Removing an Intersubject Variance Component in a General Linear Model Improves Multiway Factoring of Event-Related Spectral Perturbations in Group EEG Studies

Jeffrey S. Spence,[1,2]* **Matthew R. Brier,**[3] **John Hart, Jr.,**[3]
and Thomas C. Ferree[4]

[1]*Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas*
[2]*Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas*
[3]*Center for BrainHealth, School of Behavioral and Brain Science, University of Texas at Dallas, Dallas, Texas*
[4]*Department of Radiology and Biomedical Engineering Program, University of Texas Southwestern Medical Center, Dallas, Texas*

◆ ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ ◆

Abstract: Linear statistical models are used very effectively to assess task-related differences in EEG power spectral analyses. Mixed models, in particular, accommodate more than one variance component in a multi-subject study, where many trials of each condition of interest are measured on each subject. Generally, intra- and intersubject variances are both important to determine correct standard errors for inference on functions of model parameters, but it is often assumed that intersubject variance is the most important consideration in a group study. In this article, we show that, under common assumptions, estimates of some functions of model parameters, including estimates of task-related differences, are properly tested relative to the intrasubject variance component only. A substantial gain in statistical power can arise from the proper separation of variance components when there is more than one source of variability. We first develop this result analytically, then show how it benefits a multiway factoring of spectral, spatial, and temporal components from EEG data acquired in a group of healthy subjects performing a well-studied response inhibition task. *Hum Brain Mapp 34:651–664, 2013.* © 2011 Wiley Periodicals, Inc.

Key words: time-frequency analysis; principal components; statistical power

◆ ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ ◆

INTRODUCTION

Time-frequency analyses of EEG data acquired in cognitive experiments cover spatial, temporal, and spectral dimensions, requiring an analytic approach that incorporates both inferential statistics and data reduction. Often the goal is to determine regional differences in brain activity between contrasts of task conditions and baseline or prestimulus intervals. Linear statistical models or general linear models (GLM) are commonly used to assess task-related differences, and principal components are often used as a data reduction tool for interpretability.

This analytic approach to handle the large "volume" of data in EEG has been addressed previously. In the context of event-related potentials, a body of work advocates using sequential principal component analysis (PCA) for data reduction, then conducting statistical tests on the resulting components [Donchin, 1966; Dien, 2006; Dien et al., 2003; Spencer et al., 1999]. In that approach, subjects and task conditions are treated as samples in the PCA, and inference between conditions is conducted by modeling the components retained after sequential PCA in the GLM. The term "sequential" refers to the sequential application of two-way PCA to the multiway array after repeated unfolding; we refer to this ERP-based approach as PCA-ANOVA.

In recent work, Ferree et al. [2009] tested the application of PCA-ANOVA to event-related spectral perturbations and found several types of instability in the results. First, the largest PCA component reflected 60 Hz noise, which was not significantly different between conditions within any single subject, but was variable across electrodes and subjects. It was concluded that this approach was not effective at isolating activity that is specific to task conditions. This problem extends beyond 60 Hz noise to include prominent features in the power spectrum (e.g., $1/f$-like behavior at low frequencies and the alpha resonance) that dominate the spectrum in each condition but may not be different between conditions. Second, the results were highly sensitive to the deletion of a single subject, with differential sensitivity across subjects. In effect, PCA-ANOVA finds regions of high intersubject variance because this is the dominant component of the PCA, which then enters the GLM as the dependent variable.

Recalling that event-related spectral perturbations are usually analyzed within subjects, using trials to test for differences between conditions, it was hypothesized that task-related differences could be better isolated by first applying the GLM on the power spectral estimates themselves within each subject, and then submitting only significant contrasts of conditions and baseline to PCA. This procedure produced more stability in the results, including a proper cancellation of 60 Hz noise, and more robustness against single-subject deletions.

This new method, termed STAT-PCA [Ferree et al., 2009], is an improvement over PCA-ANOVA, because intersubject variance does not contribute to the test on condition differences, thus better isolating task-related differences. However, in that first instantiation of STAT-PCA subjects were still included as samples in the PCA. Subsequent investigations reported here showed that a single subject can, in some cases, yield a dominant PCA component so that the spatial, temporal, and spectral pattern of activation reflects the pattern of the single subject rather than the activation pattern common to the group.

For group studies, in which the subjects are sampled from populations of interest (e.g., healthy controls or patient groups), an implicit goal is to assess task-related changes in brain activity that are characteristic of the theoretical population from which the sample is taken.

This is best accomplished by incorporating the subjects in appropriate mixed effects linear models at the inferential stage (STAT) before the PCA. These models, explicitly stated in the Theory section, are more general in the sense that they allow estimates of parameter effects and more than one variance component within the framework of a single model rather than adopting an inferential model separately for each subject. In addition, the correct separation of variance components still allows the testing of task-related differences relative to the intrasubject variance component, and the incorporation of subjects in a single model removes subjects from the PCA so that only space, time and frequency dimensions need to be reduced by PCA.

Our motivation for the development of the models described below is generated by the desire to extend the original STAT-PCA procedure to allow a more flexible set of experimental designs while improving its robustness properties. This article demonstrates two distinct advantages of the proposed models: (1) the statistical power to detect task-related changes in brain activation is markedly improved when there is more than one source of variability, and (2) the PCA data reduction step more accurately displays the group pattern of activations since the intersubject variances are correctly accounted for during the inferential stage before PCA. The general linear models and the separation of variance components from several sources of variability are first described theoretically, then its implementation is evaluated by comparing this extended STAT-PCA time-frequency analysis to the original STAT-PCA procedure in a simple response inhibition task for which the main result is known from the literature.

THEORY

Statistical Models

As noted above the results of EEG power spectral analyses typically span space, time and frequency (STF). Depending on the spatial, temporal and spectral resolution, the array of results is typically of an order of magnitude similar to an fMRI statistic map image with each STF combination a point in the three-dimensional array, or a "voxel" in fMRI terminology. For ease of notation we suppress the reference to the particular STF "voxel" in the data array and model the logarithm of the power spectral density (PSD) at each STF "voxel." By defining $y_{jkl} = \log PSD_{jkl}$ we describe the observed responses y_{jkl} as deriving from an experimentally imposed task condition and random variation in both subjects and trials using the standard linear model notation

$$y_{jk\ell} = \mu + b_j + \gamma_k + \varepsilon_{jk\ell}, \qquad (1)$$

where $j = 1,\ldots,n$ subjects; $k = 1,\ldots,c$ conditions; and $l = 1,\ldots,t_{jk}$ trials in each subject and condition. This is a mixed effects linear model with subjects as a random effect, b_j, and task condition as a fixed effect, γ_k. Thus, there are two

variance components, σ_ε^2 and σ_b^2, with the distributional assumptions $\varepsilon_{jkl} \overset{i.i.d.}{\sim} N(0,\sigma_\varepsilon^2)$ and $b_j \overset{i.i.e.}{\sim} N(0,\sigma_b^2)$, independent of ε_{jkl}. These variance components are the intra- and inter-subject variances, respectively. Note that this is a simple extension of a within-subject linear model $y_{kl} = \mu + \gamma_k + \varepsilon_{kl}$, which is used implicitly in standard t-tests on task condition differences. The additional random effect term for subjects, b_j, allows the accommodation of this additional source of variability.

Model 1 asserts no *subject-by-condition* interaction. That is, it assumes each subject, sampled from a single theoretical population, responds similarly to the task conditions. Since the sample of subjects in the response inhibition task, described in the Methods section, is presumably drawn from a single population (e.g., healthy controls between 18 and 25 years of age), there is no a priori reason to suspect a *subject-by-condition* interaction. However, if this interaction is suspected, then Model 1 can be modified as

$$y_{jkl} = \mu + b_j + \gamma_k + (b\gamma)_{jk} + \varepsilon_{jkl}, \qquad (2)$$

with the interaction term also a random effect, adding a third variance component, $\sigma_{b\gamma}^2$, corresponding to the independent random variable $(b\gamma)_{jk} \overset{i.i.d.}{\sim} N(0,\sigma_{b\gamma}^2)$. Considerable variability among subjects with respect to the effect of task condition, indicating a sample drawn from more than one population, might justify this additional interaction term. For example, if some subjects respond to an inhibition task with an increase in PSD at a given STF voxel, while others respond with a decrease, and still others show no response at all, an interaction term will "capture" the extra variability due to an inhomogeneous group of subjects (i.e., when the sample is, in fact, not drawn from a single population). Otherwise, a systematic effect of task condition across all subjects from a homogeneous group ($\sigma_{b\gamma}^2 = 0$) allows one to generalize the condition effect to the population from which the sample is taken.

Models 1 and 2 both take advantage of two features of the acquired EEG data: (1) the condition effect, γ_k, is a crossed factor; and (2) the variability in the EEG PSD can be linearly decomposed due to the independence of within-subject and between-subject sources. The first feature simply means that all levels k of the task condition occur in each subject; the second feature allows the separation of these sources of variability, which improves statistical efficiency. To see how these models can substantially increase statistical power to detect differences due to task condition when the variance of y_{jkl} includes more than one source of variability, we first demonstrate analytically how the separate variance components enter into an analysis of variance (ANOVA) framework and then develop the t-statistic for any contrast of interest among the levels of task condition.

F-statistics in an ANOVA are ratios of mean squares or variances, each estimated by a quadratic form of the data vector comprised of elements y_{jkl}. Since the models we are positing are mixed-effects models having more than one

TABLE I. Expected values of mean square estimates for each effect in linear Models 1 and 2

Effect	EMS
Model 1	
Subject	$\sigma_\varepsilon^2 + ct\,\sigma_b^2$
Condition	$\sigma_\varepsilon^2 + \phi^2\,(\gamma)$
Error	σ_ε^2
Model 2	
Subject	$\sigma_\varepsilon^2 + t\sigma_{b\gamma}^2 + ct\sigma_b^2$
Condition	$\sigma_\varepsilon^2 + t\sigma_{b\gamma}^2 + \phi^2\,(\gamma)$
Subj × cond	$\sigma_\varepsilon^2 + t\sigma_{b\gamma}^2$
Error	σ_ε^2

variance component, it is important to have the appropriate mean square in the denominator of the F-statistic for correct inference. These can be determined by taking expected values of the mean square estimates or expected mean squares (EMS) corresponding to each effect in the linear model. Table I shows these expected values for each model and for $t_{jk} = t$ [see Milliken and Johnson, 2009].

What is important to note about Table I is the fact that the EMS for *condition*—the only fixed effect in either model—does not contain the intersubject variance component, σ_b^2. In other words, for both models the intersubject variance, σ_b^2, is not one of the variance components that comprise the total variability when assessing the effect of task conditions. This means that under the null hypothesis of no *condition* effect, the quadratic function of the fixed effect parameter, $\phi^2(\gamma)$, is zero; and the appropriate denominator in the F-statistic, which tests the effect due to task condition, is an estimate of one of the following expected mean squares: σ_ε^2 if Model 1 is posited, or $\sigma_\varepsilon^2 + t\sigma_{b\gamma}^2$, if Model 2 is posited.

Contrasts among the levels of task condition take the form $\gamma_k - \gamma_{k'}$ in Model 1 or $\gamma_k - \gamma_{k'} + (\overline{b\gamma})_{.k} - (\overline{b\gamma})_{.k'}$ in Model 2 for $k \neq k'$. It is known that the best linear unbiased estimator of any task condition contrast in both models is $\bar{y}_{.k.} - \bar{y}_{.k'.}$ [Searle, 1971], where the averages are taken over the indices j and l, i.e., over the subjects and trials. To form a t-statistic one needs the variance of these contrasts. This is easily found, using the EMS Table I, as

$$\mathrm{Var}(\bar{y}_{.k.} - \bar{y}_{.k'.}) = 2\sigma_\varepsilon^2/nt \quad \text{or}$$
$$\mathrm{Var}(\bar{y}_{.k.} - \bar{y}_{.k'.}) = 2(\sigma_\varepsilon^2 + t\sigma_{b\gamma}^2)/nt$$

for Models 1 and 2, respectively. Estimates of the variances are obtained by replacing the EMS components with mean square values (MS) taken directly from an ANOVA table, and the square root of the variance estimates constitute the denominator of the t-statistic, t_s.

$$t_s = (\bar{y}_{.k.} - \bar{y}_{.k'.})/\sqrt{2 \cdot \mathrm{MS}_{\text{error}}/nt} \quad \text{or}$$
$$t_s = (\bar{y}_{.k.} - \bar{y}_{.k'.})/\sqrt{2 \cdot \mathrm{MS}_{\text{subj×cond}}/nt} \qquad (3)$$

In Model 1 the *t*-statistic has $nct - n - c + 1$ degrees of freedom; in Model 2 the *t*-statistic has $(n - 1)(c - 1)$ degrees of freedom. Again, note that the variance of task condition contrasts does not involve the intersubject variance component, σ_b^2, in either model. This is precisely the source of the increase in statistical power for detecting differences in task conditions when using either Model 1 or 2.

Calculation of Statistical Power

If one does not account for the separation of variance components in a full GLM as we have done explicitly in these linear mixed effects models, then the intersubject variance component will be included in the estimate of error variance and will necessarily contribute a positive constant to the denominator of *F*- and *t*-statistics, thereby reducing statistical power to detect task condition differences. The magnitude of the reduction in statistical power will depend on the relative contributions of these variance components comprising the total $\text{Var}(y_{jkl}) = \sigma_\varepsilon^2 + \sigma_{b\gamma}^2 + \sigma_b^2$ (or $\sigma_\varepsilon^2 + \sigma_b^2$ if $\sigma_{b\gamma}^2 = 0$). If the intersubject variance component, σ_b^2, is a significant proportion of $\text{Var}(y_{jkl})$, then the reduction in statistical power can be substantial.

Let $\sigma_{b\gamma\varepsilon}^2$ denote the inflated error variance when the intersubject variance component is not appropriately removed using Model 1 or 2. Thus, $\sigma_{b\gamma\varepsilon}^2$ in the reduced GLM contains all variance components and the inflated contrast variances are $\sigma_\varepsilon^2 + t\sigma_b^2$ (Model 1) or $\sigma_{b\varepsilon}^2 + t\sigma_{b\gamma}^2 + t\sigma_b^2$ (Model 2), each multiplied by $2/nt$. By further defining the relative efficiency (*r.e.*) as the ratio of variances $\sigma_\varepsilon^2/\sigma_{b\gamma\varepsilon}^2$ (Model 1) or $(\sigma_\varepsilon^2 + t\sigma_{b\gamma}^2)/\sigma_{b\gamma\varepsilon}^2$ (Model 2), one can easily quantify the loss of statistical power to detect a condition difference. The *r.e.* is at most one when $\sigma_b^2 = 0$ and decreases as the relative contribution of σ_b^2 increases.

Statistical power, $\Phi(\lambda)$, as a function of the noncentrality parameter, λ, of Student's *t*-distribution on ν degrees of freedom and denoted $\phi_\lambda(\cdot;\nu)$, is calculated as [Casella and Berger, 1990]

$$\Phi(\lambda) = \int_{-\infty}^{t_{\alpha/2}} \phi_\lambda(x;\nu)dx + \int_{t_{1-\alpha/2}}^{\infty} \phi_\lambda(x;\nu)dx,$$

where α is the desired voxel-level error rate. The reduction in statistical power due to the inflated variance is $\Phi(\lambda^*)$, where $\lambda^* = \sqrt{r.e.} \times \lambda$.

Good theoretical and applied sources that provide confirmation and considerably more detail for all the preceding material are the following: Casella and Berger [1990], Milliken and Johnson [2009], and Searle [1971].

METHODS

Cognitive Task

To first demonstrate the efficacy of the full GLM, then compare the original STAT-PCA procedure with our
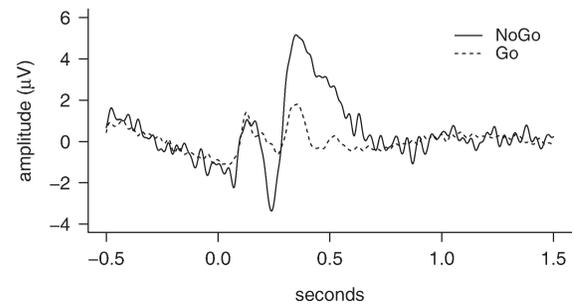
Figure 1.

The ERP signature, averaged over subjects and trials, in the frontocentral electrode (FCz) following NoGo and Go conditions.

extended modifications using the full GLM, we implemented a simple inhibition task—"Go/NoGo"—that elicits a well-known response. The ERP signature of that response is a frontocentral response known as the N2 [Jodo and Kayama, 1992; Smith et al., 2007] and P3 [Bruin et al., 2001]. That ERP has a period in the range of theta oscillations [Luu and Tucker, 2001], and in fact the NoGo condition results in a larger theta band increase than the Go condition [Yamanaka and Yamamoto, 2010, see Fig. 1].

Participants were instructed to push a button if they were presented with an arrow ("Go" trials constitute 80% of trials) and to withhold a response if presented with an octagon ("NoGo" trials constitute 20% of trials). Stimuli were presented for 300 msec followed by 1,700 msec of blank screen. Baseline trials were obtained from the prestimulus intervals for each subject.

Participants

Twenty-six subjects (12 males and 14 females, ages 18–25) participated in the Go/NoGo experiment while EEG was recorded. All were free of neurological deficits by self-report. All were right handed and gave informed consent before participation in accordance with the Institutional Review Board of The University of Texas at Dallas. This study was conducted at the UT Dallas Center for BrainHealth, following the Good Clinical Practice Guidelines, the Declaration of Helsinki, and the U.S. Code of Federal Regulations.

Data Acquisition and Preprocessing

Continuous EEG was recorded from a 64-electrode Neuroscan Quickcap using a Neuroscan SynAmps2 amplifier and Scan 4.3.2 software sampled at 1 kHz and hardware filtered at 200 Hz with impedances typically below 10 kΩ. An experienced EEG technician preprocessed the data manually. First, data recorded from poorly functioning electrodes were visually identified and removed. Second, eye blink artifacts were removed by a spatial filtering algorithm in the Neuroscan Edit software (Compumedics,

Inc.), using the option to preserve the background EEG. Third, the data were passed through an automated artifact removal program [Junghöfer et al., 2000] which detects individual channel and global artifacts based on the recording reference and average reference, respectively. The data from each subject and electrode were epoched by condition, and re-referenced to the spline-based average reference [Ferree, 2006]. The Fourier power spectrum was calculated within the peristimulus intervals using 0.5-s wide windows moving in 0.05-s steps. In each window the time series was linearly detrended, cosine tapered, and zero-padded to 1.0-s duration to achieve 1 Hz resolution. Finally, the squared modulus was normalized to obtain the power spectral density (PSD) in units of $\mu V^2/Hz$. For this simple response inhibition task we analyzed frequencies only up to 25 Hz.

Inference From the Full GLM

The logarithm of the power spectral density was modeled at each STF voxel as indicated in Models 1 and 2 with $n = 26$ subjects, $c = 3$ conditions—baseline, Go, NoGo—and t_{jk} trials; i.e., a variable number of trials which depended on subject and condition. To assess which of the two models was more appropriate for our EEG data we estimated each of the variance components by restricted maximum likelihood (REML) and calculated the percentage of each component to the total variance of y_{jkl}. Thus, the relative contributions of the intra-subject variance, σ_ε^2, the interaction variance, $\sigma_{b\gamma}^2$, and the intersubject variance, σ_b^2, were estimated across the STF voxels as the ratios $\hat{\sigma}_\varepsilon^2/(\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{b\gamma}^2 + \hat{\sigma}_b^2)$, $\hat{\sigma}_{b\gamma}^2/(\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{b\gamma}^2 + \hat{\sigma}_b^2)$, and $\hat{\sigma}_b^2/(\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{b\gamma}^2 + \hat{\sigma}_b^2)$, respectively.

For the response inhibition task described above, primary interest was centered on the contrast between the two task conditions, Go and NoGo. Hence, t-statistics were calculated similar to those shown in Eq. (3) but modified appropriately to accommodate the unequal t_{jk} [Milliken and Johnson, 2009]. All analyses were done using the mixed procedure in SAS (Cary, NC) with the degree-of-freedom method indicated in Kenward and Roger [1997] to adjust the downward bias in variances when testing for fixed effects in mixed models. This constituted the inferential stage (STAT) of the STAT-PCA procedure. Only those STF voxels (62 electrodes × 21 time points × 25 frequencies) with significant task-condition differences were passed to the PCA stage of the STAT-PCA analysis. This was done as follows.

Let s be an index of STF-voxel location, and let STF(s) be the estimate of $\bar{y}_{\cdot NoGo\cdot} - \bar{y}_{\cdot Go\cdot}$ at voxel s. An indicator array, denoted

$$I(s) = \begin{cases} 1 & \text{if } q < q^* \\ 0 & \text{if } q \geq q^*; \end{cases}$$

where q is the false discovery rate (FDR) q-value [Benjamini and Hochberg, 1995; Storey, 2003], provided a "significance mask." Finally, the array calculated by element-wise multi-

plication, STF(s)·I(s), was then passed to the PCA stage, which is described below. In this study we set a significance threshold at $q^* = 0.01$.

Sequential PCA

To obtain the most salient subsets from the full set of potential results a PCA of the statistically significant results was conducted sequentially with two levels of "unfolding." Specifically, the STF(s)·I(s) array, which spans three dimensions, was arranged as a two-dimensional data matrix with frequencies as columns and concatenated electrodes/time points as rows. The first PCA in the sequence was performed, followed by component selection using parallel analysis [Horn, 1965] and a *varimax* rotation [Kaiser, 1958] of the subspace of retained components. The corresponding scores were determined by projecting the data matrix onto the rotated components. For each spectral component the corresponding score was rearranged, again, as a two-dimensional data matrix with electrodes as columns and time points as rows. A second PCA was performed, followed by a second round of component selection by parallel analysis and *varimax* rotation. The temporal components following the second PCA were taken to be the scores corresponding to the retained spatial components.

Much more detail of the sequential PCA procedures can be found in Ferree et al. [2009]. An important distinction, however, between the sequential PCA in Ferree et al. [2009] and this article is that the former included a third temporal PCA with time points as columns in the rearranged data matrix and subjects as rows. Since, in this article, subjects were included in the full GLM at the inferential stage of the STAT-PCA procedure, the number of steps in the sequential PCA were reduced to two. The sequential PCA, parallel analysis and varimax rotation were implemented in the R statistical computing language (available at: http://www.r-project.org) and MATLAB (available at: http://www.mathworks.com).

Visualization of Principal Components

Once the sequential PCA was completed we had, for each contrast of interest, a "branching" of spectral, spatial and temporal components. We refer to this as "branching" because each spectral component may be associated with more than one spatial component, These triplets of spectral, spatial, and temporal components can be plotted in side-by-side panels showing the time-course of the dominant frequencies associated with unique topographic maps on the surface of the head. In addition, because earlier work on STAT-PCA showed good agreement between the temporal component and the time course of the group-averaged power at the peak frequency and electrode, time-courses of individual subjects were also investigated to assess their individual contributions to the group average.
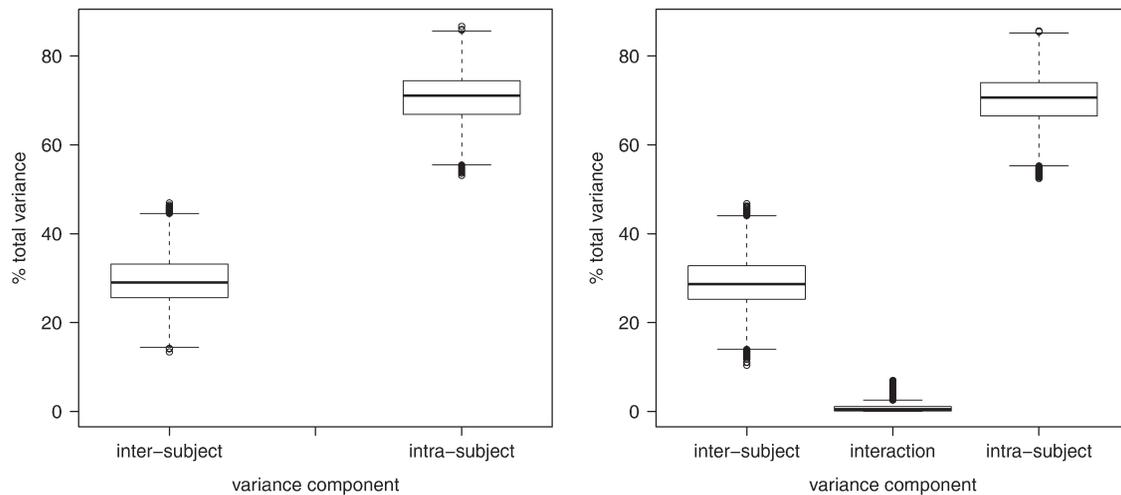
Figure 2.

Box plots of estimates of the variance components in Model 1 (left) and Model 2 (right) as a percentage of total variance over all STF voxels. The *subject* × *condition* interaction variance component does not contribute a significant proportion to the total variance (median = 0.52%). The inter- and intrasubject variance components constitute an average 30% and 70% of the total variance, respectively, regardless of which model is posited. Hence, Model 1 is sufficient to explain the variability inherent in the response inhibition task for our sample.

RESULTS

Assessing the Contributions of Each Variance Component to the Total Variance

Model 2 allows for the possibility that our sample of subjects is not from a single population and so responds differently to the inhibition task, through the *subject* × *condition* interaction term. In that case it would not be legitimate to extrapolate the findings to a single assumed population. Evidence that $\sigma_{b\gamma}^2 = 0$, therefore, would indicate that the effect due to response inhibition is consistent across subjects within the bounds of intrasubject variability and that this finding would legitimately generalize to a single population. We first explore the variance components implicit in Models 1 and 2.

To assess which of the two models was more appropriate for these data, we estimated all three variance components in Model 2 and calculated the contributions and $\hat{\sigma}_b^2$, $\hat{\sigma}_{b\gamma}^2$, $\hat{\sigma}_\varepsilon^2$ as a percentage of total variance. We expected the contribution of the *subject* × *condition* interaction variance component to be small since there was no a priori reason to expect heterogeneity among the subjects in this task.

Figure 2 shows box plots of percentages of total variance for each component, where the percentiles are based on distributions over all STF voxels. As expected the interaction variance component is near zero. The 25th, 50th, and 75th percentiles are 0.12%, 0.52%, and 1.10%, respectively; and nearly all (99.9%) of STF voxels have interaction variance component estimates comprising less than 5% of the total variance. The left panel of Figure 2 shows the distribution summaries from separate estimates of the inter- and intrasubject variance components in Model 1. They are nearly identical to those from Model 2, indicating that Model 2 does not provide important additional information about the variability of the subjects' response to inhibition in our sample.

All subsequent analyses reported here, therefore, utilize Model 1, the two-variance-component model, for the inference stage of STAT-PCA.

Statistical Power

Statistical power curves were calculated as formulated in the Theory section, using mean REML estimates of σ_b^2 and σ_ε^2 in Model 1 from the response inhibition task, and a conservative value of t equal to the harmonic mean $((t_h = c \cdot \{\sum_{k=1}^c 1/\bar{t}_{\cdot k}\}^{-1}))$. The estimated relative efficiency, *r.e.*, based on those values is $1/(1 + t_h\hat{\sigma}_b^2/\hat{\sigma}_\varepsilon^2)$.

Figure 3 demonstrates the dramatic differences in statistical power to detect an increase in the mean EEG power spectral density in response to the "NoGo" stimulus at the peak electrode and time point when Var(y_{jkl}) is comprised of intersubject and intrasubject variances. Without appropriately separating the inter- and intrasubject variance components using Model 1, there is a large decrease in *r.e.* and, therefore, a large decrease in statistical power. For example, a 20% increase in the mean EEG power spectral density due to the "NoGo" condition can be detected with statistical power 0.95 using Model 1. In our sample, to reach comparable statistical power in a reduced GLM, where
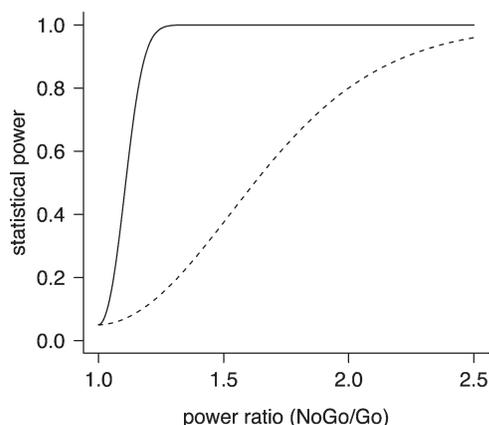
Figure 3.

Statistical power curves as a function of the increase in the EEG power spectral density due to the NoGo response at peak electrode and time poststimulus (based on REML estimates of the variance components in the response inhibition task and [α] = 0.05). Appropriate separation of the variance components based on Model 1 yields the solid curve. The loss of statistical power—dashed curve—results from the dramatic decrease of relative efficiency when the inter-subject variance component is not accounted for in the GLM.

subject-level variability is incorporated incorrectly into the test for condition differences, the increase in the mean spectral density would have to be as high as 150% (see Fig. 3).

Comparison of Methods in the Analysis of the Response Inhibition Task

The response inhibition task provides a benchmark against which we compare the original STAT-PCA procedure to our extension of it based on Model 1 (or Model 2.) As a point of reference, Figure 4 shows the baseline power spectral density (PSD) estimates for two separate electrodes and two separate frequency ranges—FCz and 5 to 6 Hz; FPz and 18 to 19 Hz. These particular electrodes and frequencies are chosen because the former is the peak location and partial frequency range of the known theta response to NoGo [Bruin et al., 2001; Jodo and Kayama, 1992; Luu and Tucker, 2001; Smith et al., 2007; Yamanaka and Yamamoto, 2010], and the latter is one of several electrodes and partial frequency ranges with no task condition effect, but where a single subject shows anomalous measures relative to the rest of the group. In the course of the analyses we find that this one subject provides the source of dominating PCA components when intersubject variance is not accounted for in the full GLM and, therefore, the subjects are not removed at the inference step.

Figure 5 shows in detail what occurs statistically at the peak electrode (FCz) within the theta band for each of three contrasts of interest: Go vs. Baseline, NoGo vs. Baseline, and NoGo vs. Go. We see that there is a small but significant increase in PSD for the Go condition relative to

baseline, particularly in the first half of the epoch. NoGo trials, however, produce a much larger increase in PSD, which, relative to the Go trials, peaks at 0.35 s poststimulus onset. Figure 5 demonstrates that, at the known location and frequency range of the response to inhibition, most subjects share the same strong PSD increase in the NoGo trials and that subject-level inference, through separate individual statistical models, or group inference, through the full GLM (Model 1 or Model 2), yields similar PSD ratios to be passed to sequential PCA after statistical thresholding. Consequently, we would expect the original STAT-PCA procedure to be sufficient in isolating the response to the NoGo task.

In contrast, Figure 6 reveals that a single subject, inconsonant with the rest of the group, can pass non-zero PSD ratios to the sequential PCA when employing the original STAT-PCA procedure because intersubject variances are not taken into account. Figure 6 is an example in a frontal electrode (FPz) at higher frequencies (18–19 Hz), where there is no known response to the conditions of the inhibition task. By accounting for the inter-subject variance component using the full GLM of Model 1 (or Model 2), Figure 6 shows that none of the PSD ratios are passed to the PCA. Otherwise, the single subject survives thresholding and contributes very low PSD ratios to PCA throughout the entire epoch, which has a strong influence on the components of the PCA, inappropriately dominating the group pattern of activation.

The utility of PCA resides in its ability to distill the global set of significant results into interpretable components. To understand from where the PCA components arise, one method of displaying all voxel-level results from a task condition contrast is to display time-frequency color maps for every electrode in a layout with similar topography as the electrode cap. Figure 7 shows these layouts for (a) the full GLM and (b) the original STAT-PCA, each comparing the NoGo response with the Go response. Frequency ranges from 1 to 25 Hz are shown vertically, and the temporal epoch is shown horizontally (0–1 s) within each channel of the electrode cap. In Figure 7a one sees from the frontal midline electrodes that the theta band increase in PSD due to the NoGo trials occurs at 0.35 s just before a lower frequency band (up to 4 Hz) increase in PSD at 0.4 s poststimulus onset. Figure 7b, on the other hand, reveals much lower PSD ratios in many electrodes: the two frontal electrodes (FPz and FP2) at nearly constant amplitude in the frequency range 16 to 25 Hz for the entire 1-s epoch; and predominantly occipital and parietal electrodes in various frequency bands around 0.6 s poststimulus onset. The lower PSD ratios in Figure 7b are all due to a single subject.

To display the global set of statistically significant results by PCA, the non-zero PSD ratios, following the inference stage from NoGo/Go, are passed to the sequential PCA. Retained frequency components from the first level are shown in the scree plots. Figure 8a follows the inference stage (STAT) using Model 1, which incorporates subjects in a single model; and Figure 8b follows the inference stage from the original procedure, where each subject has a separate
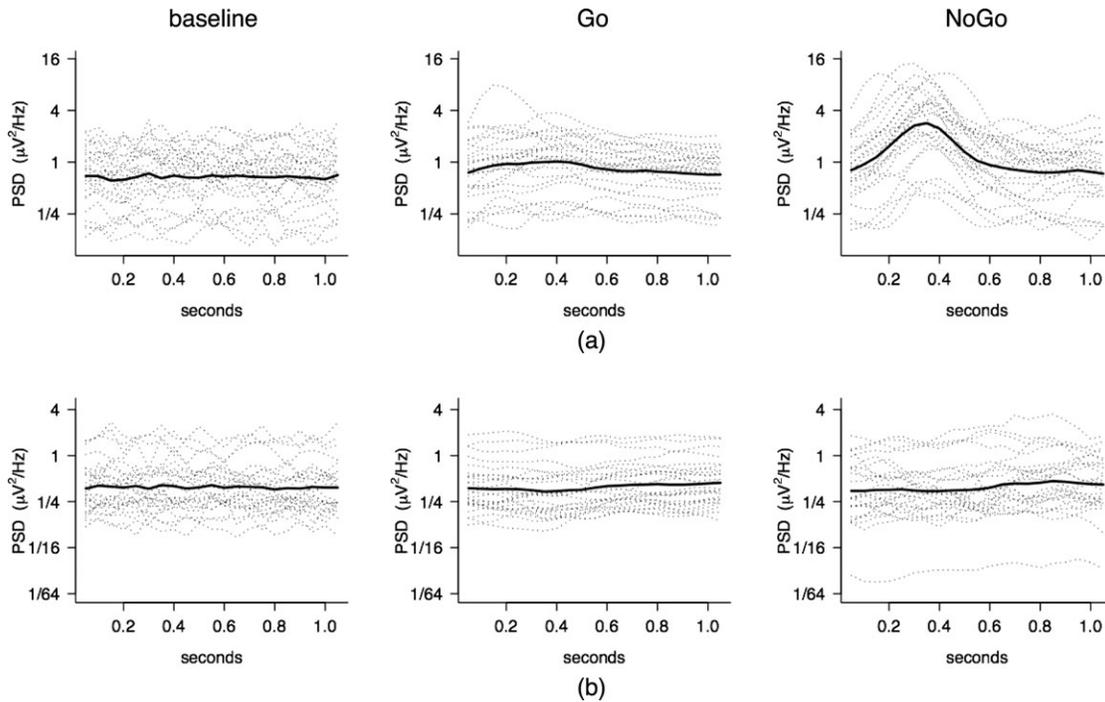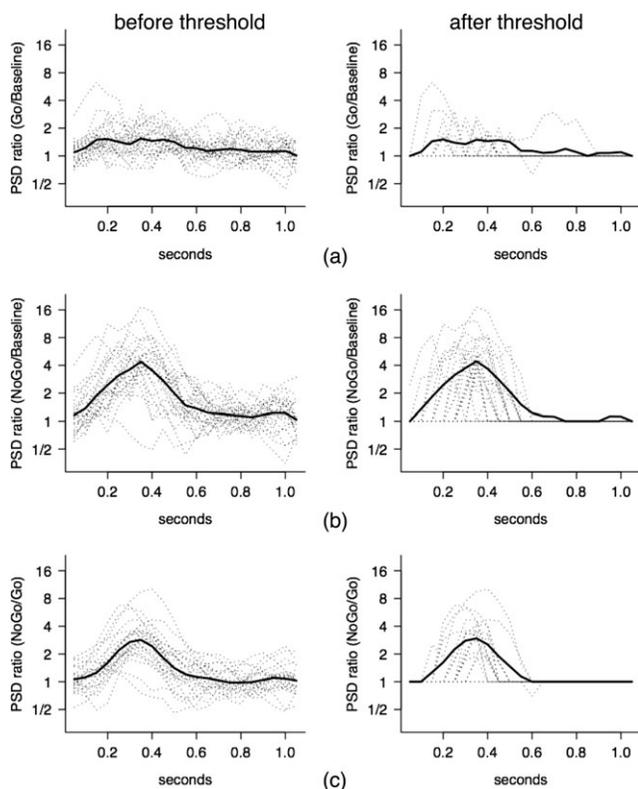
Figure 4.

(**a**) PSD estimates in each of the three task conditions—base-line, Go, and NoGo—from the peak electrode (FCz) averaged over part of the theta frequency range 5 to 6 Hz. (**b**) PSD estimates in each of the three task conditions from the frontal electrode (FPz) averaged over the frequency range 18 to 19 Hz. Each subject is represented as a dotted line, and the group average is shown as solid lines. Note the increase in PSD for theta oscillations in the NoGo condition from the peak electrode 0.35 s from stimulus onset (upper right panel), which is a pattern shared by most subjects. Note also that there is no effect of task condition on PSD in the frontal electrode at 18 to 19 Hz but that a single subject is substantially lower than the rest in the NoGo condition during the entire epoch (lower right panel).

model. Four components are retained in Figure 8a—two delta band components and two theta band components; five components are retained in Figure 8b—the same delta components, a theta component, and two beta components (16–25 Hz and 10–16 Hz). Again, as predicted from Figure 7b, many of these frequency components retained in Figure 8b are derived entirely from the single subject.

Following the second level of sequential PCA to obtain the subspace of spatial components and temporal scores, we derive an easily interpretable display of the increase in PSD due to the NoGo task, which agrees nicely with the results found in the literature covering simple inhibition tasks. By (1) including subjects in a single statistical model and (2) properly separating the variance components from the several sources of variability, we obtain Figure 9. We see succinctly and unequivocally the group finding that NoGo elicits an increase in PSD beyond the slight increase due to Go alone in the frontal midline electrodes, centered at FCz. At this location we see that the increase occurs in the theta band with peak at 0.35 s and at a delta frequency range with peak at 0.4 s from the onset of stimulus. In addition, the increase in delta band extends slightly toward the parietal electrodes bilaterally.

Conversely, the original STAT-PCA procedure leaves subjects in the sequential PCA, and, by the vagaries of a single subject in our sample, yields components dominated by that subject. This masks the group pattern and leaves ambiguous which are group results and which are single individual results. Figures 10 and 11 reveal this ambiguity. The two delta components are identical to those found from the full GLM as are the peaks at electrode FCz and 0.4 s; however, three additional spatial components from the second level are derived from the "delta-branch" of the first level of sequential PCA. All three come from one subject only, revealing the 0.6-s peak in the decrease of PSD at occipital and parietal electrodes. The theta band result is also recovered (Fig. 11) at FCz but peaks a little sooner (0.25 s) due to a larger increase in the PSD ratio from another single subject at 0.25 s. The high beta band decrease in the two frontal electrodes, due to the single subject, is clearly seen by PCA, and the temporal scores mimic this subject's time-course at those electrodes and frequencies (see Figs. 4c and 6c). Finally, the lower beta band component yields three spatial components from the same subject with a decrease in PSD ratios, peaking near 0.7 s in the occipital and left temporal electrodes, then 0.8 s on the right.
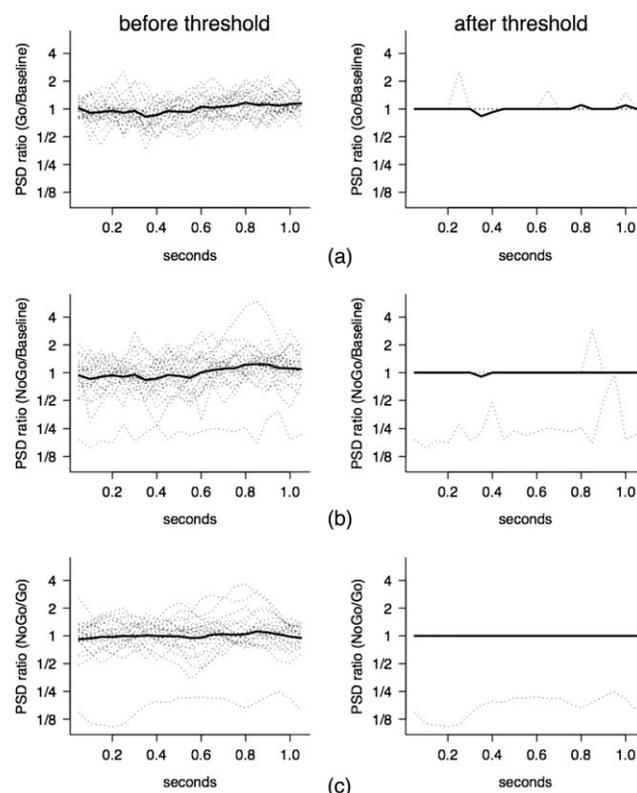
Figure 5.

Task condition contrasts in the peak electrode (FCz) before and after statistical thresholding for (**a**) Go versus baseline, (**b**) NoGo versus baseline, and (**c**) NoGo versus Go at the theta band frequency range. Using Model 1, only the group averages (solid lines) that survive the inference threshold are passed to PCA. In the original STAT-PCA procedure all subjects (dotted lines) that survive their individual respective thresholds are passed to PCA. In this case there is good agreement among most subjects, and the average captures the essential information to be passed to PCA.

DISCUSSION

The statistical models presented in this article impart two important advantages to analyses of event-related spectral perturbations. The first is purely statistical. For experimental designs having several sources of variability, statistical power is maximal when variance components in the mixed model are appropriately separated in tests of condition differences. The second follows serendipitously for sequential PCA. When subjects are included in the statistical model and, therefore, not in the PCA, the results of PCA more accurately reflect the group behavior rather than individual outlying subjects.

The original motivation for STAT-PCA in Ferree et al. [2009] was to achieve some robustness in isolating task related activation by utilizing PCA descriptively only after the inference stage (STAT). This was accomplished in large

measure not because subject-level variability was accounted for at the inference stage, but because subject-level variability was ignored at the inference stage. Tests for condition effects were applied to individuals, a separate set of voxel-level tests for each subject. In effect, this strategy transfers subject-level variability from the inference stage to the descriptive PCA stage. Since, in this context, PCA is intended to compactly display the results of the statistical inference, a desirable feature should be that the display "honors" the group behavior. If, however, subject-level variability in spectral, spatial, and/or temporal specificity occurs due to a small number of subjects, the PCA may incorporate the subject-level variance in the

Figure 6.

Task condition contrasts in the frontal electrode (FPz) before and after statistical thresholding for (**a**) Go versus baseline, (**b**) NoGo versus baseline, and (**c**) NoGo versus Go at the frequency range 18 to 19 Hz. One would not expect a response to the task conditions at this location and frequency range. Therefore, we would expect none of the average PSD ratios to be passed to PCA. This is indeed the case using Model 1 to account for intersubject variance, shown by flat solid lines in the right column for each of the three separate contrasts. However, if subjects are included in the PCA, the single outlying subject surviving the threshold (dotted line in right column) will be passed to PCA and will contribute significantly to the resulting spectral, spatial, and temporal components for both the NoGo versus baseline and NoGo versus Go contrasts.

Anterior

L   R

Posterior
(a)

Anterior

L   R

Posterior
(b)
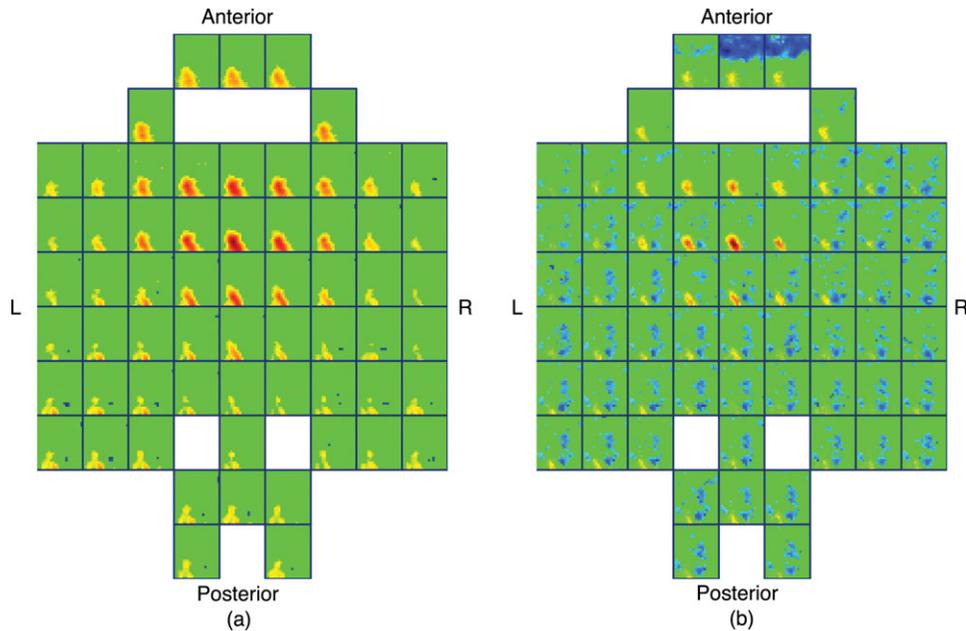
Figure 7.

Sixty-two-channel flattened layouts matching approximately the layout of the electrode cap. Each channel shows the thresholded PSD ratios of the NoGo response relative to the Go response for each frequency/time voxel, where frequencies (1–25 Hz) are represented vertically and the temporal epoch (0–1 s) is represented horizontally—bottom to top and left to right, respectively. (**a**) Shows the main findings using Model 1, where subjects are included in the GLM. The largest NoGo response, measured by the largest PSD ratios (shown in red color), occurs in FCz and neighboring electrodes at two frequency ranges: 1 to 4 Hz peaking 0.4 s and 5 to 8 Hz peaking 0.35 s from stimulus onset. (**b**) Shows the same results from the original STAT-PCA procedure. The group finding is recovered, but a single subject contributes very low PSD ratios (shown in blue) not only in FPz and FP2 at higher frequencies, as seen in Figures 4 and 6, but also in several frequency bands in a majority of electrodes near 0.6 s from stimulus onset.

form of a dominant eigenvalue and corresponding factor loadings. If this occurs, then the PCA is displaying not group-level inference, but a host of individual inferences.

A reliable display of group-level inference by PCA mandates that group-level inference occurs at the first stage (STAT). Clearly, subjects should be included in the statistical model. However, when subjects are included, we incorporate additional sources of variability to the model and, therefore, proper accommodation and separation of variance components is essential to maximize statistical power. Model 1 and Model 2 accomplish both so that PCA can reduce the large volume of results from correlated voxel-level inference and display only the essential features of group patterns of activation.

We have presented these statistical models within the framework of STAT-PCA as developed in Ferree et al. [2009] because our group had a motivation to improve the robustness of the procedure and increase the flexibility by making the statistical inference (STAT) as general as possible. It should be emphasized, however, that the two parts of STAT-PCA are independent of each other, with "STAT" comprising statistical inference only and "PCA" comprising the method by which the large volume of results are displayed. The statistical models presented in this paper
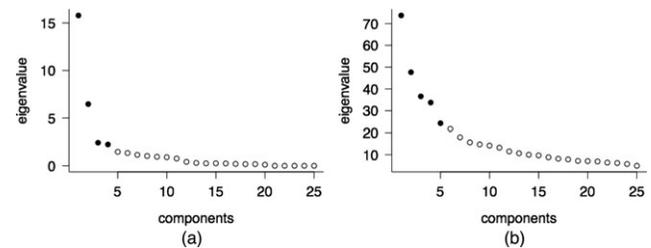
Figure 8.

Scree plots from the first level of sequential PCA. Four components in (**a**) are retained (filled circles) from parallel analysis following statistical inference using the full GLM (Model 1). Low frequencies (1–4 Hz) yield the two largest components, and the theta frequency range (5–8 Hz) yields the last two retained components. In (**b**) the original STAT-PCA procedure retains five components following parallel analysis: the two largest in the range 1 to 4 Hz, the next at 5 to 6 Hz, followed by one at 16 to 25 Hz, and the last between 10 and 16 Hz. In (b) all but the second component is determined either partially or completely by the single subject (see Figs. 10 and 11).
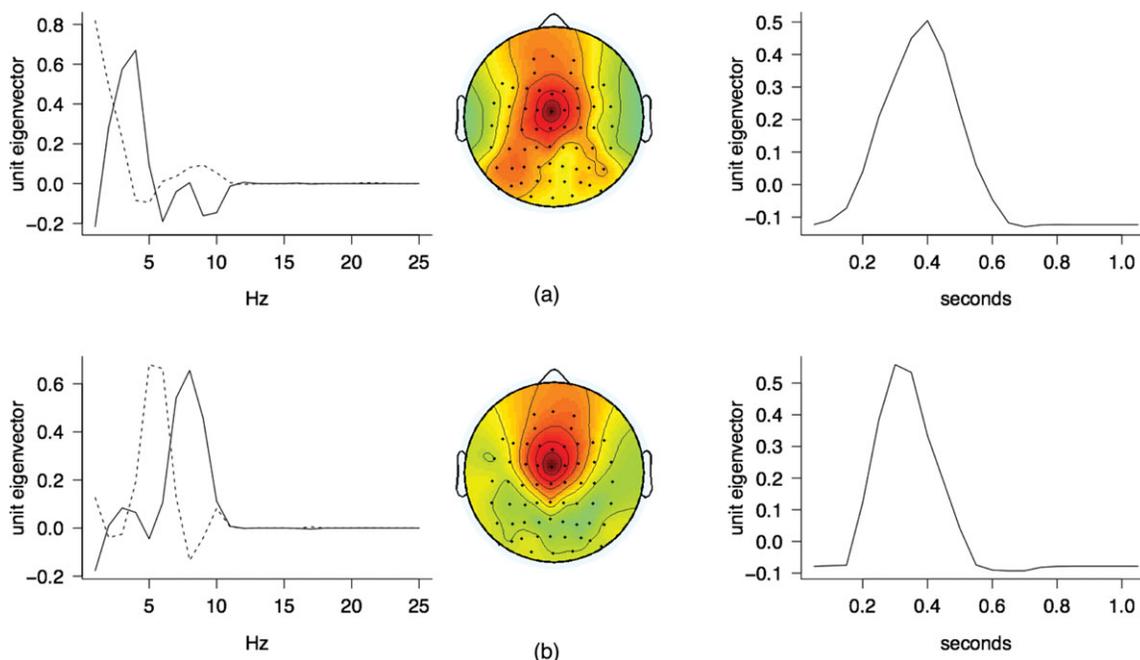
Figure 9.

PCA results following the inference stage using Model 1. Spectral components (left), spatial components (middle), and temporal components (right) confirm the main finding that the NoGo task results in an increase in PSD relative to the Go task at two frequency ranges centered at FCz: (**a**) two components in the delta band (with some power at posterior electrodes), peaking 0.4 s postonset; (**b**) two components in the theta band, peaking 0.35 s postonset. Note that the PCA distills the findings in Figure 7a into succinct easily interpretable factors, and the temporal component in (b) matches the group average time course at FCz (5–6 Hz) in Figure 5c.

do not require a subsequent PCA (Fig. 7 is an example of a non-PCA display of results). We believe that these models would benefit many EEG studies and many fMRI studies, regardless of how an investigator would like to present the high-dimensional results obtained from the models. However, we also believe that for high-volume studies similar to the one presented in this paper, PCA is a very convenient and reliable method for distilling voxel-level results to a manageable and interpretable visual presentation.

One may argue that a much simpler model could be utilized to accommodate subjects in the "STAT" stage before PCA, thus reducing the computational burden of variance component estimation across STF voxels. For each subject and condition, one could first average across trials to obtain a reduced dependent variable set, so that each subject would have a single measurement per condition. Then difference measures—NoGo−Go, for example— could be modeled by a reduced GLM, yielding, for our response inhibition task, a paired *t*-test analysis. By differencing the pairs of task condition levels for each subject, the intersubject variance component is effectively removed in the same fashion as is accomplished by Model 1. An analysis based on this reduced model is equivalent to that based on Model 1 *if and only if* $t_{jk} = t$ for all j, k; $k = 2$

specifically; and σ_ε^2 is known. Although this reduced model may be more familiar in current practice, we would recommend against this approach for several reasons.

1. More commonly in EEG $t_{jk} \neq t$; i.e., there is a variable number of trials for each subject and condition. In the response inhibition task, for example, the Go condition occurs about four times as often as NoGo. Consequently, the dependent measures entering this reduced model would have unequal variances across conditions, in which case differencing the pairs of dependent variables would *not* remove the inter-subject variance component. Rather, a proportion of σ_b^2 would be contained in the variance of differences, increasing the denominator in a paired *t*-test. An analysis based on either Model 1 or Model 2 accommodates these unequal variances explicitly.

2. If $k > 2$, then several of these reduced models would be utilized implicitly in paired *t*-tests—one for each possible pairwise combination of condition levels, consequently reducing the statistical efficiency of parameter estimation. Model 1 or Model 2 utilizes all the data in a single model for any number of condition levels, maximizing statistical efficiency of parameter estimation.
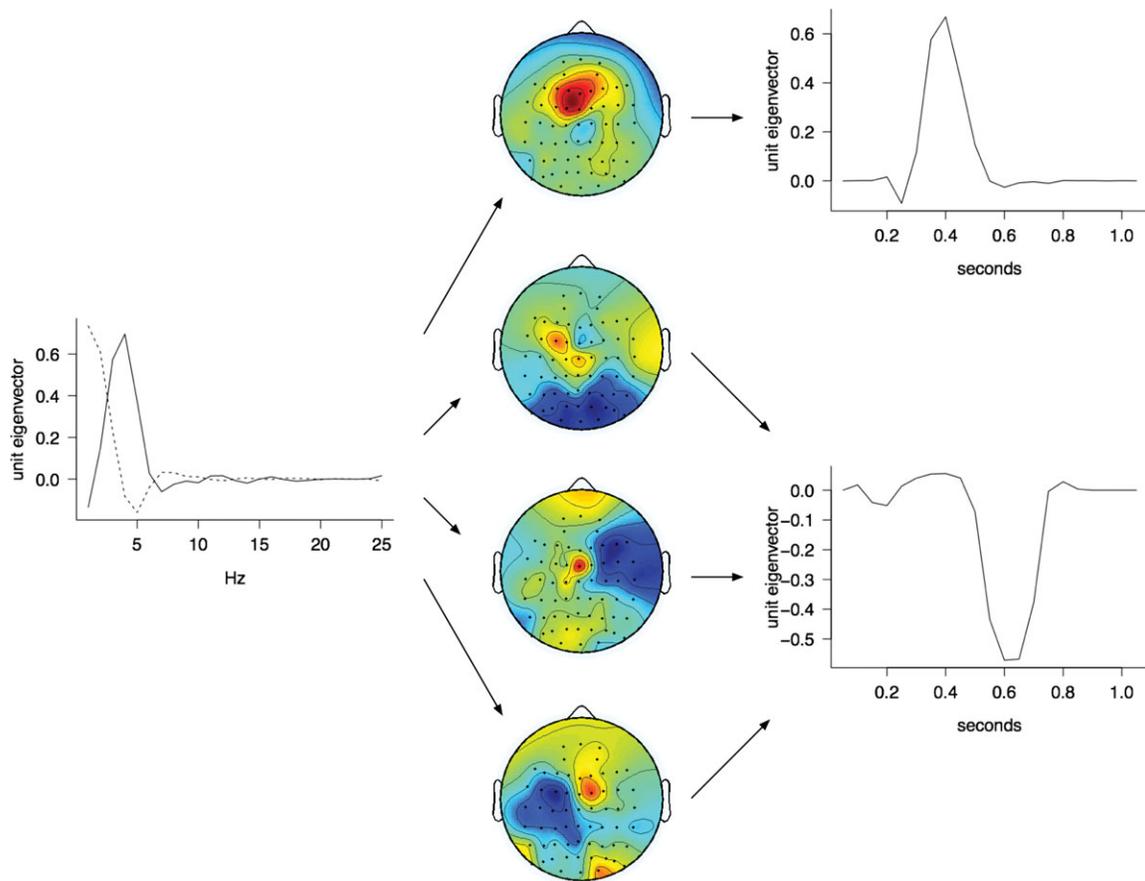
Figure 10.

PCA results from the first two retained spectral components of the original STAT-PCA procedure. The two delta frequency components yield four spatial components: spatial component 1 is centered at FCz and peaks 0.4 s postonset, which matches the group result for the NoGo increase in PSD; the other three spatial components are derived from the single subject, and reveal a 0.6-s peak for a NoGo decrease in PSD at posterior, right, and left electrodes, respectively. This result is consistent with the results shown in Figure 7b, strongly influenced by the single subject.

3. The reduced model could not assess the presence of a *subject × condition* interaction variance component, $\sigma^2_{b\gamma}$. If this variance component exists, the extrapolation of inference to a conceptual population of subjects would not be valid. Model 2 explicitly accounts for the possibility of sample heterogeneity due to more than one conceptual population.

4. Since σ^2_ε is not known, it must be estimated. In this reduced model the estimation occurs across the subject differences with $(n - 1)$ degrees of freedom. In Model 1 the estimation occurs across all subjects, conditions and trials with $nct - n - c + 1$ degrees of freedom. Not only does the latter confer much greater statistical efficiency, it also confers a practical advantage, if, as is common in human studies, the number of subjects enlisted is low. Variance estimates based on only five subjects, for example, would have only 4 degrees of freedom in the reduced model; the same estimates would have nearly 800 degrees of freedom based on Model 1, greatly reducing the variance of the variance estimates themselves. In addition, extrapolative inference is valid, even in smaller sample sizes, as long as $\sigma^2_{b\gamma} = 0$, and if the subjects can be considered a random sample.

5. Finally, Models 1 and 2 easily extend to accommodate multiple subject groups, as is common when, for example, patient groups are compared to a matched control group.

To illustrate this final point, Model 1 extends as

$$y_{ijk\ell} = \mu + \alpha_i + b_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + \varepsilon_{ijk\ell},$$

with $i = 1, \ldots, a$ groups. In addition to the fixed effect for *group*, a fixed effect for the *group × condition* interaction is included. Note that this interaction term is not the
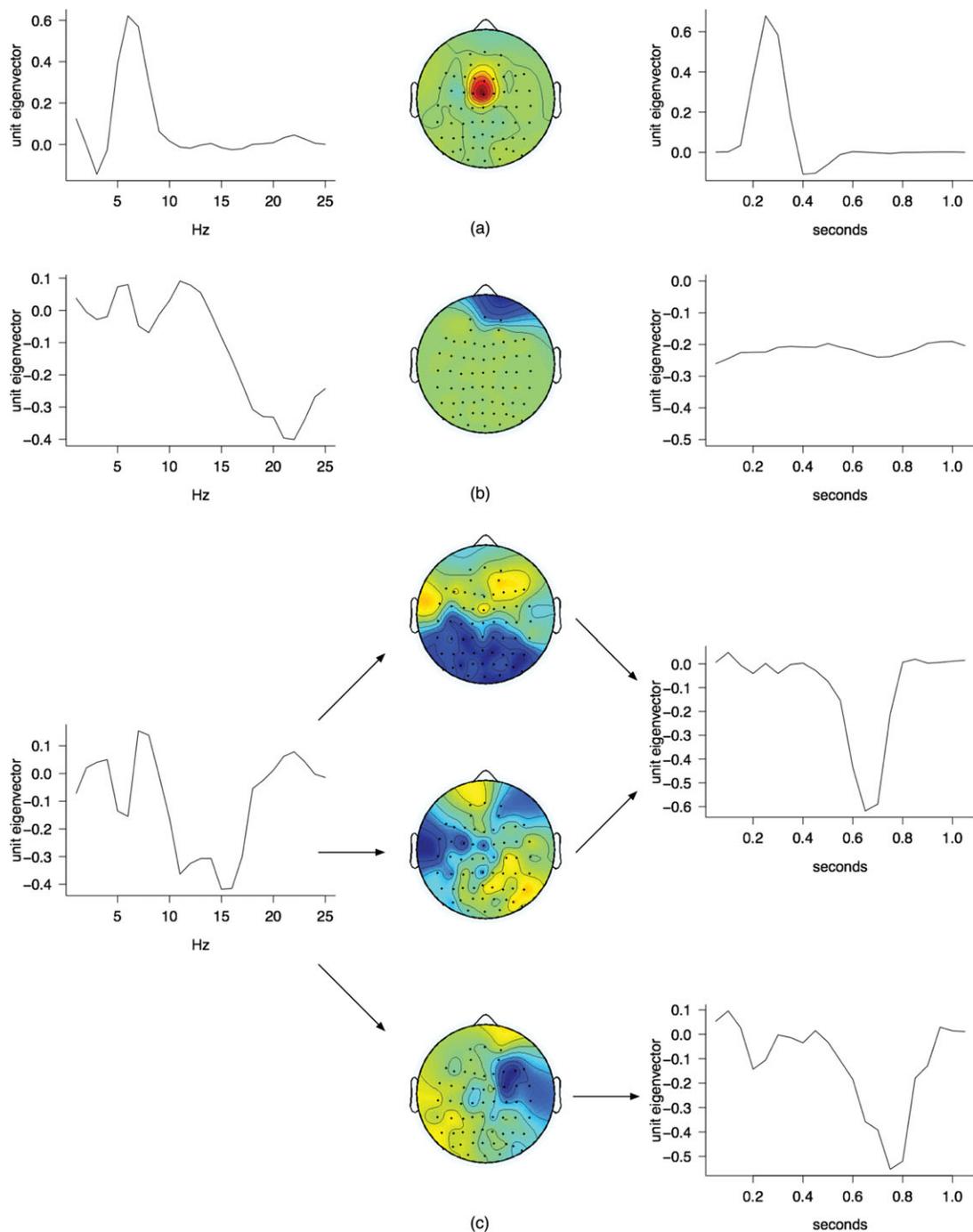
Figure 11.

PCA results from the last three retained spectral components of the original STAT-PCA procedure. Triplet (**a**) shows the group result for the expected NoGo response in the theta band centered at FCz, though one subject causes the peak increase in PSD to occur slightly sooner than expected at 0.25 to 0.3 s. The single subject dominates all other components: (**b**) reveals the decrease in PSD following the NoGo stimulus in the two frontal electrodes at high beta frequency range and at constant amplitude across the entire epoch, mirroring this subject's time course in Figures 4 and 6; (**c**) shows the same subject's influence in the range 10 to16 Hz, where the decrease in the PSD ratios occurs in the posterior and left-sided electrodes at 0.65 to 0.7 s postonset, then right-sided electrodes 0.75 to 0.8 s postonset. As in Figure 10, these components are derived from a single subject, masking the true group result.

interaction term that appears in Model 2. The important point in the context of this paper is the fact that tests on how the groups differ with respect to their task-condition differences *are also* relative to the intrasubject variance component. That is, the EMS for the fixed-effect interaction term and variances of interaction contrasts do not contain σ_b^2, the inter-subject variance component. If intersubject variances are included in tests of task-condition change-scores between groups, then statistical power suffers due to a reduction in relative efficiency.

When *is* σ_b^2 important? Testing the *group* effect within a single condition or averaged over the levels of condition in the extended model would indeed involve a linear combination of intra- and intersubject variances. Those tests must include σ_b^2 for valid inference. Examples would include tests of differences between two or more groups (patient vs. control, for example) only at the NoGo condition or only at baseline. Similarly, if one averaged over all task conditions for each group, then tested for group differences on the condition averages, one must include σ_b^2 for valid inference. Often, however, the *group × condition* interaction contrast is of primary importance. For example, would the increase in the power spectral density due to the NoGo condition, relative to the Go condition, occur to the same extent or occur at all in one group versus another? Those are group differences of change scores. To answer with good statistical power, these interaction contrasts should be tested appropriately relative to a function of σ_ε^2 only.

We conclude with a quotation from Ferree et al. [2009] that "... STAT-PCA provides a basis for the reduction of the results of time-frequency analysis of multielectrode EEG data into concise components that facilitate cognitive interpretation." In this article we have emphasized and developed further the statistical inference (STAT) step, not only to improve the robustness of the procedure, but to improve its capacity to generalize to any type of experimental design. In doing so, we have provided a general procedure for analyzing high volume EEG time-frequency data and, perhaps more importantly, provided the basis for reducing the need for multi-level approximations to mass-univariate procedures.

ACKNOWLEDGMENTS

REFERENCES

Benjamini Y, Hochberg Y (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300.

Bruin KJ, Wijers AA, van Staveren AS (2001): Response priming in a go/nogo task: Do we have to explain the go/nogo N2 in terms of response activation instead of inhibition? Clin Neurophysiol 112:1660–1671.

Casella G, Berger RL (1990): Statistical Inference. California: Duxbury Press.

Dien J, Spencer KM, Donchin E (2003): Localization of the event-related potential novelty response as defined by principle components analysis. Cogn Brain Res 17:637–650.

Dien J (2006): Progressing towards a consensus on PCA of ERPs. Clin Neuropsychol 117:695–707.

Donchin E (1966): A multivariate approach to the analysis of average evoked potentials. IEEE Trans Biomed Eng 13:131–139.

Ferree TC (2006): Spherical splines and average referencing in scalp electroencephalography. Brain Topogr 19:43–52.

Ferree TC, Brier MR, Hart J, Kraut MA (2009): Space-time-frequency analysis of EEG data using within subject statistical tests followed by sequential PCA. Neuroimage 45:109–121.

Horn JL (1965): A rationale and test for the number of factors in factor analysis. Psychometrika 30:179–185.

Jodo E, Kayama Y (1992): Relation of negative ERP component to response inhibition in a go/no-go task. Electroencephalogr Clin Neurophysiol 82:477–482.

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.

Kenward MG, Roger JH (1997): Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983–997.

Luu P, Tucker DM (2001): Regulating action: Alternating activation of midline frontal and motor cortical networks. Clin Neurophysiol 112:1295–1306.

Milliken GA, Johnson DE (2009): Analysis of Messy Data, Vol. 1: Designed Experiments. Florida: Chapman & Hall/CRC.

Searle S (1971): Linear Models., New York: John Wiley & Sons, Inc.

Smith JL, Johnstone SJ, Barry RJ (2007): Response priming in the Go/NoGo task: The N2 reflects neither inhibition nor conflict. Clin Neurophysiol 118:343–355.

Spencer KM, Dien J, Donchin E (1999): A componential analysis of the ERP elicited by novel events using a dense electrode array. Psychophysiology 36:409–414.

Storey JD (2003): The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat 31:2013–2035.

Yamanaka K, Yamamoto Y (2010): Single-trial EEG power and phase dynamics associated with voluntary response inhibition. J Cogn Neurosci 22:714–727.