Dien, J. (1997). The varieties of deviant experience: ERP manifestations of deviance processors. In: G.J.M. van Boxtel and K.B.E. Bocker (Eds.), *Brain and Behavior: Past, Present, and Future*, Tilburg: Tilburg University Press, 67-91.
- Donoho D.L. and Johnstone I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.
- Dornhege G., Blankertz B., Krauledat M., Losch F., Curio G., Müller K.-R. (2006). Combined optimization of spatial and temporal filters for improving Brain-Computer Interfacing. *IEEE Transactions on Biomedical Engineering*, 53(11), 2274—2281.

- Duffy, F.H., Jones, K., Bartels, P., Albert, M., McAnulty, G.B. and Als, H. (1990). Quantified neurophysiology with mapping: Statistical inference, exploratory and confirmatory data analysis. *Brain Topography*, 3(1): 3-12.
- Dyrholm M, Christoforou C., Parra L.C., (2007). Bilinear discriminant component analysis, *Journal of Machine Learning Research*, vol. 8, 1097-1111
- Effern A, Lehnertz K, Fernandez G, Grunwald T, David P, Elger CE. (2000). Single trial analysis of event related potentials: non-linear de-noising with wavelets. *Clinical Neurophysiology*; 11:2255–2263.
- Fabiani, M., Gratton, G. , Karis, D., & Donchin, E. (1987). Definition, identification and reliability of measurement of the P300 component of the event-related brain potential. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.) *Advances in psychophysiology*, Vol. 2 (pp. 1-78). Greenwich , CT : JAI Press.
- Farwell L.A., Martinerie J.M., Bashore T.R., Rapp P.E., and Goddard P.H. (1993). Optimal digital filters for long-latency components of the event-related brain potential. *Psychophysiology*, 30:306-315.
- Fisher, R.A., (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10, 507-521.
- Freeman W., (1975). *Mass activation in the nervous system*. Academic Press, New York
- Gaillard AWK, Ritter WK. (1983). *Tutorials in event-related potential research: endogenous components*. Amsterdam: North-Holland Publishing Company.
- Gansler T. and Hansson M. (1991). Estimation of event-related potentials with GSVD. *Proceedings 13th Annual Conference. IEEE-EMBS*, 423-424.
- Garoosi V. and Jansen B. (2000). Development and evaluation of the piecewise prony method for evoked potential analysis. *IEEE Transaction on Biomedical Engineering*. 47(12): 1549-1554.
- Glaser EM, Ruchkin DS. (1976). *Principles of neurobiological signal analysis*. New York: Academic Press.
- Gonsalvez CL, Polich J (2002) P300 amplitude is determined by target-to-target interval. *Psychophysiology* 39:388-396
- Gratton, G., Coles, M.G.H., Donchin, E., (1983). A new method for off-line removal of ocular artifact. *Electroencephalography Clinical Neurophysiology*. 55, 468 – 484
- Gratton G., Kramer A.F., Coles M.G.H., and Donchin E. (1989). Simulation studies of latency measures of components of the event-related brain potential. *Psychophysiology*, 26:233-248.

- Hansson M. and Cedholt T. (1990). Estimation of event related potentials. *Proceedings 12th Annual Conference IEEE-EMBS*, 901-902.
- Hansson M., Gansler T., and Salomonsson G. (1996). Estimation of single event-related potentials utilizing the prony method. *IEEE Transactions on Biomedical Engineering*, 43(10):973-981.
- Hendrickson AE, White PO. (1964). Promax: a quick method for rotation to oblique simple structure. *Br J Stat Psychol*;17:65–70.
- Hesse C.W. and James C.J., (2006). On Semi-Blind Source Separation using Spatial Constraints with Applications in EEG Analysis, *IEEE Transactions on Biomedical Engineering*. 53(12): 2525-2534.
- Horn JL. (1965). A rationale and test for the number of factors in factor-analysis. *Psychometrika*; 30(2):179–85.
- Huber P., (1981). *Robust Statistics*. New York: Wiley.
- Hyvärinen A., Oja E. (1997). A Fast Fixed-Point Algorithm for Independent Component Analysis, *Neural Computation*, vol. 9: 1483–1492.
- Hyvaerinen A., (1999). Survey on independent component analysis. *Neural Computation. Sur.*, vol. 2: 94–128.
- Hyvarinen A., Karhunen J. and Oja E. (2001). Independent Component Analysis. New York: Wiley
- James, C. and Hesse, C. (2004). Independent component analysis for biomedical signals. *Physiological Measurement*, 26, (1), 15-39.
- Jaskowski P. and Verleger R. (1999). Amplitudes and latencies of single-trial ERP's estimated by a maximum-likelihood method. *IEEE Transactions on Biomedical Engineering*, 46(8):987-993.
- Johnson, R., Jr. (1986). A triarchic model of P300 amplitude. *Psychophysiology*, 23, 367 - 384.
- Johnson R. (1989). Developmental evidence for modality-dependent P300 generators: a normative study. *Psychophysiology* 26: 651-667.
- Jung T-P, Makeig S, Westerfield M, Townsend J, Courchesne E, and Sejnowski TJ, (1999). Analyzing and Visualizing Single-trial Event-related Potentials. *Advances in Neural Information Processing Systems*, 11:118-124.

- Jung T-P, Makeig S, Lee T-W, McKeown M.J., Brown G., Bell, A.J. and Sejnowski TJ, (2000). Independent Component Analysis of Biomedical Signals, *The 2nd Int'l Workshop on Independent Component Analysis and Signal Separation*, 633-644.
- Jung T-P, Makeig S, McKeown M.J., Bell, A.J. , Lee T-W, and Sejnowski TJ, (2001). Imaging Brain Dynamics Using Independent Component Analysis, *Proceedings of the IEEE*, 89(7):1107-1122.
- Junghöfer M, Elbert T, Leiderer P, Berg P, Rockstroh B (1997) Mapping EEG-potentials on the surface of the brain: a strategy for uncovering cortical sources. *Brain Topography* 9:203-217
- Junghöfer M, Elbert T, Tucker DM, Rockstroh B (2000) Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology* 37:523-532
- Kaiser HF. (1958). The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*. 23:187–200.
- Kalman R.E. (1960). A new approach to linear filtering and prediction problems. *Transactions ASME, Journal of Basic Engineering*. 82D:35-45.
- Karjalainen P.A., Kaipio J.P., Koistinen A.S., and Karki T. (1996). Recursive Bayesian estimation of single trial evoked potentials. *Proceedings 18th Annual Conference IEEE-EMBS*, Amsterdam
- Karjalainen PA. (1997). Regularization and Bayesian Methods for Evoked Potential Estimation. *Ph.D. Thesis*, Kuopio University Publications C. Natural and Environmental Sciences 61
- Karjalainen P.A., Kaipio J.P., Koistinen A.S., Vauhkonen M. (1999). Subspace regularization method for the single-trial estimation of evoked potentials. *IEEE Transaction on Biomedical Engineering* 46(7):849-60.
- Keil A, Bradley MM, Junghoefer M, Russmann T, Lowenthal W, Lang PJ (2007) Cross-modal Attention Capture by Affective Stimuli: Evidence from Event-Related Potentials. *Cognitive, Affective, & Behavioral Neuroscience* 7:18-24
- Koch C., Poggio T., Torre V., (1983). Nonlinear interactions in a dendritic tree: localization, timing, and role in information processing. *Proceedings National Academy of Science USA* 80: 2799–2802
- Koles ZJ, Lazar MS, Zhou SZ. (1990). Spatial patterns underlying population differences in the background EEG. *Brain Topography*; 2(4):275-284.
- Koles ZJ, Lind JC, Flor-Henry P. (1994). Spatial patterns in the background EEG underlying mental disease in man. *Electroenceph Clinical Neurophysiology*, 91:319-328.

- Kok A., (2001) On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology* 38:557-577
- Krieger S., Timmer J., Lis S., and Olbrich H.M. (1995). Some considerations on estimating event-related brain signals. *Journal of Neural Trans Gen Sect*, 99(1-3):103-129.
- Lang, P., Bradley, M. M., and Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. In P. Lang, R. F. Simons, & M. Balaban (Eds.), *Attention and orienting: Sensory and motivational processes*. Hillsdale, NJ: Erlbaum. 97–136.
- Lang PJ, Bradley MM, Cuthbert BN (2005) International Affective Picture System: Technical Manual and Affective Ratings. In. Gainesville, FL: NIMH Center for the Study of Emotion and Attention
- Lange D.H. (1996). Variable single-trial evoked potential estimation via principal component identification. *Proceedings 18th Annual Conference IEEE-EMBS*, Amsterdam.
- Lee T. W., Girolami M., and Sejnowski T. J., (1999). Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and superGaussian sources. *Neural Computation*., vol. 11, 417–441.
- Lee T.W., Girolami M., Bell A. J., and Sejnowski T. J., (2000). A unifying information-theoretic framework for independent component analysis. *Computing Math Application*., vol. 39, 1–21.
- Lemm S., Blankertz B., Curio G., and Müller K.-R., (2005). Spatio-spectral filters for improved classification of single trial EEG, *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 9, 1541–1548.
- Li R., Principe J., (2006). Blinking Artifact Removal in Cognitive EEG Data using ICA, *International Conference of Engineering in Medicine and Biology Society*, 6273-6278.
- Li R., Principe J.C., Bradley M., Ferrari V., (2007) Robust single-trial ERP estimation based on spatiotemporal filtering. *Proceedings IEEE EMBS Conference*. 5206-5209
- Li R., Principe J.C., Bradley M., Ferrari V., (2008) A spatiotemporal filtering methodology for single-trial ERP component estimation. *IEEE Transactions on Biomedical Engineering*, In Press
- Liberati D., Bertolini L., and Colombo D. C. (1991). Parametric method for the detection of inter- and intrasweep variability in VEP processing. *Medical and Biological Engineering and Computing*, 29:159-166,.
- Linden D. (2005). The P300: Where in the brain is it produced and what does it tell us? *Neuroscientist*, 11(6): 563-576.

- Liu W., Pokharel P., Principe J., (2007). Correntropy: Properties and Applications in Non-Gaussian Signal Processing, *IEEE Transactions on Signal Processing*, Vol. 55, Issue 11
- Luo A. and Sajda P. (2006) Learning discrimination trajectories in EEG sensor space: application to inferring task difficulty, *Journal of Neural Engineering*, 3 (1) L1-L6.
- Maccabee P.J., Pinkhasov E. I., and Cracco R. Q. (1983). Short latency evoked potentials to median nerve stimulation: Effect of low-frequency filter. *Electroencephalography and Clinical Neurophysiology*, 55:34-44.
- Madhavan G. P., Bruin H. de, and Upton A.R.M. (1984). Evoked potential processing and pattern recognition. *Proceedings 6th Annual Conference IEEE-EMBS*, 699-702.
- Madhavan G. P., Bruin H. de, and Upton A. R. M. (1986) Improvements to adaptive noise cancellation. *Proceedings 8th Annual Conference IEEE-EMBS*, Fort Worth, TX, 482-486.
- Madhavan G. P. (1988). Comments on adaptive filtering of evoked potentials. *IEEE Transactions on Biomedical Engineering*, 35:273-275.
- Madhavan G. P. (1992). Minimal repetition evoked potentials by modified adaptive line enhancement. *IEEE Transactions on Biomedical Engineering*, 39(7):760-764.
- Makeig, S., Bell, A., Jung, T., Sejnowski, T., (1996). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, vol. 8. MIT Press, 145–151.
- Makeig S, Jung T-P, Bell AJ, and Sejnowski TJ, (1997). Blind Separation of Auditory Event-related Brain Responses into Independent Components. *Proceedings of National Academy of Sciences*, 94:10979-10984.
- Makeig S, Westerfield M, Jung T-P, Covington J, Townsend J, Sejnowski TJ, and Courchesne E, (1999). Functionally Independent Components of the Late Positive Event-Related Potential during Visual Spatial Attention," *Journal of Neuroscience*, 19: 2665-2680.
- Makeig S., Jung T.-P., Ghahremani D., and Sejnowski T. J. (2000). Independent component analysis of simulated ERP data. in *Hum. High. Func. I: Adv. Meth.*, T. Nakada, Ed.
- Makeig S, Westerfield M, Jung T-P, Enghoff S, Townsend J, Courchesne E, Sejnowski TJ. (2002). Dynamic brain sources of visual evoked responses. *Science*, 295:690-694.
- Makeig S, Delorme A, Westerfield M, Jung T-P, Townsend J, Courchesne E, Sejnowski TJ. (2004). Electroencephalographic brain dynamics following visual targets requiring manual responses, *PLOS Biology*, 2(6):747-762.
- Marple S. Lawrence Jr. (1987). Digital Spectral Analysis with Applications. Prentice Hall, Englewood Cliffs.

- McGillem C.D. and Aunon J.I. (1987). Analysis of event-related potentials, chapter 5, Elsevier Science Publisher. 131-169.
- Mocks J, Verleger R. (1991). Application of principal component analysis to event-related potentials. In Weitkunat R., editor. Multivariate methods in biosignal analysis: *Digital Biosignal Processing*. Amsterdam: Elsevier; 399–458.
- Müller-Gerking J., Pfurtscheller G., and Flyvbjerg H., (1999). Designing optimal spatial filters for single-trial EEG classification in a movement task, *Clinical Neurophysiology*, 110: 787-798.
- Naatanen, R. and Picton, T. W. (1987). The N 1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24, 375-425.
- Nishida S., Nakamura M., and Shibasaki H. (1993). Method for single-trial recording of somatosensory evoked potentials. *Journal on Biomedical Engineering*, 150:257-262.
- Parra, L.C., Alvino, C., Tang, A., Pearlmutter, B., Young, N., Osman, A., Sajda, P., (2002). Linear spatial integration for single-trial detection in encephalography. *NeuroImage* 17, 223– 230
- Parra, L.C., Sajda, P., (2003). Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*. 4, 1261–1269.
- Parra, L., Spence, C., Gerson, A., Sajda, P., (2005). Recipes for the linear analysis of EEG. *Neuroimage*, 28, 326-341.
- Parzen, E., (1962) On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, Vol. 33, No. 3. 1065-1076
- Patterson R. D., Robinson K., Holdsworth J., McKeown D., Zhang C., and Allerhand M. H., (1992) Complex sounds and auditory images, In *Auditory Physiology and Perception*, (Eds.) Y Cazals, L. Demany, K.Horner, Pergamon, Oxford, 429-446
- Pham D. T., Mocks J., Kohler W., and Gasser T. (1987). Variable latencies of noisy signals: Estimation and testing in brain potential data. *Biometrika*, (74) 525–533.
- Picton TW, Bentin S, Berg P, Donchin E, Hillyard SA, Johnson R Jr, Miller GA, Ritter W, Ruchkin DS, Rugg MD, Taylor MJ. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology*. 37(2):127-52

- Polich J, Alexander JE, Bauer LO, Kuperman S, Morzorati S, O'Connor SJ, Porjesz B, Rohrbaugh J, Begleiter H (1997) P300 topography of amplitude/latency correlations. *Brain Topogr* 9:275-282
- Polich J. (1999) P300 in clinical applications. *Electroencephalography: basic principles, clinical applications and related fields* (Eds. E. Niedermayer and F. Lopes de la Silva). Urban and Schwarzenberger, Baltimore-Munich. 1073-1091.
- Quiñan Quiroga R., García H. (2003). Single-trial event-related potentials with wavelet denoising. *Clinical Neurophysiology*. 114: 376–390.
- Ramoser, H., Mueller-Gerking, J., Pfurtscheller, G., (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transaction on Rehabilitation Engineering*. 8 (4), 441– 446.
- Reilly, J., (1992). *Applied Bioelectricity*. Springer. New York
- Roberts S. and Everson E. (2001) Independent Component Analysis: Principles and Practice Cambridge: Cambridge University Press
- Ruchkin D.S. and Glaser E.M. (1978). Simple digital filters for examining CNV and P300 on a single trial basis. (ed. Otto D.A.) *Multidisciplinary perspectives on event-related brain potential research*, U.S. Government Printing Office, Washington DC. 579-581.
- Ruchkin D.S. (1987). Measurement of event-related potentials. In *Human Event-Related Potentials*, volume 3 of *Handbook of Electroencephalography and Clinical Neurophysiology*, Elsevier. 7-44.
- Samar VJ, Swartz KP, Raghuveer MR. (1995). Multiresolution analysis of event related potentials by wavelet decomposition. *Brain Cognition*; 27:398-438.
- Sams, M., Alho, K., Näätänen, R. (1984). Short-term habituation and dishabituation of the mismatch negativity of the ERP. *Psychophysiology*, 21, 434-441.
- Scherg, M. and Picton, T. W. (1991). Separation and identification of event-related potential components by brain electric source analysis. *Event-related brain research, EEG Suppl.* 42, (ed. C. H. Brunia, G. Mulder, and M. N. Verbaten), Elsevier, Amsterdam. 24-37.
- Schiff SJ, Aldrouby A, Unser M, Sato S. (1994). Fast wavelet transform of EEG. *Electroenceph Clinical Neurophysiology*; 91:442-455.
- Silva F.H. Lopes da. (1993). Event-related potentials: Methodology and quantification. *Electroencephalography: Basic principles, clinical applications and related fields*, Williams & Wilkins, 877-886.

- Silverman B.W. (1986), *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London
- Spencer KM, Dien J, Donchin E. (1999). A componential analysis of the ERP elicited by novel events using a dense electrode array. *Psychophysiology*;36:409–14.
- Spencer KM, Dien J, Donchin E. (2001). Spatiotemporal analysis of the late ERP responses to deviant stimuli. *Psychophysiology*; 38(2):343–58.
- Spreckelsen M. and Bromm B. (1988). Estimation of single-evoked cerebral potentials by means of parametric modeling and filtering. *IEEE Transactions on Biomedical Engineering*, 35:691-700.
- Steeger G.H., Herrmann O., and Spreng M. (1983). Some improvements in the measurements of variable latency acoustically evoked potentials in human EEG. *IEEE Transactions on Biomedical Engineering*, 30:295-303.
- Sutton, S., Braren, M., Zubin, J., and John, E. R. (1965). Evoked potential correlates of stimulus uncertainty. *Science*, 150, 1187-8.
- Tang A., Pearlmutter B., Malaszenko N., Phung D., Reeb B., (2002) Independent components of magnetoencephalography: localization. *Neural Computation*. 14 (8), 1827– 1858
- Thakor N.V. (1987). Adaptive filtering of evoked potentials. *IEEE Transactions on Biomedical Engineering*, 34(1):6-12.
- Thakor N, Xin-rong G, Yi-Chun S, Hanley D. (1993). Multiresolution wavelet analysis of evoked potentials. *IEEE Transactions on Biomedical Engineering*; 40:1085–1094.
- Truccolo, W., Knuth, K.H., Shah, A., Schroeder, C., Bressler, S.L., Ding, M. (2003) Estimation of single-trial multi-component ERPs: differentially Variable Component Analysis (dVCA). *Biological Cybernetics*, 89, 426-438.
- Tsai AC, Liou M, Jung TP, Onton JA, Cheng PE, Huang CC, Duann JR, Makeig S. (2006). Mapping single-trial EEG records on the cortical surface through a spatiotemporal modality. *Neuroimage*. 32(1):195-207.
- van Boxtel GJM. (1998). Computational and statistical methods for analyzing event-related potential data. *Behavioral Research Methods Instrumental Computation*. 30(1): 87–102.
- Vigario R., Jousmäki V., Hämäläinen M., Hari R., and Oja E., (1998). Independent component analysis for identification of artifacts in magnetoencephalographic recordings, *Proceedings NIPS*, Cambridge, MA, MIT Press, 229-235.

- Vigario R, Sarela J, Jousmaki V, Hamalainen M, Oja E. (2000). Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transaction on Biomedical Engineering*. 47(5):589-593.
- Vila C.E. Da, Welch A.J., and Rylander H.G. III. (1986). Adaptive estimation of single evoked potentials. *Proceedings 8th Annual Conference IEEE-EMBS*, 406-409.
- Wan Eric A. and Nelson Alex T., (1996). Dual Kalman Filtering Methods for Nonlinear Prediction, Estimation, and Smoothing, in *Advances in Neural Information Processing Systems 9*.
- Widrow B. and Stearns S. D. (1985). Adaptive Signal Processing. Englewood Cliffs, NJ: Prentice-Hall.
- Wood CC, McCarthy G. (1984). Principal component analysis of event-related potentials: Simulation studies demonstrate misallocation of variance across components. *Electroencephalography Clinical Neurophysiology*. 59:249–60.
- Wood, C. C. (1987). Generators of event-related potentials. *A textbook of clinical neurophysiology* (ed. A. M. Halliday, S. R. Butler, and R. Paul), Wiley, New York. 535-567.
- Woody C.D., (1967) Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. *Medical and Biological Engineering and Computing*, 5:539-553
- Yu K. and McGillem C.D. (1983). Optimum filters for estimating evoked potential waveforms. *IEEE Transactions on Biomedical Engineering*, 30:730-737.

BIOGRAPHICAL SKETCH

Ruijiang Li was born in Shandong, China. He received the B.S. degree in automation with emphasis on systems and control in 2004, from Zhejiang University, Hangzhou, China. Since 2004, he has been working toward his Ph.D. at the Electrical and Computer Engineering Department at the University of Florida, under the supervision of Jose C. Principe. His current research interests include statistical signal processing, machine learning and their applications in biomedical engineering.