Stochastic Variational Inference for the EEG Inverse Problem

Fatima Afzali

IDH3035: Biophysics of Neural Computation, Florida International University

Introduction

In the EEG/MEG neuroimaging paradigms, electromagnetic fields generated by populations of pyramidal neurons are measured on the surface of the scalp. The locations of these populations, known as the **sources**, are related to the measured signals, by a linear mapping known as the **lead field**:

$$B_t = LS_t + \mathcal{E}$$

where B_t is a vector of b signals measured at time t, S_t is a vector of s dipoles at a set of points in the brain generating these signals, L is the $b \times 3s$ matrix relating S_t to B_t, and $\mathcal{E}$ is an error term. There are two main problems in estimating the parameters of this model:

Forward Problem

Given S_t, how do we determine B_t? In most approaches, a *head model* is constructed to describe how electromagnetic signals conduct; given geometric information describing the shape of a subject's head, this model is constructed using methods such as the Boundary Element Method (BEM) or the Finite Element Method (FEM); the lead field L is then constructed from this model, and applied to a given set of sources.

Inverse Problem

Given $B_{(t)}$ and L, how do we determine S_t? We wish to maximize $p(B_t \mid S_t)$ w/r/t S_t for an observed B_t, *while also* placing "reasonableness" constraints on S_t, which amounts to minimizing $S_t^T \Sigma_s^{-1} S_t$ for some weighting matrix Σ_s. In *linear* formulations, Σ_s is fixed, allowing us to find an analytic solution. In *Bayesian* formulations, Σ_s itself is probabilistically generated, and, while we may be given a prior, we must fit Σ_s to the data as well. This is the approach we take.

Statistical Model

Suppose that in a short timespan we measure n scalp observations, $\{B_n\}$. Compile these into a matrix $B \in \mathbb{R}^{b \times n}$. We then wish to find, for each column, the generating sources $\{S_n\}$, which will be compiled into the matrix $S \in \mathbb{R}^{s \times n}$.

$$p(B \mid S) \propto \exp\left(-\frac{1}{2}\operatorname{Tr}\left[(B - LS)^T \Sigma^{-1} (B - LS)\right]\right) \qquad (1)$$

The sources S will be regularized by placing a distribution on it as well: given a family of possible covariance matrices Σ_s parameterized by $\gamma = \{\gamma_i\}$,

$$p(S \mid \gamma) \propto \exp\left(-\frac{1}{2}\operatorname{Tr}\left[S^T \Sigma_s^{-1} S\right]\right)$$

The covariance matrix Σ_s will be assumed to be diagonal, giving us $3s$ covariance basis matrices $C_i = e_i e_i^T, i = 1, \ldots, 3s$. Each basis will have magnitude γ_i; we then place an inverse-Gamma prior on γ_i:

$$p(\gamma_i \mid a_i, b_i) = \text{Inv-Gamma}(a_i, b_i) = \Gamma(a_i)^{-1} b_i^{a_i} \gamma_i^{-(a_i+1)} e^{-b_i/\gamma_i}$$

where $a = [a_1, \ldots, a_{3s}]^T$ and $b = [b_1, \ldots, b_{3s}]^T$ are the model parameters to be fit to B.

We set the initial parameters $\Sigma = I_b$; $a, b = \mathbf{1}_{3s}$; $\Sigma_s = \sum_{i=1}^{3s} \gamma_i C_i$.

Classical Variational Inference

As detailed below, the posterior distribution $p(S, \gamma \mid B)$ will be approximated by the probability distribution $q(S, \gamma)$, which one assumes factorizes:

$$q(S, \gamma) = q_S(S) \cdot q_\gamma(\gamma)$$

These distributions are adjusted to minimize the Kullback-Leibler Divergence, a measure of statistical similarity, between the approximate and true posteriors. In this paper, we demonstrated that $q_S(S)$ was a product of multivariate Gaussian distributions, $q_\gamma(\gamma)$ a product of inverse Gamma distributions, and derived a *variational EM* algorithm to iteratively update $q_S(S)$ and $q_\gamma(\gamma)$, as well as to estimate the covariance Σ:

$$q_S^{(new)}(S) \leftarrow \prod_{i=1}^{n} \mathcal{N}\left(\Sigma_s L^T (L\Sigma_s L^T + \Sigma)^{-1} B_i, (\Sigma_s^{-1} + L^T \Sigma^{-1} L)^{-1}\right)$$

$$\Sigma \leftarrow n^{-1} LSS^T L^T$$

$$q_\gamma^{(new)}(\gamma) \leftarrow \prod_{i=1}^{3s} \text{Inv-Gamma}\left(a_i + \frac{n}{2}, b_i + \frac{n}{2}\mathbb{E}_{q_S^{(new)}(S)}\left[S_i^T S_i\right]\right)$$

where

$$\begin{aligned}\mathbb{E}\left[S_i^T S_i\right] &= \operatorname{Tr}\left[(\Sigma_s^{-1} + L^T \Sigma^{-1} L)^{-1}\right] \\ &\quad + B_i^T (\Sigma + L\Sigma_s L^T)^{-1} L \Sigma_s^2 L^T (\Sigma + L\Sigma_s L^T)^{-1} B_i\end{aligned}$$

We repeat these updates until convergence.

Background: Variational Inference

Suppose that an observed random variable x depends on a set of parameters $\theta = (\theta_1, \ldots, \theta_k)$, giving us a probability distribution $p(x \mid \theta)$. The question Bayesian inference asks is: *how do we determine θ given x?* Upon observing x, Bayes' Law gives us the probability of θ given x, known as the posterior distribution:

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{\int p(x \mid \theta) p(\theta)\, d\theta}$$

where $p(\theta)$ represents what we believed about θ before observing x (our prior), and the integral in the denominator, which is equal to $p(x)$, is known as the model evidence. While this is not a function of θ, we may set hyperparameters λ and treat θ as random variables distributed as $p(\theta \mid \lambda)$, in which case the model evidence does become a function of λ. (In our model, $x = B, \theta = S, \lambda = \gamma$.) While the model evidence is in general computationally intractable, we may approximate it:

$$\ln p(x) = \mathcal{L}(q(\theta, \lambda)) + \mathrm{KL}\left[q(\theta, \lambda) \mid p(\theta, \lambda \mid x)\right] \geq \mathcal{L}(q(\theta, \lambda))$$

Maximizing the approximation $\mathcal{L}(q(\theta, \lambda))$ w/r/t λ is equivalent to minimizing the Kullback-Leibler (KL) divergence. This is also computationally intractable in general, though, so we must make an additional assumption: namely, that q factors over its parameters. This is known as the *mean field approximation*:

$$q(\theta, \lambda) = q_\theta(\theta) q_\lambda(\lambda)$$

We then fit these distributions in an iterative manner.

Stochastic Variational Inference

In this paper, we demonstrated that the Fisher information metric took on a block diagonal form, and derived an analytic solution for the natural gradient. Denoting for convenience $\Omega_i = \mathbb{E}_{q_S(S)}\left[S_i^T S_i\right]$ (see left panel),

$$\widetilde{\nabla}_{a_i} \mathcal{L}(a_i) = \frac{b_i \Omega_i}{2(a_i \psi_1(a_i) - 1)(a_i - 1)} \left(\frac{\psi_1(a_i) + a_i^2 - a_i - 1}{a_i \psi_1(a_i) - \psi_1(a_i) - a_i^2 + a_i}\right)$$

$$\widetilde{\nabla}_{b_i} \mathcal{L}(b_i) = \frac{b_i^2 \Omega_i}{2(a_i \psi_1(a_i) - 1)(a_i - 1)} \left(\frac{1}{a_i - 1} - \psi_1(a_i)\right)$$

In order to use these natural gradients to infer sources, we require a timestep regime $\{\rho_i\}$ that tells us how to balance new information with old information; given this, our update equations for a, b for the $(t+1)^{th}$ update are:

$$a_i^{(t+1)} = a_i^{(t)} + \frac{\rho_t}{n} \widetilde{\nabla}_{a_i^{(t)}} \mathcal{L}(a_i^{(t)})$$

$$b_i^{(t+1)} = b_i^{(t)} + \frac{\rho_t}{n} \widetilde{\nabla}_{b_i^{(t)}} \mathcal{L}(b_i^{(t)})$$

If we want, we can compute MAP estimates for S, γ:

$$\gamma_i := \frac{b_i}{a_i - 1} \qquad S := \Sigma_s L^T (L\Sigma_s L^T + \Sigma)^{-1} B$$

Background: Information Geometry

Most forms of inference work by *gradient ascent* on the coordinates of the variational distribution q. To do this, you need to work in a space with a *metric* that allows the measurement of distances. Generally, this is flat Euclidean space $\mathbb{R}^k$ with the metric $d(x, y) = \sqrt{(x-y)^T (x-y)}$. We can do better!

Consider the $k-$dimensional manifold $\mathcal{M}$ whose points are members of some family of probability distributions. There is a unique Riemannian metric invariant under mappings that preserve sufficient statistics, known as the *Fisher Information Metric*. If $\mathcal{M}$ is parametrized by $\theta = [\theta_1, \ldots, \theta_k]^T$, and $p(x \mid \theta)$ is a function over this manifold, then this metric, at a point θ, is:

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[(\nabla_\theta \ln p(x \mid \theta))\,(\nabla_\theta \ln p(x \mid \theta))^T\right]$$

This matrix is symmetric and (in general) non-singular; placing the metric $\mathcal{I}(\theta)$ on $\mathbb{R}^k$ yields the *natural gradient* of p:

$$\widetilde{\nabla}_\theta p(x \mid \theta) = \mathcal{I}(\theta)^{-1} \nabla_\theta p(x \mid \theta)$$

As this takes into account not just the distance between parameters but the distance between the probability distributions *represented* by those parameters, it yields significantly faster convergence!

In the case of variational inference, we may use the natural gradient to revise the evidence lower bound (ELBO) $\mathcal{L}(q(\theta, \lambda))$ by treating the ELBO as a function of λ and performing coordinate ascent on $\mathcal{L}(\lambda)$ by computing the natural gradient $\widetilde{\nabla}_\lambda \mathcal{L}(\lambda)$.

Computational Results

Synthetic EEG data was created by simulating the EEG signals produced by a pair of sources slowly moving through the brain; a Python program was then written to very naively implement these algorithms. While inference is simultaneously performed over large batches of data, the magnitude of the source dipoles at 200 candidate locations is displayed below only for a specific time. Due to a lack of optimization for space and time complexity (many such optimizations can be made in these algorithms, given that most matrices here are block diagonal and/or have closed-form inverses), the program took up more space and time than necessary, possibly several orders of magnitude more. Despite this, it still gains accuracy quickly, as shown in the bottom figure, which was plotted after two iterations of the stochastic variational inference algorithm.

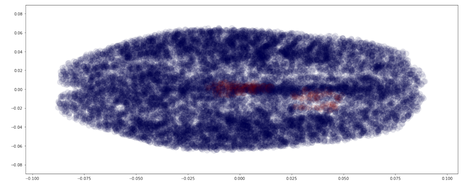

Figure: Artificial EEG sources (axial view)

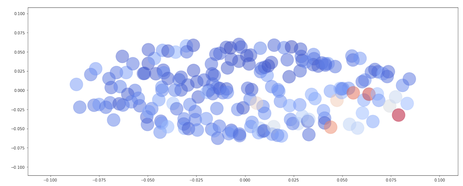

Figure: Computed sources after two iterations (axial view)

References

[1] Christopher M Bishop.
Pattern Recognition and Machine Learning.
Springer, 2006.

[2] David Wipf and Srikantan Nagarajan.
A unified bayesian framework for meg/eeg source imaging.
NeuroImage, 44(3):947–966, 2009.

[3] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols.
Statistical Parametric Mapping: the Analysis of Functional Brain Images.
Elsevier, 2011.

[4] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley.
Stochastic variational inference.
The Journal of Machine Learning Research, 14(1):1303–1347, 2013.

[5] Shun-ichi Amari.
Information geometry and its applications, volume 194.
Springer, 2016.