Hierarchical Latent Dictionaries for Models of Brain Activation

Alona Fyshe
Machine Learning
Carnegie Mellon University

Emily Fox
Department of Statistics
The Wharton School
University of Pennsylvania

David Dunson
Statistical Science
Duke University

Tom Mitchell
Machine Learning
Carnegie Mellon University

Abstract

In this work, we propose a hierarchical latent dictionary approach to estimate the time-varying mean and covariance of a process for which we have only limited noisy samples. We fully leverage the limited sample size and redundancy in sensor measurements by transferring knowledge through a hierarchy of lower dimensional latent processes. As a case study, we utilize Magnetoencephalography (MEG) recordings of brain activity to identify the word being viewed by a human subject. Specifically, we identify the word category for a single noisy MEG recording, when only given limited noisy samples on which to train.

1 Introduction

The interpretation of noisy time series data is a challenge encountered in many application domains. From speech processing to weather forecasting, the regime of low signal to noise ratio (SNR) hinders data analysis. In such scenarios, replicates or *repeated trials* can improve the ability of an algorithm to uncover the underlying signal. Within the problem setting, a key challenge is how to fully leverage the multiple time series in order to optimally share knowledge between them. The problem is compounded when the time series is of high dimension and there are few replicates.

As a motivating example, consider Magnetoencephalography (MEG) recordings of brain activity (described further in Section 2). Due to the recording mechanism, the SNR is extremely low and recording many replicates of a given stimulus is a costly task. A further obstacle is the sheer dimensionality of the time series, typically on the order of more than 100 sensors recordings per time step. However, the close spatial proximity of the sensors leads to redundancies that can be harnessed in conjunction with the repeated trials. This situation is common to many high-dimensional time series domains.

Motivated by the structure of our high-dimensional time series, we propose a Bayesian nonparametric dynamic latent factor model (DLFM). A DLFM assumes that the non-idiosyncratic variations in our observations are governed by dynamics evolving in a lower dimensional subspace. To transfer knowledge between the multiple trials and better recover the signal from few noisy samples, we *hierarchically couple* the latent trajectories. To capture the MEG signal's long-range dependencies we take the latent trajectories, or *dictionary elements*, to be Gaussian process random functions. This hierarchical latent dictionary formulation is a main contribution of this paper.

In many application domains it is insufficient to assume that the *correlations* between the elements of the observation vector are static. For example, the spatial correlation of the MEG sensor recordings change as the co-activation pattern of brain regions evolves in time. In such cases, one needs a *heteroscedastic* model. Within the DLFM framework, this is achieved by extending the standard model to have a *time-varying* mapping from the lower dimensional subspace to the full observation space.

Though our model is general enough to be applied in many domains, we focus here on the task of predicting the category of word a person is viewing based on MEG recordings of their brain activity. We show a subject a set of concrete nouns (see Table 1), and collect multiple recordings of their brain activity for each word. We then wish to predict the word based on one low-SNR MEG recording.

Single trial MEG classification is an inherently challenging task due to large inter-trial variability and the susceptibility of MEG sensors to interference. Still, successful single trial analyses have been performed

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

Table 1: The 20 concrete nouns used in this experiment, sorted by category.

Animals	Buildings	Food	Tools
bear	apartment	carrot	chisel
cat	barn	celery	hammer
cow	church	corn	pliers
dog	house	lettuce	saw
horse	igloo	tomato	screwdriver

in the past, mostly through decompositional methods like principal component analysis [13] or discriminative classification algorithms [6, 18]. Our work aims to produce a generative model that characterizes the MEG signal's time-varying mean and covariance. A generative model allows us to predict not only what stimulus caused a specific MEG recording, but also what MEG signal (and thus what neuronal activation) would be observed in response to a given stimulus. The approach we develop in this paper is generic, but shows significant promise – it forms the foundation for more intricate future generative models that incorporate other characteristics of the MEG signal (e.g., frequency and phase, lagged correlation, sensor drift), more elaborate representations of the stimulus, and assumptions about the cognitive subprocesses that give rise to observed brain activity.

2 The Magnetoencephalography Data

When neurons in the brain fire in a coordinated fashion, a weak magnetic field can be detected outside of the skull. The MEG gradiometer measures the spatial gradient of this magnetic activity (i.e. the change in magnetic field strength in space) measured in Teslas per meter (T/m) [15]. Gradiometers are arranged within a helmet, at 102 locations around the head (Figure 5 illustrates the layout) [1]. As mentioned, the MEG signal is incredibly noisy, as is apparent in Figure 4. To increase the signal to noise ratio (SNR), researchers typically collect multiple trials (samples) of subjects performing a task (e.g. reading a word), and analyze the sample mean MEG signal over trials. While the maximum likelihood estimate (MLE) may perform well in scenarios with large amounts of data, time and subject fatigue constrains the number of trials that can be obtained. Thus, we seek a model that can efficiently learn the subtle signal from a few very noisy replicates.

MEG sensors produce redundant recordings of underlying cognitive processes; adjacent sensors are often highly correlated. For this reason techniques often seek to explain the data with a small number of latent sources (e.g. Equivalent Current Dipole (ECD) methods [14]). Recently, Bayesian approaches to source localization have been developed [16, 26, 30]. The success of such methods indicates that there is an accurate lower dimensional representation for the brain activity captured by MEG. The model described herein learns a lower dimensional representation of the observed MEG activity, but focuses on the accuracy of fit rather than the localization of the latent sources.

[1] Our MEG machine has three sensors at each helmet position: two gradiometers and one magnetometer. To reduce the dimensionality of the problem we consider only one gradiometer per helmet location.

3 Background

We provide a brief review of some key elements of our generative model outlined in Section 4: Gaussian processes and dynamic latent factor models.

Gaussian Processes A Gaussian process provides a distribution over real-valued functions $f : \mathbb{T} \to \mathbb{R}$, with the property that the function evaluated at any finite collection of points is jointly Gaussian. The Gaussian process, denoted $\mathrm{GP}(m, c)$, is uniquely defined by its *mean function* m and *kernel function* c. So, $f \sim \mathrm{GP}(m, c)$ if and only if

$$p(f(t_1), \ldots, f(t_n)) \sim N_n(\mu, K), \tag{1}$$

with $\mu = [m(t_1), \ldots, m(t_n)]$ and K the $n \times n$ *Gram matrix* with entries $K_{ij} = c(t_i, t_j)$. The properties (e.g., continuity, smoothness, periodicity, etc.) of functions drawn from a given Gaussian process are determined by the kernel function. One example kernel leading to smooth functions is the squared exponential kernel:

$$c(t, t') = d \exp(-\kappa ||t - t'||_2^2), \tag{2}$$

where d is a *scale* hyperparameter and κ the *bandwidth*, which determine the extent of the correlation in f over $\mathbb{T}$. See [21] for further details.

Factor Analysis A latent factor model assumes that the non-idiosyncratic variations in the observations are determined by a smaller collection of latent variables. Specifically,

$$y_i = \Lambda \eta_i + \epsilon_i, \quad \eta_i \sim N_k(0, I), \quad \epsilon_i \sim N_p(0, \Sigma_0), \tag{3}$$

where y_i is a p-dimensional observation, η_i is a k-dimensional *latent factor* with $k << p$, Λ is the *factor loadings* matrix, and ϵ_i is idiosyncratic Gaussian noise with $\Sigma_0 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Marginally, $y_i \sim N_p(0, \Sigma)$ with $\Sigma = \Lambda\Lambda' + \Sigma_0$. That is, the $p \times p$-dimensional covariance matrix which, in general, has $p(p+1)/2$ unique elements is assumed to have a decomposition in terms of a low rank component plus a diagonal matrix resulting in $p(k+1)$ unique elements. For large p domains, this approach represents a substantial reduction in parameters. Inferring the dimension of a latent space has been explored in many works including [2, 4, 17].

Dynamic Latent Factor Models The latent factor model of Equation (3) assumes that the latent factors are independent. Latent representations have also been explored in the time-series domain by assuming a latent factor *process*. Such dynamic latent factor models have a rich history. Typically, the dynamics of the latent factors are assumed to follow a simple Markov evolution with a time-invariant parameterization [19, 28]:

$$\begin{aligned} \eta_t &= \Gamma\eta_{t-1} + \nu_t, \quad \nu_t \sim N_k(0, I) \\ y_t &= \Lambda\eta_t + \epsilon_t, \quad \epsilon_t \sim N_p(0, \Sigma_0). \end{aligned} \tag{4}$$

Assuming a stationary process on η_t, the resulting model leads to $y_t \sim N_p(0, \Sigma)$ with $\Sigma = \Lambda\Sigma_\eta\Lambda' + \Sigma_0$. Here, Σ_η denotes the marginal covariance of η_t.

4 Hierarchical Latent Dictionary Learning

To cope with the high-dimensionality of the MEG time series, one might propose to use the dynamic latent factor model (DLFM) of Equation (4). However, such a model assumes Markov dynamics. Furthermore, the time-invariant parameterization leads to a homoscedastic model (i.e., the covariance does not evolve in time.) As we will see, these assumptions are insufficient to capture the key long-range dependencies and heteroscedasticity inherent in MEG data. Furthermore, we wish to develop a model that can share information between the noisy single trials. In this section, we develop a DLFM in which the latent dynamics evolve *nonparametrically* and the latent trajectories, or *latent dictionary elements*, are hierarchically coupled. Furthermore, we allow for time-varying factor loadings, leading to a heteroscedastic model.

The model outlined herein, and depicted in Figure 1,2 and 3, is generic to a variety of signal types and predictor spaces (e.g., weather patterns evolving *spatially*), but for ease of exposition we will restrict our description to the MEG application of interest. As such, we take the predictor space to be time and the collection of signals to be single trials of MEG responses to some single word stimulus $w \in \mathcal{W}$. Assume we use p MEG sensors. We use τ to denote continuous time, and t is the discrete time index. All variables, and their dimensions, appear in Table 3 of the Appendix.

Observation Model We take each MEG signal to be Gaussian distributed with time-varying mean and covariance. The mean is assumed to be trial-specific whereas the covariance is taken to be shared between trials of a given word stimulus. Let $y_t^{(i,w)} \in \mathbb{R}^p$ be a p-dimensional MEG response at time τ_t for the ith single trial of word stimulus w. Conditioned on the trial-specific mean process $\mu^{(i,w)}(\cdot)$ and word-specific

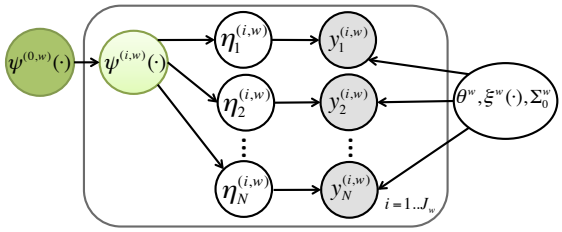


Figure 1: A graphical representation of the model outlined in Section 4 for one word w, and its trials $i = 1 \dots J_w$. The mean of the child latent process $\psi^{(i,w)}$ (light green) is given by the parent latent process $\psi^{(0,w)}$ (dark green). The latent factors $\eta_t^{(i,w)}$ are centered about $\psi^{(i,w)}$. The marginal mean of $y^{(i,w)}$ is governed by Θ^w, ξ^w and $\psi^{(i,w)}$ while the covariance of $y^{(i,w)}$ is governed by Θ^w, ξ^w and Σ_0^w as in Equation 10.

covariance process $\Sigma^{(w)}(\cdot)$, our model specifies

$$y_t^{(i,w)} \sim N_p(\mu^{(i,w)}(\tau_t), \Sigma^{(w)}(\tau_t)), \tag{5}$$

independently for each word w, trial i and time t.

Classical approaches focus on modeling the time-varying mean $\mu(\tau)$, which captures changes in levels of MEG signal *magnitude*. Typically, a homoscedastic approach is taken in which the covariance $\Sigma^{(w)}$ is static. However, as we show in Section 7, capturing the time-varying *correlations* amongst the sensors is key. As such, we develop a *heteroscedastic* approach with time-varying covariance $\Sigma^{(w)}(\tau)$. Additionally, allowing for a time-varying covariance matrix provides insight into the changing coordination of neuronal activity between brain areas, an idea of *dynamic functional connectivity* fundamentally different from those explored via network analyses [5]. Our dictionary approach, outlined below, also has the ability to automatically infer the frequency of latent components, a characteristic of MEG data currently of great interest [8].

Our specific choice of trial-specific mean, but shared covariance is justified experimentally by the fact that the trial-to-trial variation of the raw activity is significant (see Figure 4) while the co-activation of brain regions is fairly similar. In many application domains, such an assumption is reasonable and dramatically reduces the parameterization of the model.

Building a Hierarchy To fully leverage the signal in each single trial, we construct a *word-specific hierarchical latent model* (See Figure 1). Specifically, we hierarchically couple the trial-specific means for a given word stimulus w. One could take each of the p components of $\mu^{(i,w)}(\cdot)$ to be a draw from a Gaussian process centered about a word-specific *global* mean process $\mu^{(0,w)}(\cdot)$ [1]. However, this approach does not harness the inherent redundancy in our high-dimensional

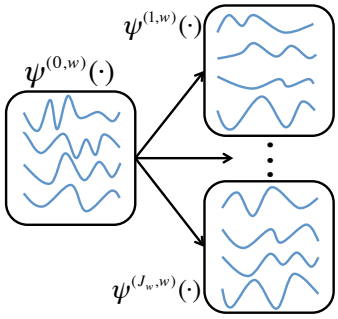


Figure 2: The hierarchy of latent processes. Each child $\psi^{(i,w)}$ is centered around the parent $\psi^{(0,w)}$, allowing sharing of information. See Equation 8.

observation vector (e.g., spatially co-located sensors measure similar neuronal activity). Instead, our goal is to discover a *latent dictionary* of Gaussian processes.

To cope with the high-dimensionality of the data, and as motivated by our MEG application, we consider a *semi-parametric* DLFM. Recall from Section 3 that a DLFM assumes that the non-idiosyncratic variations in our observations are governed by dynamics evolving in a lower dimensional subspace. To model long-range dependencies, we assume that the latent factors $\eta_t^{(i,w)} \in \mathbb{R}^k$ evolve *nonparametrically*, while heteroscedasticity is captured via a time-evolving factor loadings matrix $\Lambda^{(w)}(\tau_t) \in \mathbb{R}^{p\times k}$. Specifically, taking $k << p$, we propose

$$\begin{aligned}\eta_t^{(i,w)} &= \psi^{(i,w)}(\tau_t) + \nu_t^{(i,w)}, \quad \nu_t^{(i,w)} \sim N_k(0, I_k)\\ y_t^{(i,w)} &= \Lambda^{(w)}(\tau_t)\eta_t^{(i,w)} + \epsilon_t^{(i,w)}, \quad \epsilon_t^{(i,w)} \sim N_p(0, \Sigma_0^{(w)}).\end{aligned} \tag{6}$$

Here, the factor loadings $\Lambda^{(w)}(\tau_t)$ (and thus the induced covariance) are shared between trials of a given word w. Figure 3 illustrates the (time-varying) projection of the noisy latent functions up to the observation space. As in Section 3, the idiosyncratic noise covariance is diagonal and, furthermore, time-invariant: $\Sigma_0^{(w)} = \text{diag}(\sigma_{w,1}^2, \sigma_{w,2}^2, \ldots \sigma_{w,p}^2)$.

The evolution of the latent factors $\eta_t^{(i,w)}$ is governed by a collection of k *latent dictionary functions*, each of which is a random function:

$$\psi^{(i,w)}(\tau) = [\psi_1^{(i,w)}(\tau), ..., \psi_k^{(i,w)}(\tau)] \tag{7}$$

To capture a smoothly evolving latent factor mean with correlations over potentially distant time points, we take the $\psi_j^{(i,w)}(\cdot)$ to be Gaussian process random functions with squared exponential correlation functions. To share structure between the single trials within this latent space, we *hierarchically couple* the trial-specific latent dictionary functions as

$$\psi_j^{(0,w)}(\cdot) \sim \text{GP}(0, c_0) \quad \psi_j^{(i,w)}(\cdot) \sim \text{GP}(\psi_j^{(0,w)}(\cdot), c_1), \tag{8}$$

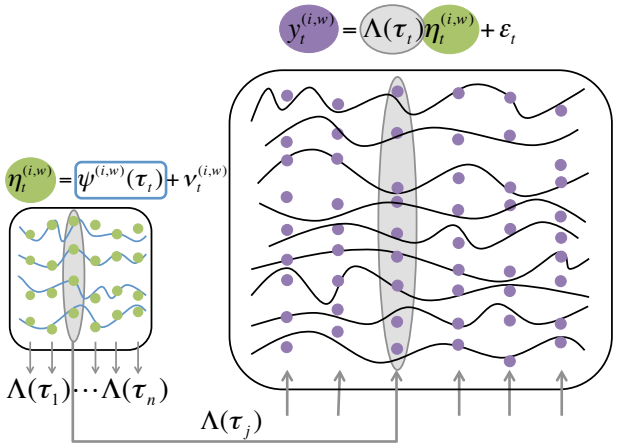


Figure 3: The latent model for one trial i of word w. Independent noise is added to the child latent process $\psi^{(i,w)}$ (see Figure 2) to produce $\eta_t^{(i,w)}$ (green circles). Then, $\eta_t^{(i,w)}$ is projected to the full dimensional space via the time-varying $\Lambda(\tau)$ (grey). The observed data ($y_t^{(i,w)}$, purple circles) is that projection plus sensor-specific noise (ϵ).

with $c_i(\xi, \xi') = d_i \exp(-\kappa||\xi - \xi'||_2^2)$ for $i = 0, 1$. The choice of d_0 and d_1 controls the amount of prior variation of the global functions from 0 and the trial-specific functions from the global functions, respectively. Conceptually, the child processes $\psi_j^{(i,w)}$ are centered about the parent process $\psi_j^{(0,w)}$ (See Figure 2). During model fitting, the child processes share information through the parent process.

We would also like to have the factor loadings $\Lambda^{(w)}(\tau)$ vary smoothly with time in order to capture the key changes in correlation amongst the p MEG sensors. However, treating the elements of $\Lambda^{(w)}(\tau)$ in a similar fashion to those of $\psi^{(i,w)}(\tau)$ requires defining $p \times k$ latent Gaussian random functions. For large p, this is methodologically and computationally impractical. To reduce the dimension further, we employ the Bayesian nonparametric heteroscedastic regression of [7] in which $\Lambda^{(w)}(\cdot)$ is modeled as a weighted combination of a much smaller set of $L \times k$ *latent covariance dictionary functions* $\xi_{lk}^{(w)}(\cdot)$. Specifically we model

$$\Lambda^{(w)}(\tau) = \Theta^{(w)}\xi^{(w)}(\tau) \tag{9}$$

with $\xi_{\ell k}^{(w)}(\cdot) \sim \text{GP}(0, c_0)$ and $\Theta^{(w)} \in \mathbb{R}^{p\times L}$ distributed according to the shrinkage prior of [2]. Specifically, a conditionally Gaussian prior is induced on Θ that flexibly shrinks the elements $\Theta_{j\ell}^{(w)}$ toward zero increasingly as ℓ grows, as controlled by a set of latent precision parameters (see Appendix). Since the ℓth column of $\Theta^{(w)}$ weights the ℓth row of covariance dictionary functions in $\xi^{(w)}$, our choice of shrinkage prior discounts the importance of the Gaussian process random functions with higher row index.

Finally, we employ the usual conditionally conjugate

inverse gamma prior on the diagonal elements of $\Sigma_0^{(0)}$: $\sigma_{j,w}^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$, independently for each $j = 1, \ldots, p$.

Marginalizing the idiosyncratic noise terms $\nu_t^{(i,w)}$ and $\epsilon_t^{(i,w)}$ induces the following mean and covariance structure of the observed signal $y_t^{(i,w)}$:

$$\begin{aligned} \mu^{(i,w)}(\tau_t) &= \Theta^{(w)}\xi^{(w)}(\tau_t)\psi^{(i,w)}(\tau_t) \\ \Sigma^{(w)}(\tau_t) &= \Theta^{(w)}\xi^{(w)}(\tau_t)\xi^{(w)}(\tau_t)'\Theta^{(w)\prime} + \Sigma_0^{(w)}. \end{aligned} \tag{10}$$

As the decomposition in Equation (10) is not unique, we learn an over-complete dictionary in which there are infinitely many ways to characterize the mean and covariance functions. For inference tasks based on the induced mean and covariance processes, as examined in Section 7, identifiability of a unique decomposition is not necessary. Avoiding identifiability constraints leads to computational and modeling advantages [11]. Although our semi-parametric DLFM captures non-Markovian dynamics and heteroscedasticity, the Gaussian process formulation comes at a computational cost relative to the traditional DLFMs of Section 3.

5 Related Work

Our hierarchical latent dictionary model takes inspiration from a wide body of work. Our model is similar to the Bayesian nonparametric heteroscedastic regression model of [7], but incorporates the idea of a latent hierarchy of Gaussian processes, the children of which need not be centered around 0. This hierarchy is central to our framework, and key in sharing information between single trials. A (non-hierarchical) Gaussian process latent factor model was considered in [25], but within a homoscedastic framework without a time-varying factor loading matrix. In [12] a hierarchy of Gaussian process parameters is utilized, but the Gaussian processes do not operate in a lower dimensional latent space.

Employing Gaussian processes for heteroscedastic modeling was also proposed in [9, 29], though based on an alternative covariance decomposition. We appeal to the framework of [7] because of the ready interpretation in terms of latent factor models.

6 Posterior Computations

Due to the form of our model, posterior computation can rely on a Gibbs sampler, which alternately samples parameters in blocks from standard distributions and is implemented in parallel for the different word-specific models. Sampling steps are fully detailed in the Appendix. The steps are similar to [7]; here we focus on the sampling of latent dictionary elements. Note that the sampler is run independently for each word w. We assume J_w single trials of word w, each of length n. The derivations of the conditional posteriors harness the fact the observation model of Equation (6) can be rewritten as

$$y_{t,j}^{(i,w)} = \sum_{m=1}^{k} \eta_{t,m}^{(i,w)} \sum_{\ell=1}^{L} \Theta_{j\ell}^{(w)} \xi_{\ell m}^{(w)}(\tau_t) + \epsilon_{t,j}^{(i,w)} \tag{11}$$

Block-Sample $\{\boldsymbol{\psi^{(i,w)}, \nu_{1:n}^{(i,w)}}\}$: For each single trial i, we sample each child process $\psi^{(i,w)}$ from its conditional posterior marginalizing $\nu_t^{(i,w)}$ and cycling through each latent dictionary function $\psi_\ell^{(i,w)}$. We then treat $\nu_{1:n}^{(i,w)}$ as auxiliary variables that are imputed conditioned on $\psi^{(i,w)}$. From Equation (6), we have that marginally

$$\begin{aligned} y_t^{(i,w)} &= \Lambda^{(w)}(\tau_t)\psi^{(i,w)}(\tau_t) + \omega_t^{(i,w)} \\ \omega_t^{(i,w)} &\sim N_p(0, \Sigma^{(w)}(\tau_t)) \\ \Lambda^{(w)}(\tau_t) &= \Theta^{(w)}\xi^{(w)}(\tau_t) \end{aligned} \tag{12}$$

and $\Sigma^{(w)}$ as in Equation (10). Recall that $\psi^{(i,w)}(\cdot) \sim \text{GP}(\psi^{(0,w)}(\cdot), c_1)$, so we could sample the parent process $\psi^{(0,w)}$ and condition on that value when sampling child processes $\psi^{(i,w)}$. However, the parent process $\psi^{(0,w)}$ can be analytically marginalized and we can instead use the predictive distribution of $\psi_\ell^{(i,w)}(\tau_{1:n})$ given $\{\psi_\ell^{(j,w)}(\tau_{1:n}), j \neq i\}$ (the dictionary function evaluated at $\tau_1, \ldots, \tau_n$). Standard Gaussian results imply that the conditional posterior of $\psi_\ell^{(i,w)}(\tau_{1:n})$ is

$$\psi_\ell^{(i,w)}(\tau_{1:n}) \mid y_{1:n}^{(i,w)}, \psi_{\setminus\ell}^{(i,w)}(\tau_{1:n}), \psi^{(0,w)}(\tau_{1:n}), \Omega^{(w)} \sim N_n\left(\tilde{\Sigma}_\psi^{(w)} \begin{bmatrix} [\Lambda^{(w)}(\tau_1)]_{\cdot\ell}'\Sigma^{-(w)}(\tau_1)\tilde{y}_1^{(i,w)} + \tilde{\mu}_1 \\ \vdots \\ [\Lambda^{(w)}(\tau_n)]_{\cdot\ell}'\Sigma^{-(w)}(\tau_n)\tilde{y}_n^{(i,w)} + \tilde{\mu}_n \end{bmatrix}, \tilde{\Sigma}_\psi^{(w)}\right)$$

where $\Omega^{(w)}$ denotes $\{\Theta^{(w)}, \xi^{(w)}, \Sigma_0^{(w)}\}$ and

$$\begin{aligned} \tilde{\Sigma}_\psi^{-(w)} = \tilde{K}^{-1} + \\ \text{diag}\Big([\Lambda^{(w)}(\tau_1)]_{\cdot\ell}'\Sigma^{-(w)}(\tau_1)[\Lambda^{(w)}(\tau_1)]_{\cdot\ell}, \ldots \\ [\Lambda^{(w)}(\tau_n)]_{\cdot\ell}'\Sigma^{-(w)}(\tau_n)[\Lambda^{(w)}(\tau_n)]_{\cdot\ell}\Big). \end{aligned} \tag{13}$$

Here, $\tilde{\mu}$ and $\tilde{K}$ are the mean and covariance of the predictive distribution of $\psi_\ell^{(i,w)}(\tau_{1:n})$ given $\{\psi_\ell^{(j,w)}(\tau_{1:n}), j \neq i\}$ marginalizing $\psi_\ell^{(0,w)}(\tau_{1:n})$, $\tilde{y}_t^{(i,w)} = y_t^{(i,w)} - \sum_{(r \neq \ell)}[\Lambda^{(w)}(\tau_t)]_{\cdot r}\psi_r^{(i,w)}(\tau_t)$, and $\psi_{\setminus\ell}^{(i,w)}$ is the set of latent dictionary functions $\psi_j^{(i,w)}$ for $j \neq \ell$. In the Appendix we provide an alternative derivation without marginalizing the parent processes $\psi^{(0,w)}$, which allows for parallelization across single trials. Such parallel sampling is advantageous when there are many single trials.

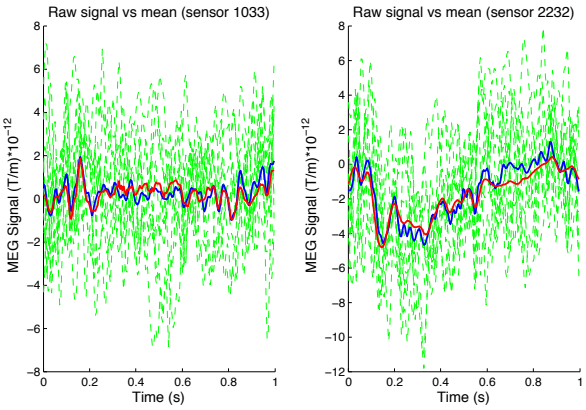


Figure 4: Examples of the estimated mean calculated using MLE (blue) and our hierarchical model ($\mu^{(0,w)}$, red) with the corresponding raw signal (green) for Subject 1, two MEG sensors and 10 trials of the word *hammer*. Note the extreme noise of the raw MEG signal and the smoothness of $\mu^{(0,w)}$ compared to μ_{MLE}.

Conditioned on $\psi^{(i,w)}$, we independently impute $\nu_t^{(i,w)}$ for each t. Such parallel sampling is advantageous when there are many single trials, and allows for block moves. Consider $\tilde{y}_t^{(i,w)} = y_t^{(i,w)} - \Lambda^{(w)}(\tau_t)\psi^{(i,w)}(\tau_t) = \Lambda^{(w)}(\tau_t)\nu_t^{(i,w)} + \epsilon_t^{(i,w)}$. Then, straightforwardly,

$$\nu_t^{(i,w)} \mid y_t^{(i,w)}, \psi^{(i,w)}(\tau_t), \Theta^{(w)}, \xi^{(w)}(\tau_t), \Sigma_0^{(w)} \sim N_k\left(\Phi^{(w)}\Lambda^{(w)'}(\tau_t)\Sigma_0^{-(w)}\tilde{y}_t^{(i,w)}, \Phi^{(w)}\right). \quad (14)$$

Where $\Phi^{(w)} = \left(I + \Lambda^{(w)'}(\tau_t)\Sigma_0^{-(w)}\Lambda^{(w)}(\tau_t)\right)^{-1}$. Subsequent steps proceed as in [7] (see Appendix).

7 MEG Word Category Classification

Recall the goal outlined in Section 1: we wish to identify the word a subject is viewing based on a single noisy MEG recording. Our MEG data was recorded while two subjects viewed 20 stimuli describing concrete nouns (both the written noun and a representative line drawing), with 20 interleaved trials per word. These concrete nouns fall into four categories: animals, buildings, food and tools (see Table 1). In terms of the model outlined in Section 4, trial i of word w for time points $t = 1 : n$ is denoted $y_{1:n}^{(i,w)}$, and $y_t^{(i,w)}$ is a $p = 102$ dimensional vector.

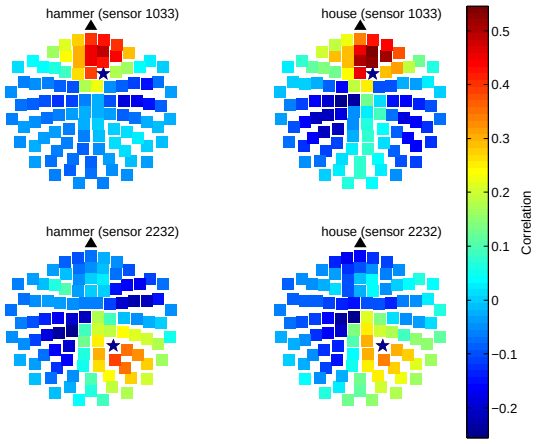
Our dataset consists of MEG single trials recorded from two subjects. Independently for each subject we trained one hierarchical latent model per word, using 15 trials as training data, and 5 trials per word for testing. This resulted in 5 subject-specific models per word category (20 word models per subject) and 100 test instances per subject. We ran the sampler for 3000 iterations. To evaluate performance we use samples thinned to 100 from iterations 2500 : 3000. A full list of settings appears in the Appendix. While the word models were trained on the full 1.7 seconds of MEG signal, we use only the 1 second after word stimulus to score models. The resulting parent mean $\mu^{(0,w)}(\tau) = \Theta^{(w)}\xi^{(w)}(\tau)\psi^{(0,w)}(\tau)$ is shown in Figure 4, along with the corresponding noisy MEG signal. Note the extreme noise of the raw MEG data and the smoothness of $\mu^{(0,w)}(\tau)$ due to its GP formulation.

To assess convergence of the sampler, we performed the modified Gelman-Rubin diagnostic [3] on the MCMC samples of the mean and variance terms for four sensors at five time points for one word model. We chose one sensor per brain lobe, with equal distribution between hemispheres and five equally spaced time points between 151 ms and 551 ms following stimulus onset. These sensors and time points were also used to create Figures 4 and 5. We then calculated the potential scale reduction factor (PSRF) [10] for these 20 sensor/time points combinations across three chains. The 20 mean variables had no PSRF above 1.2 (maximum PSRF = 1.04). In two cases the variance had PSRF above 1.2. However, the maximum PSRF for the variance was only 1.26. This analysis indicates convergence of the sampler. Anecdotally, we found the mean converges very quickly (within 500 iterations), but fitting the covariance requires more iterations.

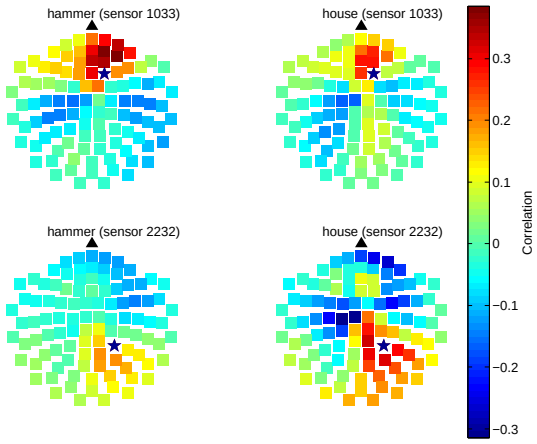
Figure 5(a) shows a representation of the posterior mean correlation at two time points as computed from samples of $\Sigma^{(w)}(\tau_t)$. One of the advantages of our heteroscedastic approach is that it allows us to explore the evolution of brain activity over time. Movies of the posterior mean correlation can be found in the supplementary material. In the movies, and in Figure 5, there is a striking spatial smoothness to the correlation. This smoothness emerges even though our model imposes no spatial smoothness constraint, nor has any information about the spatial arrangement of the sensors. The left and right columns of Figure 5(a) and Figure 5(b) show the correlations for the word *hammer* and *house*, respectively. The top and bottom rows depict correlation for sensors covering the occipital lobe and parietal lobes of the brain, respectively. At 150 ms it is expected that there will be few differences between words, as the mind has not processed the meaning of the word. By 500 ms it is generally accepted [15] that the brain has begun to process the meaning of the word, and so we would expect the correlation patterns to differ between words. Both of these patterns of activation can be confirmed in Figure 5.

7.1 Evaluation Metrics

After models are learned for each of the 20 words, we can classify a single trial as belonging to a particular word category by choosing the category of the word model with the maximal score. We explored different



(a) Correlation at time = 151 ms



(b) Correlation at time = 501 ms

Figure 5: Posterior mean correlations computed from samples of $\Sigma^{(w)}(\tau)$ for Subject 1 and words *hammer* and *house* between a selected sensor (star) and all other sensors, shown as positioned in the MEG helmet viewed from above. The black triangle indicates the position of the subject's nose. The left and right columns of (a) and (b) show the correlations for the word *hammer* and *house*, respectively. The top and bottom rows are for sensors located near the frontal and parietal lobes, respectively. At 151 ms the mind has not processed the word's meaning, and there are few differences between brain activation patterns for different words. By 500 ms the brain is processing the meaning of the word, and the correlation patterns differ.

Table 2: Accuracy for classifying single MEG trials into one of four word categories. Our method appears in the shaded row. p_h indicates the hierarchical Monte Carlo integration of Equation (16), ll is log likelihood, $\hat{\Sigma}^{(w)}$ is the static covariance matrix, $\hat{\Sigma}_t^{(w)}$ is the kernel estimate of time-varying covariance and $\mu_{MLE}^{(w)}$ is the MLE estimate of the mean for word w. SVM is a linear support vector machine. (Binomial confidence intervals: * p=0.05, ** p=0.01)

Classification Rule	μ	Σ	**Accuracy, Subj 1**	**Accuracy, Subj 2**
Chance	-	-	0.25	0.25
$\underset{w}{\text{argmax}}\{ll(y_{1:n}^{(*)}\|\mu_{MLE}^{(w)}, \hat{\Sigma}^{(w)})\}$	$\mu_{MLE}^{(w)}$	$\hat{\Sigma}^{(w)}$	0.27	0.20
$\underset{w}{\text{argmax}}\{ll(y_{1:n}^{(*)}\|\mu_{MLE}^{(w)}, \hat{\Sigma}_t^{(w)})\}$	$\mu_{MLE}^{(w)}$	$\hat{\Sigma}_t^{(w)}$	0.33*	0.27
SVM			0.35*	0.32
$\underset{w}{\text{argmax}}\{p_h(y_{1:n}^{(*)}\|\{y_{train}^{(w)}\})\}$	$\mu^* \sim N(\mu^{(0,w)}, K)$	$\Sigma^{(w)}(\tau)$	0.39**	0.34*

scoring methods depending on the comparison being made.

Predictive Likelihood For our Bayesian hierarchical model of Section 4, we compute the predictive likelihood of each single trial as follows. Let $y_{1:n}^{(*)}$ denote the data associated with the single trial we wish to classify. For each word w, we wish to compute

$$p(y_{1:n}^{(*)} \mid \{y_{1:n}^{(i,w)}\}_{i=1}^{J_w}) = \int \left(\int p(y_{1:n}^{(*)} \mid \Omega^{(w)}, \psi^*) p(\psi^* \mid \psi^{(0,w)}) d\psi^* \right) p(\Omega^{(w)}, \psi^{(0,w)} \mid \{y_{1:n}^{(i,w)}\}_{i=1}^{J_w}) d\Omega, \quad (15)$$

where $\Omega^{(w)}$ denotes $\{\Theta^{(w)}, \xi^{(w)}, \Sigma_0^{(w)}\}$. Though the integral over ψ^* has a closed form, its computation involves an unwieldy matrix of size $(np)^2$. For this reason, we perform Monte Carlo integration by drawing Gaussian process samples $\psi_m^{*(i,w)}(\tau_{1:n}) \sim N_n(\psi_m^{(0,w)}, K)$ where $\psi_m^{(0,w)}$ is a MCMC sample of $\psi^{(0,w)}$. We then compute the likelihood of $y_{1:n}^{(*)}$ given the sampled $\psi_m^{*(i,w)}$. The integral over $\Omega^{(w)}$ is approximated by samples from our Gibbs sampler (see Section 6.) Thus, the predictive likelihood of the single trial data under word w is approximated as

$$p(y_{1:n}^{(*)} \mid \{y_{1:n}^{(i,w)}\}_{i=1}^{J_w}) \approx \frac{1}{300 \cdot M} \sum_{m=1}^{M} \left(\sum_{n=1}^{300} p(y_{1:n}^{(*)} | \Omega_m^{(w)}, \psi^{*(i,m)}) \right), \quad (16)$$

where M is the number of Gibbs samples considered, $\Omega_m^{(w)}$ is the mth Gibbs sample of $\{\Theta^{(w)}, \xi^{(w)}, \Sigma_0^{(w)}\}$, and 300 is the number of new Gaussian process samples we draw.

Log Likelihood Under MLE As a comparison to our Bayesian model-based approach, we classified single MEG trials based on non-latent-factor Gaussian formulations with maximum likelihood estimates (MLE) of the associated parameters. For each of the MLE-based models, we calculate the mean (μ_{MLE}) by averaging the training trials per word. An example of μ_{MLE} can be seen in Figure 4, alongside the estimation from our model and the noisy MEG signal. We estimate sample covariance matrices in two ways: (i) using all trials and time points to compute one static covariance matrix ($\hat{\Sigma}^{(w)}$), and (ii) with a kernel estimation method. The kernel estimation of the time-varying covariance ($\hat{\Sigma}_t^{(w)}$) was obtained by computing the covariance within a sliding window of size 40 ms. To both matrices a small diagonal component ($10^{-3}I_p$) was added to ensure positive definiteness. For each of these MLE-based comparisons, we compute the likelihood of the test single trial assuming a Gaussian with the specified MLE mean and covariance.

7.2 Word Category Classification Performance

To ascertain which characteristics of the signal are most important for classification performance, we evaluated the mean and covariance components in turn. The results are summarized in Table 2. There is a clear trend with both subjects: the MLE of the mean ($\mu_{MLE}^{(w)}$) with a static covariance ($\hat{\Sigma}^{(w)}$) was not powerful enough to represent the MEG activity, and does not yield statistically significant performance for either subject. Introducing a time-varying covariance matrix ($\hat{\Sigma}_t^{(w)}$) improves performance, giving statistically significant results for Subject 1, but not Subject 2. Note that it is exceedingly difficult to estimate covariance in high dimensions with little data, a hindrance overcome with our Bayesian latent factor approach.

Our method of fitting an instantaneous mean and covariance outperforms all models, and performs significantly above chance, $p = 0.01$ for Subject 1, and $p = 0.05$ for Subject 2. Our method is the only one to perform above chance for Subject 2.

Discriminative methods often produce very powerful and accurate classifiers, but they lack interpretability and extendibility. Still, for comparison, we include here the performance of a one-vs-all SVM (ties broken by distance to the hyperplane). We used a linear, RBF and third order polynomial kernel and found that the linear kernel performed best. The SVM performance is below that of our generative approach, and not statistically significant for Subject 2. Moreover, taking a generative approach rather than a discriminative one allows for extensions upon which we elaborate in Section 8.

8 Discussion

In this paper we introduced a method that identifies the signal amongst the noise in MEG recordings of brain activity. Our model outperforms discriminative methods and affords many opportunity for extensions. For example, a natural extension to our hierarchical model adds a layer to the hierarchy, building a parent model for each word *category*. In this way we can harness the signal in many more trials when approximating the average response to a particular word category. A similar technique could be used to fit a subject-specific response template, while hierarchically sharing information from many trials across many subjects.

Though not explored here, our method also allows for the prediction and interpolation of missing data, another challenge often encountered when working with sensor-derived data. Indeed, during MEG recording sessions, single sensors may become uncalibrated or artifacts introduced by eye blinks may temporarily obscure the signal. Our latent factor approach can cope with such lost data without relying on imputing the missing values.

Furthermore, our generative model supports sampling from the learned distribution, which allows us to explore changes in brain activity in relation to stimuli. In future work, we plan to explore brain activity fluctuations in response to the semantic components of a word, as in [20]. A generative model based on word semantics will allow us to model the mental processes used to represent concepts, which could lead to a better understanding of the human brain.

Acknowledgements

The authors would like to thank Gustavo Sudre and Leila Wehbe for their help with data acquisition, preprocessing, and useful suggestions. This research was partially supported by the W. M. Keck Foundation, NSF under Grant IIS-0835797, AFOSR under Grant FA9550-10-1-0501, and the National Institute of Environmental Health Sciences (NIEHS) of the NIH under Grant R01 ES017240. Alona Fyshe was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] S. Behseta, Robert E. Kass, and Garrick L. Wallstrom. Hierarchical models for assessing variability among functions. *Biometrika*, 92(2):419–434, June 2005. ISSN 0006-3444.

[2] A. Bhattacharya and D.B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

[3] S.P. Brooks and A. Gelman. General methods for general methods for monitoring convergence of iterative

simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[4] Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, December 2008. ISSN 0162-1459.

[5] Stavros I Dimitriadis, Nikolaos A Laskaris, Vasso Tsirka, Michael Vourkas, Sifis Micheloyannis, and Spiros Fotopoulos. Tracking brain dynamics via time-dependent network analysis. *Journal of Neuroscience Methods*, 193(1):145–155, 2010.

[6] Guido Dornhege, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE transactions on bio-medical engineering*, 51(6):993–1002, June 2004. ISSN 0018-9294.

[7] Emily Fox and David Dunson. Bayesian Nonparametric Covariance Regression, 2011. URL `http://arxiv.org/abs/1101.2017v2`.

[8] Roman Freunberger, Markus Werkle-Bergner, Birgit Griesmayr, Ulman Lindenberger, and Wolfgang Klimesch. Brain Oscillatory Correlates of Working Memory Constraints. *Brain research*, 1375:93–102, December 2010. ISSN 1872-6240.

[9] Alan E. Gelfand, Alexandra M. Schmidt, Sudipto Banerjee, and C. F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13(2):263–312, December 2004. ISSN 1133-0686. doi: 10.1007/BF02595775.

[10] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[11] Joyee Ghosh and David B Dunson. Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009.

[12] Adrian R Groves, Michael A Chappell, and Mark W Woolrich. Combined spatial and non-spatial prior for inference on MRI time-series. *NeuroImage*, 45(3):795–809, 2009.

[13] Marcos Perreau Guimaraes, Dik Kin Wong, E Timothy Uy, Logan Grosenick, and Patrick Suppes. Single-trial classification of MEG recordings. *IEEE transactions on bio-medical engineering*, 54(3):436–43, March 2007. ISSN 0018-9294.

[14] M. Hamalainen, R. Hari, R.J. Ilmoniemi, J. Knuutila, and O.V. Lounasmaa. Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2), 1993.

[15] Peter Hansen, Morten Kringelbach, and Riitta Salmelin. *MEG: An Introduction to Methods*. Oxford University Press, USA, 2010. ISBN 0195307232.

[16] Stefan J Kiebel, Jean Daunizeau, Christophe Phillips, and Karl J Friston. Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG. *NeuroImage*, 39(2):728–41, January 2008. ISSN 1053-8119.

[17] David Knowles and Zoubin Ghahramani. Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In *International Conference on Independent Component Analysis and Signal Separation*, pages 381–388, 2007.

[18] Steven Lemm, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE transactions on bio-medical engineering*, 52(9):1541–8, September 2005. ISSN 0018-9294.

[19] Hedibert Freitas Lopes, Esther Salazar, and Dani Gamerman. Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4):759–792, 2008. ISSN 19316690.

[20] Mark Palatucci, Geoffrey Hinton, Dean Pomerleau, and Tom M Mitchell. Zero-Shot Learning with Semantic Output Codes. *Advances in Neural Information Processing Systems*, 22:1410–1418, 2009.

[21] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*, volume 14. The MIT Press, April 2006.

[22] S Taulu and J Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology*, 51:1–10, 2006.

[23] Samu Taulu and Riitta Hari. Removal of magnetoencephalographic artifacts with temporal signal-space separation: demonstration with single-trial auditory-evoked responses. *Human brain mapping*, 30(5):1524–34, May 2009. ISSN 1097-0193.

[24] Samu Taulu, Matti Kajola, and Juha Simola. The Signal Space Separation method. *ArXiv Physics*, 2004. URL `http://arxiv.org/abs/physics/0401166`.

[25] Y.W. Teh, Matthias Seeger, and M.I. Jordan. Semiparametric latent factor models. In R. G. Cowell and Z. Ghahramani, editors, *Workshop on Artificial Intelligence and Statistics*, volume 10, pages 333–340. Citeseer, 2005.

[26] Nelson J Trujillo-Barreto, Eduardo Aubert-Vázquez, and Pedro a Valdés-Sosa. Bayesian model averaging in EEG/MEG imaging. *NeuroImage*, 21(4):1300–19, April 2004. ISSN 1053-8119.

[27] M A Uusitalo and R J Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Medical & biological engineering & computing*, 35(2):135–40, March 1997. ISSN 0140-0118.

[28] Mike West. Bayesian Factor Regression Models in the Large p , Small n Paradigm. *Bayesian Statistics*, 7 (2003):723–732, 2003.

[29] Andrew Wilson and Zoubin Ghahramani. Generalised wishart processes. In *The 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, Barcelona, Spain, July 2011.

[30] David P Wipf, Julia P Owen, Hagai T Attias, Kensuke Sekihara, and Srikantan S Nagarajan. Robust Bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using MEG. *NeuroImage*, 49(1):641–55, January 2010. ISSN 1095-9572.

Table 3: The variables of the model described in this paper, and their dimensions. In the context of the MEG data, p is the number of sensors, n is the number of time samples. L and k are parameters of the model that determine the number of latent factors.

Variable name	Dimension
$y_t^{(i,w)}$	$p \times 1$
$\mu^{(i,w)}(\tau_t)$	$p \times 1$
$\Sigma^{(w)}(\tau_t)$	$p \times p$
$\Sigma_0^{(w)}$	$p \times p$
$\eta_t^{(i,w)}$	$k \times 1$
$\psi^{(i,w)}(\tau_t)$	$k \times 1$
$\psi^{(0,w)}(\tau_t)$	$k \times 1$
$\nu_t^{(i,w)}$	$k \times 1$
$\epsilon_t^{(i,w)}$	$p \times 1$
$\Lambda^{(w)}(\tau_t)$	$p \times k$
$\xi_{(w)}(\tau_t)$	$L \times k$
$\Theta^{(w)}$	$p \times L$
$\omega_t^{(i,w)}$	$p \times 1$
$\tilde{\Sigma}_\xi$	$n \times n$
$\tilde{\Sigma}_\psi^{(w)}$	$n \times n$

A Model Details

Here we cover in greater detail the sampling methodology used to estimate the time-varying mean and covariance of the MEG signal $y_t^{(i,w)}$. Code is available at `http://www.cs.cmu.edu/~afyshe/papers/aistats2012/`. For reference, the variables appear in Table 3, along with their dimensions. Recall our latent factor model:

$$\begin{aligned}\eta_t^{(i,w)} &= \psi^{(i,w)}(\tau_t) + \nu_t^{(i,w)}, \quad \nu_t^{(i,w)} \sim N_k(0, I_k) \\ y_t^{(i,w)} &= \Lambda^{(w)}(\tau_t)\eta_t^{(i,w)} + \epsilon_t^{(i,w)}, \quad \epsilon_t^{(i,w)} \sim N_p(0, \Sigma_0^{(w)}).\end{aligned} \tag{17}$$

With this formulation, we can marginalize out the idiosyncratic noise terms $\nu_t^{(i,w)}$ and $\epsilon_t^{(i,w)}$ to induce the following mean and covariance structure:

$$\begin{aligned}\mu^{(i,w)}(\tau_t) &= \Theta^{(w)}\xi^{(w)}(\tau_t)\psi^{(i,w)}(\tau_t) \\ \Sigma^{(w)}(\tau_t) &= \Theta^{(w)}\xi^{(w)}(\tau_t)\xi^{(w)}(\tau_t)'\Theta^{(w)'} + \Sigma_0^{(w)}.\end{aligned} \tag{18}$$

A.1 Prior selection

To ameliorate the burden of setting L (the number of latent dictionary functions), we seek a prior on Θ that favors many values of Θ being close to zero. Then we may choose L larger than the expected number of dictionary functions (also controlled by the latent factor dimension k). As proposed in [2], we use the following shrinkage prior:

$$\begin{aligned}&\Theta_{j\ell} \mid \phi_{j\ell}, \zeta_\ell \sim \mathcal{N}(0, \phi_{j\ell}^{-1}\zeta_\ell^{-1}) \quad \phi_{j\ell} \sim \mathrm{Ga}(3/2, 3/2) \\ &\delta_1 \sim \mathrm{Ga}(a_1, 1), \quad \delta_h \sim \mathrm{Ga}(a_2, 1),\ h \geq 2, \quad \zeta_\ell = \prod_{h=1}^{\ell} \delta_h.\end{aligned} \tag{19}$$

Choosing $a_2 > 1$ implies that δ_h is greater than 1 in expectation so that ζ_ℓ tends stochastically towards infinity as ℓ goes to infinity, thus shrinking the elements $\Theta_{j\ell}$ toward zero increasingly as ℓ grows. The $\phi_{j\ell}$ precision parameters allow for flexibility in how the elements of Θ are shrunk towards zero by incorporating local shrinkage specific to each element of Θ, while ζ_ℓ provides a global column-wise shrinkage factor.

We specify the prior on Σ_0 via the usual inverse gamma priors on the diagonal elements of $\Sigma_0^{(W)}$. That is,

$$\sigma_{w,j}^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma) \tag{20}$$

independently for each $j = 1, \ldots, p$.

A.2 Sampling

Now we cover the sampling procedure in its entirety. Recall that the sampler is run independently for each set of J_w single trials for word w, and each single trial has length n. The derivations of the conditional posteriors harness the fact the observation model of Equation (17) can be rewritten as

$$y_{t,j}^{(i,w)} = \sum_{m=1}^{k} \eta_{t,m}^{(i,w)} \sum_{\ell=1}^{L} \Theta_{j\ell}^{(w)} \xi_{\ell m}^{(w)}(\tau_t) + \epsilon_{t,j}^{(i,w)}. \tag{21}$$

Again, we use K_j to denote the Gram matrix with elements $c_j(t, t')$ for $j = 0, 1$.

Step 1: Block-Sample $\{\psi^{(i,w)}, \nu_{1:n}^{(i,w)}\}$ For each single trial i, we sample $\psi^{(i,w)}$ from its conditional posterior marginalizing $\nu_t^{(i,w)}$ and cycling through each latent dictionary function $\psi_\ell^{(i,w)}$. We then treat $\nu_{1:n}^{(i,w)}$ as auxiliary variables that are imputed conditioned on $\psi^{(i,w)}$. From Equation (17), we have that marginally $y_t^{(i,w)} = \Lambda^{(w)}(\tau_t)\psi^{(i,w)}(\tau_t) + \omega_t^{(i,w)}$ with $\omega_t^{(i,w)} \sim N_p(0, \Sigma^{(w)}(\tau_t))$, $\Lambda^{(w)}(\tau_t) = \Theta^{(w)}\xi^{(w)}(\tau_t)$, and $\Sigma^{(w)}(\tau_t)$ as in Equation (18). Standard Gaussian results imply that the conditional posterior of $\psi_\ell^{(i,w)}(\tau_{1:n})$ (i.e., the dictionary function evaluated at $(\tau_1, \ldots, \tau_n)$ is

$$\begin{aligned}&\psi_\ell^{(i,w)}(\tau_{1:n}) \mid y_t^{(i,w)}, \psi_{\setminus\ell}^{(i,w)}(\tau_{1:n}), \psi^{(0,w)}(\tau_{1:n}), \Omega^{(w)} \sim \\ &N_n\left(\tilde{\Sigma}_\psi^{(w)} \begin{bmatrix} [\Lambda^{(w)}(\tau_1)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_1)\tilde{y}_1^{(i,w)} + \psi^{(0,w)}(\tau_1) \\ \vdots \\ [\Lambda^{(w)}(\tau_n)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_n)\tilde{y}_n^{(i,w)} + \psi^{(0,w)}(\tau_n) \end{bmatrix}, \tilde{\Sigma}_\psi^{(w)}\right),\end{aligned} \tag{22}$$

where we use $\Omega^{(w)}$ to denote $\{\Theta^{(w)}, \xi^{(w)}, \Sigma_0^{(w)}\}$ and

$$\begin{aligned}\tilde{\Sigma}_\psi^{-(w)} = K_1^{-1} + \\ \mathrm{diag}\Big([\Lambda^{(w)}(\tau_1)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_1)[\Lambda^{(w)}(\tau_1)]_{\cdot\ell}, \ldots \\ [\Lambda^{(w)}(\tau_n)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_n)[\Lambda^{(w)}(\tau_n)]_{\cdot\ell}\Big)\end{aligned}$$

$\tilde{y}_t^{(i,w)} = y_t^{(i,w)} - \sum_{(r\neq\ell)}[\Lambda^{(w)}(\tau_t)]_{\cdot r}\psi_r^{(i,w)}(\tau_t)$, and $\psi_{\setminus\ell}^{(i,w)}$ is the set of latent dictionary functions $\psi_j^{(i,w)}$ for $j \neq \ell$.

Conditioned on $\psi^{(i,w)}$, we independently impute $\nu_t^{(i,w)}$ for each t. Consider

$$\begin{aligned}\tilde{y}_t^{(i,w)} &= y_t^{(i,w)} - \Lambda^{(w)}(\tau_t)\psi^{(i,w)}(\tau_t) \\ &= \Lambda^{(w)}(\tau_t)\nu_t^{(i,w)} + \epsilon_t^{(i,w)} \end{aligned} \quad (23)$$

Then, straightforwardly,

$$\begin{aligned}\nu_t^{(i,w)} \mid y_t^{(i,w)}, \psi^{(i,w)}(\tau_t), \Theta^{(w)}, \xi^{(w)}(\tau_t), \Sigma_0^{(w)} \sim \\ N_k\left(\Phi^{(w)}\Lambda^{(w)'}(\tau_t)\Sigma_0^{-(w)}\tilde{y}_t^{(i,w)}, \Phi^{(w)}\right). \end{aligned} \quad (24)$$

Where $\Phi = \left(I + \Lambda^{(w)'}(\tau_t)\Sigma_0^{-(w)}\Lambda^{(w)}(\tau_t)\right)^{-1}$

Step 2: Sample $\psi^{(0,w)}$ Conditioned on $\{\psi_\ell^{(i,w)}\}_{i=1}^{J_w}$ the parent latent process $\psi_\ell^{(0,w)}$ has form $\psi_\ell^{(0,w)}(\tau_{1:n}) \sim N_n(\Omega^{-(w)}\phi_\ell^w, \Omega^{-(w)})$ where $\Omega^w = K_0^{-1} + J_w K_1^{-1}$ and $\phi_\ell^w = K_1^{-1}\sum_i \psi_\ell^{(i,w)}(\tau_{1:n})$.

Step 3: Sample $\xi^{(w)}$ Conditioning on $\xi_{\setminus\ell m}^{(w)}$ (i.e., all latent covariance dictionary elements not equal to $\xi_{\ell m}^{(w)}$), our Gaussian process prior implies the following conditional posterior:

$$\begin{aligned}&\xi_{\ell m}^{(w)}(\tau_{1:n})|\{y_t^{(i,w)}\}, \psi^{(i,w)}, \{\nu_t^{(i,w)}\}, \Theta^{(w)}, \xi_{\setminus\ell m}^{(w)}, \Sigma_0^{(w)} \\ &\sim N_n\left(\tilde{\Sigma}_\xi \sum_{i=1}^{J_w} \begin{bmatrix} \eta_{1,m}^{(i,w)} \sum_{j=1}^p \Theta_{j\ell}^{(w)} \sigma_{j,w}^{-2} \tilde{y}_{1,j}^{(i,w)} \\ \vdots \\ \eta_{n,m}^{(i,w)} \sum_{j=1}^p \Theta_{j\ell}^{(w)} \sigma_{j,w}^{-2} \tilde{y}_{n,j}^{(i,w)} \end{bmatrix}, \tilde{\Sigma}_\xi\right) \end{aligned} \quad (25)$$

where

$$\begin{aligned}\tilde{\Sigma}_\xi^{-1} = K_1^{-1} + \\ \sum_{i=1}^{J_w} \mathrm{diag}\left(\left(\eta_{1,m}^{(i,w)}\right)^2 \sum_{j=1}^p \left(\Theta_{j\ell}^{(w)}\right)^2 \sigma_{j,w}^{-2}, \dots \right. \\ \left. \left(\eta_{n,m}^{(i,w)}\right)^2 \sum_{j=1}^p \left(\Theta_{j\ell}^{(w)}\right)^2 \sigma_{j,w}^{-2}\right) \end{aligned} \quad (26)$$

and $\tilde{y}_{t,j}^{(i,w)} = y_{t,j}^{(i,w)} - \sum_{(r,s)\neq(\ell,m)} \Theta_{jr}^{(w)} \xi_{rs}^{(w)}(\tau_t)$.

Step 4: Sample $\Sigma_0^{(w)}$ Let $\Theta_{j\cdot}^{(w)} = \begin{bmatrix}\Theta_{j1}^{(w)} & \dots & \Theta_{jL}^{(w)}\end{bmatrix}$ and $\eta_t^{(i,w)}$ be as in Equation (17). Since the diagonal elements of $\Sigma_0^{(w)}$ have prior $\sigma_{w,j}^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma)$, standard conjugate posterior analysis yields

$$\begin{aligned}&\sigma_{w,j}^{-2} \mid \{y_t^{(i,w)}\}, \psi^{(i,w)}, \nu_t^{(i,w)}, \Theta^{(w)}, \xi^{(w)} \sim \\ &\mathrm{Ga}\left(a_\sigma + \frac{nJ_w}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{J_w}\sum_{t=1}^{n}(y_{t,j}^{(i,w)} - \Theta_{j\cdot}^{(w)}\xi^{(w)}(\tau_t)\eta_t^{(i,w)})^2\right) \end{aligned}$$

Step 5: Sample $\Theta^{(w)}$ Conditioned on a set of latent precision parameters $\phi^{(w)}$ and $\zeta^{(w)}$, the shrinkage prior of [2] reduces to a Gaussian prior on Θ. The posterior decomposes along the rows of Θ as follows:

$$\begin{aligned}\Theta_{j\cdot}^{(w)} \mid \{y_{t,j}^{(i,w)}\}, \psi^{(i,w)}, \nu_{1:n}^{(i,w)}, \xi^{(w)}, \phi^{(w)}, \zeta^{(w)} \sim \\ N_L\left(\sigma_{w,j}^{-2}\tilde{\Sigma}_\Theta^{(w)}\tilde{\eta}^{(w)'}y_{\cdot,j}, \tilde{\Sigma}_\Theta^{(w)}\right), \end{aligned} \quad (27)$$

where $\tilde{\eta}^{(w)'}$ is the concatenation of matrices $\xi^{(w)}(\tau_{1:n})\eta_{1:n}^{(i,w)}$ for $i = 1 \dots J_w$.

$$\begin{aligned}\tilde{\eta}^{(w)'} = \left[\xi^{(w)}(\tau_1)\eta_1^{(1,w)} \dots \xi^{(w)}(\tau_n)\eta_n^{(1,w)} \dots \right. \\ \left. \xi^{(w)}(\tau_1)\eta_1^{(J_w,w)} \dots \xi^{(w)}(\tau_n)\eta_n^{(J_w,w)}\right] \end{aligned} \quad (28)$$

$y_{\cdot,j}$ concatenates the jth sensor measurements for *all* T single trials into a column vector, and $\tilde{\Sigma}_\Theta^{-(w)} = \sigma_{w,j}^{-2}\tilde{\eta}^{(w)'}\tilde{\eta}^{(w)} + \mathrm{diag}(\phi_{j1}^{(w)}\zeta_1^{(w)}, \dots, \phi_{jL}^{(w)}\zeta_L^{(w)})$.

The hyperparameters $\phi^{(w)}$ and $\zeta^{(w)}$ are updated as in [2].

A.3 Parameter Settings

Recall the squared exponential kernel $c_i(t,t') = d_i \exp(-\kappa||t-t'||_2^2)$. We parameterized with $\kappa = 500$, $d_0 = 8$ and $d_1 = 4$.

To ensure that L=10 was set high enough, we verified that a large number of the columns of Θ with higher column index were sufficiently close to 0 in L^2 norm. For our choice of the k=10, we performed sensitivity analysis and got comparable results.

Following [7] we set $a_1 = 2$, $a_2 = 5$. For the idiosyncratic noise variances, we place a diffuse prior with $a_\sigma = 1$ and $b_\sigma = 0.1$.

B MEG Data Acquisition

All subjects gave their written informed consent approved by the University of Pittsburgh (protocol PRO09030355) and Carnegie Mellon (protocol HS09-343) Institutional Review Boards. MEG data were recorded using an Elekta Neuromag device (Elekta Oy). While the machine has 306 sensors, to reduce the dimension of the data, only recordings from the second gradiometers were used for these experiments (arbitrarily chosen). The data was acquired at 1 kHz, high-pass filtered at 0.1 Hz and low-pass filtered at 330 Hz. Eye movements (horizontal and vertical eye movements as well as blinks) were monitored by recording the differential activity of muscles above, below, and beside the eyes. At the beginning of each session we recorded the position of the subject's head with four head position indicator (HPI) coils placed on the subjects scalp. The HPI coils, along with three cardinal points (nasion, left and right pre-auricular), were digitized into the system.

The data were preprocessed using the Signal Space Separation method (SSS) [22, 24] and temporal extension of SSS (tSSS) [23] to remove artifacts and noise unrelated to brain activity. In addition, we used tSSS to realign the head position measured at the beginning of each block to a common location. The MEG signal was then low-pass filtered to 50 Hz to remove the contributions of line noise and downsampled to 200 Hz. The Signal Space Projection method

Algorithm 1 One iteration of the sampling algorithm. For compactness, some distributions are given in information form. Recall that w indexes the word model, J_w the number of trials for a word model, p the number of sensors, n the number of time points, and L and k are the dimensions of the latent dictionaries.

for $i \leftarrow 1, J_w$ **do**

  for $\ell \leftarrow 1, L$ **do**

$$\tilde{\Sigma}_{\psi}^{(w)} = \tilde{K}^{-1} + \operatorname{diag}\left([\Lambda^{(w)}(\tau_1)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_1)[\Lambda^{(w)}(\tau_1)]_{\cdot\ell}, \ldots [\Lambda^{(w)}(\tau_n)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_n)[\Lambda^{(w)}(\tau_n)]_{\cdot\ell}\right)$$

$$\tilde{y}_t^{(i,w)} = y_t^{(i,w)} - \textstyle\sum_{(r\neq\ell)}[\Lambda^{(w)}(\tau)]_{\cdot r}\psi_r^{(i,w)}(\tau)$$

$$\psi_\ell^{(i,w)}(\tau_{1:n}) \sim N_n^{-1}\left(\begin{bmatrix} [\Lambda^{(w)}(\tau_1)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_1)\tilde{y}_1^{(i,w)} + \psi_1^{(0,w)} \\ \vdots \\ [\Lambda^{(w)}(\tau_n)]'_{\cdot\ell}\Sigma^{-(w)}(\tau_n)\tilde{y}_n^{(i,w)} + \psi_n^{(0,w)} \end{bmatrix}, \tilde{\Sigma}_{\psi}^{(w)}\right)$$

  for $t \leftarrow 1, n$ **do**

$$\tilde{y}_t^{(i,w)} = y_t^{(i,w)} - \Lambda^{(w)}(\tau)\psi^{(i,w)}(\tau)$$

$$\nu_t^{(i,w)} \sim N_k^{-1}\left(\Lambda^{(w)'}(\tau)\Sigma_0^{-(w)}\tilde{y}_t^{(i,w)}, I + \Lambda^{(w)'}(\tau)\Sigma_0^{-(w)}\Lambda^{(w)}(\tau)\right)$$

for $\ell \leftarrow 1, L$ **do**

$$\psi_\ell^{(0,w)}(\tau_{1:n}) \sim N_n^{-1}(K_1^{-1}\textstyle\sum_{i=1}^{J_w}\psi_\ell^{(i,w)}(\tau_{1:n}), K_0^{-1} + J_w K_1^{-1})$$

for $\ell \leftarrow 1, L$ **do**

  for $m \leftarrow 1, K$ **do**

$$\tilde{\Sigma}_\xi = K_1^{-1} + \textstyle\sum_{i=1}^{J_w}\operatorname{diag}\left(\left(\eta_{1,m}^{(i,w)}\right)^2\sum_{j=1}^p\left(\Theta_{j\ell}^{(w)}\right)^2\sigma_{j,w}^{-2}, \ldots \left(\eta_{n,m}^{(i,w)}\right)^2\sum_{j=1}^p\left(\Theta_{j\ell}^{(w)}\right)^2\sigma_{j,w}^{-2}\right)$$

$$\tilde{y}_{t,j}^{(i,w)} = y_{t,j}^{(i,w)} - \textstyle\sum_{(r,s)\neq(\ell,m)}\Theta_{jr}^{(w)}\xi_{rs}^{(w)}(\tau).$$

$$\xi_{\ell m}^{(w)}(\tau_{1:n}) \sim N_n^{-1}\left(\sum_{i=1}^{J_w}\begin{bmatrix} \eta_{1,m}^{(i,w)}\sum_{j=1}^p\Theta_{j\ell}^{(w)}\sigma_{j,w}^{-2}\tilde{y}_{1,j}^{(i,w)} \\ \vdots \\ \eta_{n,m}^{(i,w)}\sum_{j=1}^p\Theta_{j\ell}^{(w)}\sigma_{j,w}^{-2}\tilde{y}_{n,j}^{(i,w)} \end{bmatrix}, \tilde{\Sigma}_\xi\right)$$

for $j \leftarrow 1, p$ **do**

$$\Theta_{j\cdot}^{(w)} = \begin{bmatrix}\Theta_{j1}^{(w)} & \ldots & \Theta_{jL}^{(w)}\end{bmatrix},\ \eta_t^{(i,w)} = \psi^{(i,w)}(\tau) + \nu_t^{(i,w)}$$

$$\sigma_{w,j}^{-2} \sim \operatorname{Ga}\left(a_\sigma + \tfrac{nJ_w}{2}, b_\sigma + \tfrac{1}{2}\textstyle\sum_{i=1}^{J_w}\sum_{t=1}^n(y_{t,j}^{(i,w)} - \Theta_{j\cdot}^{(w)}\xi^{(w)}(\tau)\eta_t^{(i,w)})^2\right)$$

for $j \leftarrow 1, p$ **do**

$$y_{\cdot,j} = [y_{(1,j)}(t_{1:n}) \ldots y_{(J_w,j)}(t_{1:n})]'$$

$$\tilde{\eta}^{(w)'} = \begin{bmatrix}\xi^{(w)}(\tau_1)\eta_1^{(1,w)} & \ldots & \xi^{(w)}(\tau_n)\eta_n^{(1,w)} & \ldots & \xi^{(w)}(\tau_1)\eta_1^{(J_w,w)} & \ldots & \xi^{(w)}(\tau_n)\eta_n^{(J_w,w)}\end{bmatrix}$$

$$\tilde{\Sigma}_\Theta^{-(w)} = \sigma_{w,j}^{-2}\tilde{\eta}^{(w)'}\tilde{\eta}^{(w)} + \operatorname{diag}(\phi_{j1}^{(w)}\zeta_1^{(w)}, \ldots, \phi_{jL}^{(w)}\zeta_L^{(w)})$$

$$\Theta_{j\cdot}^{(w)} \sim N_L\left(\sigma_{w,j}^{-2}\tilde{\Sigma}_\Theta^{(w)}\tilde{\eta}^{(w)'}y_{\cdot,j}, \tilde{\Sigma}_\Theta^{(w)}\right)$$

(SSP) [27] was then used to remove signal contamination by eye blinks or movements, as well as MEG sensor malfunctions or other artifacts. Each MEG repetition starts 260 ms before stimulus onset, and ends 1440 ms after stimulus onset, for a total of 1.7 seconds and 340 time points of data per sample. MEG recordings are known to drift with time, so we corrected our data by subtracting the mean signal amplitude during the 200ms before stimulus onset, for each sensor/repetition pair. Because the magnitude of the MEG signal is very small, we multiplied the signal by 10^{12} to avoid numerical precision problems.

Due to recording error, 2 trials were lost. Words *lettuce* and *cat* have only 19 trials.

C MEG Videos

Four videos of the posterior mean correlation (computed from samples of $\Sigma^{(w)}(\tau)$) for Subject 1 are included in the supplementary material (available at `http://www.cs.cmu.edu/~afyshe/papers/aistats2012/`). The videos are named "S1_hierarchical_helmet_word1_word2_sensorNum.avi" where word1 and word2 correspond to the word models used to create the video and sensorNum is the number of the sensor for which the correlation is plotted. Note that the videos for *hammer* and *house* correspond to the snapshots shown in Figure 5 of the paper.