Multiresolution Gaussian Processes

Emily B. Fox
Dept of Statistics, University of Washington
ebfox@stat.washington.edu

David B. Dunson
Dept of Statistical Science, Duke University
dunson@stat.duke.edu

Abstract

We propose a multiresolution Gaussian process to capture long-range, non-Markovian dependencies while allowing for abrupt changes and non-stationarity. The multiresolution GP hierarchically couples a collection of smooth GPs, each defined over an element of a random nested partition. Long-range dependencies are captured by the top-level GP while the partition points define the abrupt changes. Due to the inherent conjugacy of the GPs, one can analytically marginalize the GPs and compute the marginal likelihood of the observations given the partition tree. This property allows for efficient inference of the partition itself, for which we employ graph-theoretic techniques. We apply the multiresolution GP to the analysis of magnetoencephalography (MEG) recordings of brain activity.

1 Introduction

A key challenge in many time series applications is capturing long-range dependencies for which Markov-based models are insufficient. One method of addressing this challenge is through employing a Gaussian process (GP) with an appropriate (non-band-limited) covariance function. However, GPs typically assume smoothness properties that can blur key elements of the signal if abrupt changes occur. The Matérn kernel enables less smooth functions, but assumes a stationary process that does not adapt to varying levels of smoothness. Likewise, a changepoint [21] or partition [8] model between smooth functions fails to capture long range dependencies spanning changepoints.

Another long-memory process is the fractional ARIMA process [5, 13]. Wavelet methods have also been proposed, including recently for smooth functions with discontinuities [2]. We take a fundamentally different approach based on GPs that allows (i) direct interpretability, (ii) local stationarity, (iii) irregular grids of observations, and (iv) sharing information across related time series.

As a motivating application, consider magnetoencephalography (MEG) recordings of brain activity in response to some word stimulus. Due to the low signal-to-noise-ratio (SNR) regime, multiple trials are often recorded, presenting a *functional data analysis* scenario. Each trial results in a noisy trajectory with key discontinuities (e.g., after stimulus onset). Although there are overall similarities between the trials, there are also key differences that occur based on various physiological phenomena, as depicted in Fig. 1. We clearly see abrupt changes as well as long-range correlations. Key to the data analysis is the ability to share information about the overall trajectory between the single trials without forcing unrealistic smoothness assumptions on the single trials themselves.

In order to capture both long-range dependencies and potential discontinuities, we propose a multiresolution GP (mGP) that hierarchically couples a collection of smooth GPs, each defined over an element of a nested partition set. The top-level GP captures a smooth global trajectory, while the partition points define abrupt changes in correlation induced by the lower-level GPs. Due to the inherent conjugacy of the GPs, conditioned on the partition points the resulting function at the bottom level is marginally GP-distributed with a partition-dependent (and thus non-stationary) covariance function. The correlation between any two observations y_i and y_j generated by the mGP at locations x_i and x_j is a function of the distance $||x_i - x_j||$ and which partition sets contain both x_i and x_j.

In a standard regression setting, the marginal GP structure of the mGP allows us to compute the marginal likelihood of the data conditioned on the partition, enabling efficient inference of the partition itself. We integrate over the hierarchy of GPs and only sample the partition points. For our

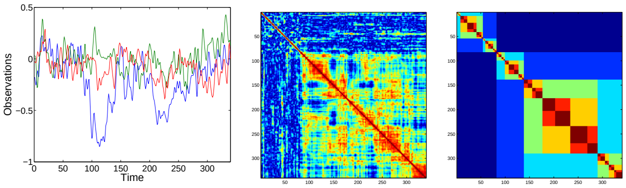

Figure 1: For sensor 1 and word *house*, *Left:* Data from three trials; *Middle:* Empirical correlation matrix from 20 trials; *Right:* Hierarchical segmentation produced by recursive minimization of normalized cut objective, with color indicating tree level.

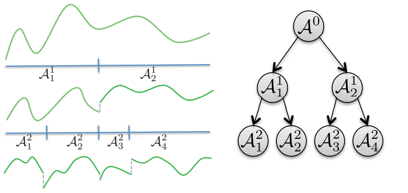

Figure 2: mGP on a balanced, binary tree partition: Parent function is split by $\mathcal{A}^1 = \{\mathcal{A}_1^1, \mathcal{A}_2^1\}$. Recursing down the tree, each partition has a GP with mean given by its parent function restricted to that set.

proposal distribution, we borrow the graph-theoretic idea of *normalized cuts* [22] often used in image segmentation. Our inferences integrate over the partition tree, allowing blurring of discontinuities and producing functions which can appear smooth when discontinuities are not present in the data.

2 Background

A GP provides a distribution on real-valued functions $f : \mathcal{X} \to \Re$, with the property that the function evaluated at any finite collection of points is jointly Gaussian. The GP, denoted $\text{GP}(m, c)$, is uniquely defined by its *mean function* m and *covariance function* c. That is, $f \sim \text{GP}(m, c)$ if and only if for all $n \geq 1$ and $x_1, \ldots, x_n$, $(f(x_1), \ldots, f(x_n)) \sim N_n(\mu, K)$, with $\mu = [m(x_1), \ldots, m(x_n)]$ and $[K]_{ij} = c(x_i, x_j)$. The properties (e.g., continuity, smoothness, periodicity, etc.) of functions drawn from a given GP are determined by the covariance function. The squared exponential kernel, $c(x, x') = d\exp(-\kappa||x - x'||_2^2)$, leads to smooth functions. Here, d is a *scale* hyperparameter and κ is the *bandwidth* determining the extent of the correlation in f over $\mathcal{X}$. See [18] for further details.

3 Multiresolution Gaussian Process Formulation

Our interest is in modeling a function g that (i) is locally smooth, (ii) exhibits long-range correlations (i.e., $\text{corr}(g(x), g(x')) > 0$ for $||x - x'||$ relatively large), and (iii) has abrupt changes. We begin by modeling a single function, but with a specification that readily lends itself to modeling a *collection* of functions that share a common global trajectory, as explored in Sec. 4.

Generative Model Assume a set of noisy observations $y = \{y_1, \ldots, y_n\}$, $y_i \in \Re$, of the function g at locations $\{x_1, \ldots, x_n\}$, $x_i \in \mathcal{X} \subset \Re^p$:

$$y_i = g(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \tag{1}$$

We hierarchically define g as follows. Let $\mathcal{A} = \{\mathcal{A}^0, \mathcal{A}^1, \ldots, \mathcal{A}^{L-1}\}$ be a nested partition, or *tree partition*, of $\mathcal{X}$ with $\mathcal{A}^0 = \mathcal{X}$, $\mathcal{X} = \bigcup_i \mathcal{A}_i^\ell$, $\mathcal{A}_i^\ell \cap \mathcal{A}_j^\ell = \emptyset$, and $\mathcal{A}_i^\ell \subset \mathcal{A}_k^{\ell-1}$ for some k. Furthermore, assume that each $\mathcal{A}_i^\ell$ is a contiguous subset of $\mathcal{X}$. Fig. 2 depicts a balanced, binary tree partition. We define a *global parent function* on $\mathcal{A}^0$ as $f^0 \sim \text{GP}(0, c^0)$. This function captures the overall shape of g and its long-range dependencies. Then, over each partition set $\mathcal{A}_i^\ell$ we independently draw

$$f^\ell(\mathcal{A}_i^\ell) \sim \text{GP}(f^{\ell-1}(\mathcal{A}_i^\ell), c_i^\ell). \tag{2}$$

That is, the mean of the GP is given by the parent function restricted to the current partition set. Due to the conditional independence of these draws, f^ℓ can have discontinuities at the partition points. However, due to the coupling of GPs through the tree, f^ℓ will maintain aspects of the shape of f^0. Finally, we set $g = f^{L-1}$. A pictorial representation of the mGP is shown in Fig. 2.

We can equivalently represent the mGP as an *additive* GP model: $\phi^\ell(\mathcal{A}_i^\ell) \sim \text{GP}(0, c_i^\ell)$, $g = \sum_\ell \phi^\ell$.

Covariance Function We assume a squared exponential kernel $c_i^\ell = d_i^\ell \exp(-\kappa_i^\ell||x - x'||_2^2)$, encouraging local smoothness over each partition set $\mathcal{A}_i^\ell$. We focus on $d_i^\ell = d^\ell$ with $\sum_{\ell=1}^{\infty}(d^\ell)^2 < 1$ for finite variance regardless of tree depth and additionally encouraging lower levels to vary less from their parent function, providing regularization and robustness to the choice of L.

We typically assume bandwidths $\kappa_i^\ell = \kappa/||\mathcal{A}_i^\ell||_2^2$ so that each child function is locally as smooth as its parent. One can think of this formulation as akin to a fractal process: zooming in on any partition, the locally defined function has the same smoothness as that of its parent over the larger partition. Thus, lower levels encode finer-resolution details. We denote the covariance hyperparameters as $\theta = \{d^0, \ldots, d^{L-1}, \kappa\}$, and omit the dependency in conditional distributions for notational simplicity. See the Supplementary Material for discussion of other possible covariance specifications.

Induced Marginal GP The conditional independencies of our mGP imply that

$$p(g \mid \mathcal{A}) = \int p(f^0) \prod_{\ell=1}^{L-1} p(f^\ell \mid f^{\ell-1}, \mathcal{A}^\ell) df^{0:L-2}. \tag{3}$$

Due to the inherent conjugacy of the GPs, one can analytically marginalize the hierarchy of GPs *conditioned on the partition tree* $\mathcal{A}$ yielding

$$g \mid \mathcal{A} \sim \text{GP}(0, c^*_{\mathcal{A}}), \quad c^*_{\mathcal{A}} = \sum_{\ell=0}^{L-1} \sum_i c_i^\ell \mathbb{I}_{\mathcal{A}_i^\ell}. \tag{4}$$

Here, $\mathbb{I}_{\mathcal{A}_i^\ell}(x, x') = 1$ if $x, x' \in \mathcal{A}_i^\ell$ and 0 otherwise. Eq. (4) provides an interpretation of the mGP as a (marginally) partition-dependent GP, where the partition $\mathcal{A}$ defines the discontinuities in the covariance function $c^*_{\mathcal{A}}$. The covariance function encodes local smoothness of g and discontinuities at the partition points. Note that $c^*_{\mathcal{A}}$ defines a *non-stationary* covariance function.

The correlation between any two observations y_i and y_j at locations x_i and x_j generated as in Eq. (1) is a function of how many tree levels contain both x_i and x_j and the distance $||x_i - x_j||$. Let r_i^ℓ index the partition set such that $x_i \in \mathcal{A}^\ell_{r_i^\ell}$ and L_{ij} the lowest level for which x_i and x_j fall into the same set (i.e., the largest ℓ such that $r_i^\ell = r_j^\ell$). Then, for $x_i \neq x_j$,

$$\text{corr}(y_i, y_j \mid \mathcal{A}) = \frac{\sum_{\ell=0}^{L_{ij}} c^\ell_{r_i^\ell}(x_i, x_j)}{\prod_{k \in \{i,j\}} (\sigma^2 + \sum_{\ell=0}^{L-1} c^\ell_{r_k^\ell}(x_k, x_k))^{\frac{1}{2}}} = \frac{\sum_{\ell=0}^{L_{ij}} d^\ell \exp(-\kappa ||x_i - x_j||_2^2 / ||\mathcal{A}^\ell_{r_i^\ell}||_2^2)}{\sigma^2 + \sum_{\ell=0}^{L-1} d^\ell}, \tag{5}$$

where the second equality follows from assuming the previously described kernels. An example correlation matrix is shown in Fig. 3(c). κ determines the width of the bands while d^ℓ controls the contribution of level ℓ. Since d^ℓ is square summable, lower levels are less influential.

Marginal Likelihood Based on a *vector* of observations $\mathbf{y} = [y_1 \cdots y_n]'$ at locations $\mathbf{x} = [x_1 \cdots x_n]'$, we can restrict our attention to evaluating the GPs at $\mathbf{x}$. Let $f^\ell(\mathbf{x}) = [f^\ell(x_1) \cdots f^\ell(x_n)]'$. By definition of the GP, we have

$$f^\ell(\mathbf{x}) \mid f^{\ell-1}(\mathbf{x}), \mathcal{A}^\ell \sim N(f^{\ell-1}(\mathbf{x}), K_\ell), \quad [K_\ell]_{i,j} = \begin{cases} c_r^\ell(x_i, x_j) & x_i, x_j \in \mathcal{A}_r^\ell \\ 0 & otherwise \end{cases}. \tag{6}$$

The level-specific covariance matrix K_ℓ is block-diagonal with structure determined by the level-specific partition $\mathcal{A}^\ell$. Observations are generated as $\mathbf{y} \mid g(\mathbf{x}) \sim N(g(\mathbf{x}), \sigma^2 I_n)$. Recalling Eq. (3), standard results yield

$$g(\mathbf{x}) \mid \mathcal{A} \sim N\left(0, \sum_{\ell=0}^{L-1} K_\ell\right) \quad \mathbf{y} \mid \mathcal{A} \sim N\left(0, \sigma^2 I_n + \sum_{\ell=0}^{L-1} K_\ell\right). \tag{7}$$

This result can also be derived from the induced mGP of Eq. (4). We see that the marginal likelihood $p(\mathbf{y} \mid \mathcal{A})$ has a closed form. Alternatively, one can condition on the GP at any level ℓ':

$$\mathbf{y} \mid f^{\ell'}(\mathbf{x}), \mathcal{A} \sim N\left(f^{\ell'}(\mathbf{x}), \sigma^2 I_n + \sum_{\ell=\ell'+1}^{L-1} K_\ell\right). \tag{8}$$

A key advantage of the mGP is the conditional conjugacy of the latent GPs that allows us to compute the likelihood of the data simply conditioned on the hierarchical partition $\mathcal{A}$ (see Eq. (7)). This fact is fundamental to the efficiency of the partition inference procedure described in Sec. 5.

4 Multiple Trials

In many applications, such as the motivating MEG application, one has a *collection* of observations of an underlying signal. To capture the common global trajectory of these trials while still allowing for trial-specific variability, we model each as a realization from an mGP with a *shared* parent function f^0. One could trivially allow for alternative structures of hierarchical sharing beyond f^0 if an application warranted. For simplicity, and due to the motivating MEG application, we additionally assume shared changepoints between the trials, though this assumption can also be relaxed.

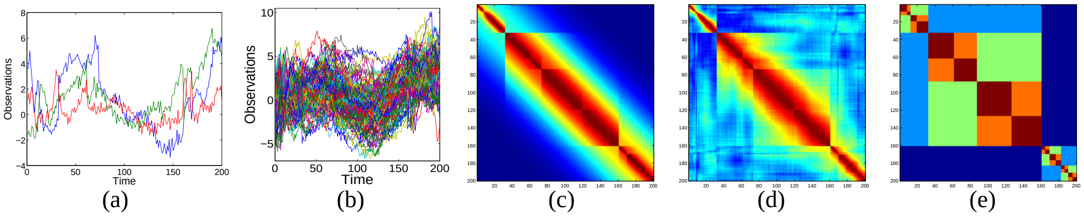

Figure 3: (a) Three trials and (b) all 100 trials of data generated from a 5-level mGP with a shared parent function f^0 and partition $\mathcal{A}$ (randomly sampled). (c) True correlation matrix. (d) Empirical correlation matrix from 100 trials. (e) Hierarchical segmentation produced by recursive minimization of normalized cut objective.

Generative Model For each trial $\mathbf{y}^{(j)} = \{y_1^{(j)}, \ldots, y_n^{(j)}\}$, we model

$$y_i^{(j)} = g^{(j)}(x_i) + \epsilon_i^{(j)}, \quad \epsilon_i^{(j)} \sim N(0, \sigma^2), \tag{9}$$

with $g^{(j)} = f^{L-1,(j)}$ generated from a trial-specific GP hierarchy $f^0 \to f^{1,(j)} \to \cdots \to f^{L-1,(j)}$ with shared parent f^0. (Again, alternative structures can be considered.) From Eq. (8) with $\ell' = 0$, and exploiting the independence of $\{f^{\ell,(j)}\}$, independently for each j

$$\mathbf{y}^{(j)} \mid f^0(\mathbf{x}), \mathcal{A} \sim N\left(\mathbf{y}^{(j)}; f^0(\mathbf{x}), \sigma^2 I_n + \sum_{\ell=1}^{L-1} K_\ell\right). \tag{10}$$

Note that with our GP-based formulation, we need not assume coincident observation locations $x_1, \ldots, x_n$ between the trials. However, for simplicity of exposition, we consider shared locations. We compactly denote the covariance by $\Sigma = \sigma^2 I_n + \sum_{\ell=1}^{L-1} K_\ell$.

Simulated data generated from a 5-level mGP with shared f^0 and $\mathcal{A}$ are shown in Fig. 3. The sample correlation matrix is also shown. Compare with the MEG data of Fig. 1. Both the qualitative structure of the raw time series as well as blockiness of the correlation matrix have striking similarities.

Posterior Global Trajectory and Predictions Based on a set of trials $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}\}$, it is of interest to infer the posterior of f^0. Standard Gaussian conjugacy results imply that

$$p(f^0(\mathbf{x}) \mid \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}, \mathcal{A}) = N\left(\left(K_0^{-1} + J\Sigma^{-1}\right)^{-1} \tilde{\mathbf{y}}, \left(K_0^{-1} + J\Sigma^{-1}\right)^{-1} \right), \tag{11}$$

where $\tilde{\mathbf{y}} = \Sigma^{-1} \sum_i \mathbf{y}^{(i)}$. Likewise, the predictive distribution of data from a new trial is

$$\begin{aligned} p(\mathbf{y}^{(J+1)} \mid \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}, \mathcal{A}) &= \int p(\mathbf{y}^{(J+1)} \mid f^0(\mathbf{x}), \mathcal{A}) p(f^0(\mathbf{x}) \mid \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}, \mathcal{A}) df^0 \\ &= N\left(\left(K_0^{-1} + J\Sigma^{-1}\right)^{-1} \tilde{\mathbf{y}}, \Sigma + \left(K_0^{-1} + J\Sigma^{-1}\right)^{-1} \right). \end{aligned} \tag{12}$$

Marginal Likelihood Since the set of trials $Y = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}\}$ are generated from a shared parent function f^0, the marginal likelihood does not decompose over trials. Instead,

$$p(Y \mid \mathcal{A}) = \frac{|K_0|^{-1/2} |\Sigma|^{-J/2}}{(2\pi)^{-nJ/2} |K_0^{-1} + J\Sigma^{-1}|^{1/2}} \exp\left(-\frac{1}{2} \sum_i \mathbf{y}^{(i)\prime} \Sigma^{-1} \mathbf{y}^{(i)} + \frac{1}{2} \tilde{\mathbf{y}}' (K_0^{-1} + J\Sigma^{-1})^{-1} \tilde{\mathbf{y}} \right). \tag{13}$$

See the Supplementary Material for a derivation. One can easily verify that the above simplifies to the marginal likelihood of Eq. (7) when $J = 1$.

5 Inference of the Hierarchical Partition

In the formulation so far, we have assumed that the hierarchical partition $\mathcal{A}$ is given. A key question is to infer the partition from the data. Assume that we have prior $p(\mathcal{A})$ on the hierarchical partition. Based on the fact that we can analytically compute $p(Y \mid \mathcal{A})$, we can use importance sampling or independence chain Metropolis Hastings to draw samples from the posterior $p(\mathcal{A} \mid Y)$.

In what follows, we assume a balanced binary tree for $\mathcal{A}$. See the Supplementary Material for a discussion of how unbalanced trees can be considered via modifications to the covariance hyperparameter specification or by considering alternative priors $p(\mathcal{A})$ such as the Mondrian process [20].

Partition Prior We consider a prior solely on the partition points $\{z_1, \dots, z_{2^{L-1}-1}\}$ rather than taking tree level into account as well. Because of our time-series analysis focus, we assume $\mathcal{X} \subset \Re$. We define a distribution F on $\mathcal{X}$ and specify $p(\mathcal{A}) = \prod_i F(z_i)$. Generatively, one can think of drawing $2^{L-1} - 1$ partition points from F and deterministically forming a balanced binary tree $\mathcal{A}$ from these. For multidimensional $\mathcal{X}$, one could use Voronoi tessellation and graph matching to build the tree from the randomly selected z_i. Such a prior allows for trivial specification of a uniform distribution on $\mathcal{A}$ (simply taking F uniform on $\mathcal{X}$) or for eliciting prior information on changepoints, such as based on physiological information for the MEG data. Eliciting such information in a level-dependent setup is not straightforward. Also, despite common deployment, taking the partition point at level ℓ as uniformly distributed over the parent set $\mathcal{A}_i^{\ell-1}$ yields high mass on $\mathcal{A}$ with small A_i^{ℓ}. This property is undesirable because it leads to trees with highly unbalanced partitions.

Our resulting inferences perform Bayesian model averaging over trees. As such, even though we specify a prior on partitions with $2^{L-1} - 1$ changepoints, the resulting functions can appear to adaptively use fewer by averaging over the uncertainty in the discontinuity location.

Partition Proposal Although stochastic tree search algorithms tend to be inefficient in general, we can harness the well-defined correlation structure associated with a given hierarchical partition to much more efficiently search the tree space. One can think of every observed location x_i as a node in a graph with edge weights between x_i and x_j defined by the magnitude of the correlation of y_i and y_j. Based on this interpretation, the partition points of $\mathcal{A}$ correspond to graph cuts that bisect small edge weights, as graphically depicted in Fig. 4. As such, we seek a method for hierarchically cutting a graph. Given a cost matrix W with elements w_{uv} defined for all pairs of nodes u, v in a set V, the *normalized cut* metric [22] for partitioning V into disjoint sets A and B is given by

$$\text{ncut}(A, B) = \text{cut}(A, B) \left[\text{assoc}(A, V)^{-1} + \text{assoc}(B, V)^{-1}\right], \tag{14}$$

where $\text{cut}(A, B) = \sum_{u \in A, v \in B} w_{uv}$ and $\text{assoc}(A, V) = \sum_{u \in A, v \in V} w_{uv}$. Typically, the cut point is selected as the minimum of the metric $\text{ncut}(A, B)$ computed over all possible subsets A and B. The normalized cut metric balances between the cost of edge weights cut and the connectivity of the cut component, thus avoiding cuts that separate small sets. Fig. 1 shows an example of applying a greedy normalized cuts algorithm (recursively minimizing $\text{ncut}(A, B)$) to MEG data.

Instead of deterministically selecting cut points, we employ the normalized cut objective as a proposal distribution. Let the cost matrix W be the absolute value of the empirical correlation matrix computed from trials $\{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(J)}\}$ (see Fig. 1). Due to the natural ordering of our locations $x_i \in \mathcal{X} \subset \Re$, the algorithm is straightforwardly implemented. We step down the hierarchy, first proposing a cut of $\mathcal{A}^0$ into $\{\mathcal{A}_1^1, \mathcal{A}_2^1\}$ with probability

$$q(\{\mathcal{A}_1^1, \mathcal{A}_2^1\}) \propto \text{ncut}(\mathcal{A}_1^1, \mathcal{A}_2^1)^{-1}. \tag{15}$$

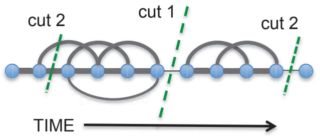

Figure 4: Illustration of cutpoints dividing contiguous segments at points of low correlation.

At level ℓ, each $\mathcal{A}_i^{\ell}$ is partitioned via a normalized cut proposal based on the submatrix of W corresponding to the locations $x_i \in A_i^{\ell}$. The probability of any partition $\mathcal{A}$ under the specified proposal distribution is simply computed as the product of the sequence of conditional probabilities of each cut. This procedure generates cut points only at the observed locations x_i. More formally, the partition point in $\mathcal{X}$ is proposed as uniformly distributed between x_i and x_{i+1}. Extensions to multi-dimensional $\mathcal{X}$ rely on spectral clustering algorithms based on the graph Laplacian [24].

Markov Chain Monte Carlo An importance sampler draws hierarchical partitions $\mathcal{A}^{(m)} \sim q$, with the proposal distribution q defined as above, and then weights the samples by $p(\mathcal{A}^{(m)})/q(\mathcal{A}^{(m)})$ to obtain posterior draws [19]. Such an approach is naively parallelizable, and thus amenable to efficient computations, though the effective sample size may be low if q does not adequately match the posterior $p(\mathcal{A} \mid Y)$. Alternatively, a straightforward independence chain Metropolis Hastings algorithm (see Supplementary Material) is defined by iteratively proposing $\mathcal{A}' \sim q$ which is accepted with probability $\min\{r(\mathcal{A}' \mid \mathcal{A}), 1\}$ where $\mathcal{A}$ is a previous sample of a hierarchical partition and

$$r(\mathcal{A}' \mid \mathcal{A}) = p(Y \mid A')p(A')q(A)/[p(Y \mid A)p(A)q(A')]. \tag{16}$$

The tailoring of the proposal distribution q to this application based on normalized cuts dramatically aids in improving the acceptance rate relative to more naive tree proposals. However, the acceptance rate tends to decrease as higher posterior probability partitions $\mathcal{A}$ are discovered, especially for trees with many levels and large input spaces $\mathcal{X}$ for which the search space is larger.

One benefit of the MCMC approach over importance sampling is the ability to include more intricate tree proposals to increase efficiency. We choose to interleave both local and global tree proposals. At each iteration, we first randomly select a node in the tree (i.e., a partition set $\mathcal{A}_i^\ell$) and then propose a new sequence of cuts for all children of this node. When the root node is selected, corresponding to $\mathcal{A}^0$, the proposal is equivalent to the global proposals previously considered. We adapt the proposal distribution for node selection to encourage more global searches at first and then shift towards a greater balance between local and global searches as the sampling progresses. Sequential Monte Carlo methods [4] can also be considered, with particles generated as global proposals.

Computational Complexity The per iteration complexity is $O(n^3)$, equivalent to a typical likelihood evaluation under a GP prior. Using dynamic programming, the cost associated with the normalized cuts proposal is $O(n^2(L-1))$. Standard techniques for more efficient GP computations are readily applicable, as well as extensions that harness the additive block structure of the covariance.

6 Related Work

Various aspects of the mGP have similarities to other models proposed in the literature that primarily fall into two main categories: (i) GPs defined over a partitioned input space, and (ii) collections of GPs defined at tree nodes. The treed GP [8] captures non-stationarities by defining independent GPs at the leaves of a Bayesian CART-partitioned input space. The related approach of [12] assumes a Voronoi tessellation. For time series, [21] examines online inference of changepoints with GPs modeling the data within each segment. These methods capture abrupt changes, but do not allow for long-range dependencies spanning changepoints nor a functional data hierarchical structure, both inherent to our multiresolution perspective. A main motivation of the treed GP is the resulting computational speed-ups of an independently partitioned GP. A two-level hierarchical GP also aimed at computational efficiency is considered by [16], where the top-level GP is defined at a coarser scale and provides a piece-wise constant mean for lower-level GPs on a pre-partitioned input space.

[10, 11] consider covariance functions defined on a phylogenetic tree such that the covariance between function-valued traits depends on both their spatial distance and evolutionary time spanned via a common ancestor. Here, the tree defines the strength and structure of sharing between a collection of functions rather than abrupt changes within the function. The Bayesian rose tree of [3] considers a mixture of GP experts, as in [14, 17], but using Bayesian hierarchical clustering with arbitrary branching structure in place of a Dirichlet process mixture. Such an approach is fundamentally different from the mGP: each GP is defined over the entire input space, data result from a GP mixture, and input points are not necessarily spatially clustered. Alternatively, multiscale processes have a long history (cf. [25]): the variables define a Markov process on a typically balanced, binary tree and higher-level nodes capture coarser level information about the process. In contrast, the higher level nodes in the mGP share the same temporal resolution and only vary in smoothness.

At a high level, the mGP differs from previous GP-based tree models in that the nodes of our tree represent GPs over a contiguous subset of the input space $\mathcal{X}$ constrained in a hierarchical fashion. Thus, the mGP combines ideas of GP-based tree models and GP-based partition models.

As presented in Sec. 3, one can formulate an mGP as an additive GP where each GP in the sum decomposes independently over the level-specific partition of the input space $\mathcal{X}$. The additive GPs of [6] instead focus on coping with multivariate inputs, in a similar vain to hierarchical kernel learning [1], thus addressing an inherently different task.

7 Results

7.1 Synthetic Experiments

To assess our ability to infer a hierarchical partition via the proposed MCMC sampler, we generated 100 trials of length 200 from a 5-level mGP with a shared parent function f^0. The hyperparameters were set to $\sigma^2 = 0.1$, $\kappa = 10$, $d^\ell = d^0 \exp(-0.5(\ell+1))$ for $\ell = 0, \dots, L-1$ with $d^0 = 5$. The data are shown in Fig. 3, along with the empirical correlation matrix that is used as the cost matrix for the normalized cuts proposals.

For inference, we set $\sigma^2 = \hat{\sigma}^2/3$ and $d^\ell = (\hat{\sigma}^2/3)\exp(-0.5\ell)$, where $\hat{\sigma}^2$ is the average time-specific sample variance. κ was as in the simulation. The hyperparameter mismatch demonstrates

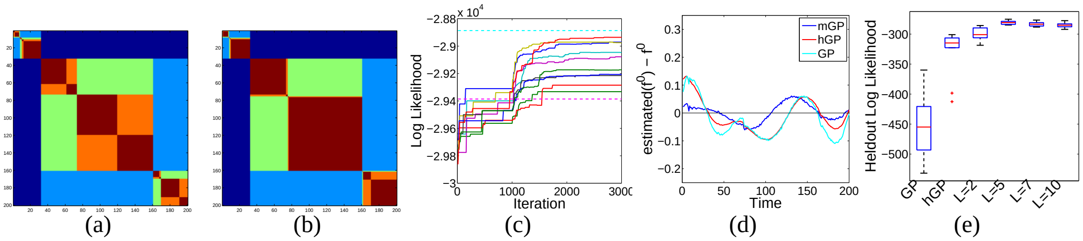

Figure 5: For the data of Fig. 3, (a) true and (b) MAP partitions. (c) Trace plots of log likelihood versus MCMC iteration for 10 chains. Log likelihood under the true partition (*cyan*) and minimized normalized cut partition of Fig. 3 (*magenta*) are also shown. (d) Errors between posterior mean f^0 and true f^0 for GP, hGP, and mGP. (e) Predictive log likelihood of 10 heldout sequences for GP, hGP, and mGP with $L = 2, 5(true), 7, 10$.

some robustness to mispecification. For a uniform prior $p(\mathcal{A})$, 10 independent MCMC chains were run for 3000 iterations, thinned by 10. The first 1000 iterations used pure global tree searches; the sampler was then tempered to uniform node proposals. The effects of this choice are apparent in the likelihood plot of Fig. 5, which also displays the true hierarchical partition and MAP estimate. Compare to the normalized cuts partition of Fig. 3, especially at the important level 1 cut. The full simulation study took less than 7 minutes to run on a single 1.8 GHz Intel Core i7 processor.

To assess sensitivity to the choice of L, we compare the predictive log-likelihood of 10 heldout test sequences under an mGP with 2, 5, 7, and 10 levels. As shown in Fig. 5(e), there is a clear gain going from 2 to 5 levels. However, overestimating L has minimal influence on predictive likelihood since lower tree levels capture finer details and have less overall effect. We also compare to a single GP and a 2-level hierarchical GP (hGP) (see Sec. 7.2). For a direct comparison, both use squared exponential kernels. Hyperparameters were set as in the mGP for the top-level GP. The total variance was also matched with the GP taking this as noise and the hGP splitting between level 2 and noise. In addition to better predictive performance, Fig. 5(d) shows the mGP's improved estimation of f^0.

7.2 MEG Analysis

We analyzed magnetoencephalography (MEG) recordings of neuronal activity collected from a helmet with gradiometers distributed over 102 locations around the head. The gradiometers measure the spatial gradient of the magnetic activity in Teslas per meter (T/m) [9]. Since the firings of neurons in the brain only induce a weak magnetic field outside of the skull, the signal-to-noise ratio of the MEG data is very low and typically multiple recordings, or *trials*, of a given task are collected. Our MEG data was recorded while a subject viewed 20 stimuli describing concrete nouns (both the written noun and a representative line drawing), with 20 interleaved trials per word. See the Supplementary Material for further details on the data and our analyses presented herein.

Efficient sharing of information between the single trials is important for tasks such as word classification [7]. A key insight of [7] was the importance of capturing the time-varying correlations between MEG sensors for performing classification. However, the formulation still necessitates a mean model. [7] propose a 2-level hierarchical GP (hGP): a parent GP captures the common global trajectory, as in the mGP, and each trial-specific GP is centered about the entire parent function[1]. This formulation maintains global smoothness at the individual trial level. The mGP instead models the trial-specific variability with a multi-level tree of GPs defined as deviations from the parent function over local partitions, allowing for abrupt changes relative to the smooth global trajectory.

For our analyses, we consider the words associated with the "building" and "tool" categories shown in Fig. 7. Independently for each of the 10 words and 102 sensors, we trained a 5-level mGP using 15 randomly selected trials as training data and the 5 remaining for testing. Each trial was of length $n = 340$. We ran 3 independent MCMC chains for 3000 iterations with both global and local tree searches. We discarded the first 1000 samples as burn-in and thinned by 10. The mGP hyperparameters were set exactly as in the simulated study of Sec. 7.1 for structure learning and then optimized over a grid to maximize the marginal likelihood of the training data.

We compare the predictive performance of the mGP in terms of MSE of heldout segments relative to a GP and hGP, each with similarly optimized hyperparameters. The predictive mean conditioned on data up to the heldout time is straightforwardly derived from Eq. (12). For the mGP, the calculation is averaged over the posterior samples of $\mathcal{A}$. Fig. 6 displays the MSEs decomposed by cortical region.

[1] The model of [7] uses an hGP in a latent space. The mGP could be similarly deployed.

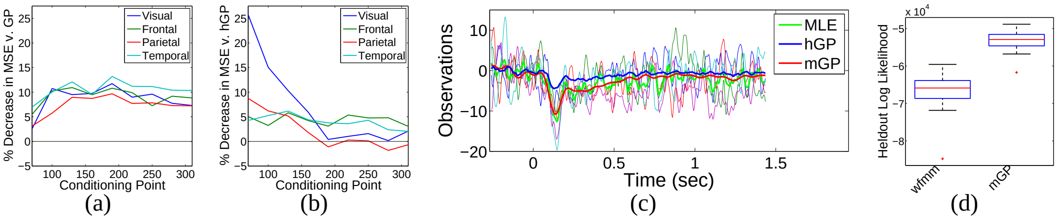

Figure 6: Per-lobe comparison of mGP to (a) GP and (b) hGP: For various values of τ, % decrease in predictive MSE of heldout $y^*_{\tau:\tau+30}$ conditioned on $y^*_{1:\tau-1}$ and 15 training sequences. (c) For a visual cortex sensor and word *hammer*, plots of test data, empirical mean (MLE), and hGP and mGP predictive mean for entire heldout $\mathbf{y}^*$. (d) Boxplots of predictive log likelihood of heldout $\mathbf{y}^*$ for the mGP and wavelet-based method of [15].

The results clearly indicate that the mGP consistently better captures the features of the data, and particularly for sensors with large abrupt changes such as in the visual cortex. The heldout trials for a visual cortex sensor are displayed in Fig. 6(c). Relative to the hGP, the mGP much better tracks the early dip in activity right after the visual stimulus onset ($t = 0$). The posterior distribution of inferred changepoints at level 1, also broken down by cortical region, are displayed in Fig. 7. As expected, the visual cortex has the earliest changepoints. Similar trends are seen in the parietal lobe that handles perception and sensory integration. The temporal lobe, which is key in semantic processing, has changepoints occurring later. These results concur with the findings of [23]: semantic processing starts between 250 and 600 ms and word length (a visual feature) is decoded most accurately very near the standard 100ms response time (“n100”).

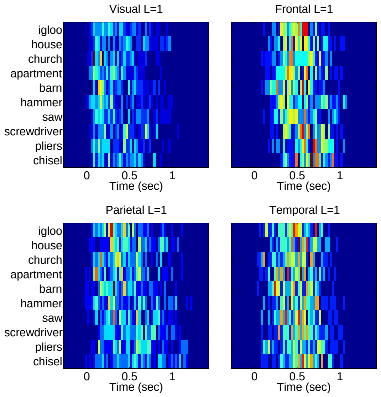

Figure 7: Inferred changepoints at level 1 aggregated over sensors within each lobe: visual (*top-left*), frontal (*top-right*), parietal (*bottom-left*), and temporal (*bottom-right*).

We also compare our predictive performance to that of the wavelet-based functional mixed model (wfmm) of [15]. The wfmm has become a standard approach for functional data analysis since it allows for spiky trajectories and efficient sharing of information between trials. One limitation, however, is the restriction to a regular grid of observations. The wfmm enables analysis in a multivariate setting, but for a direct comparison we simply apply the wfmm to each word and sensor independently. Fig. 6(d) shows boxplots of the predictive heldout log likelihood of the test trials under the mGP and wfmm. The results are over 5 heldout trials, 102 sensors, and 10 words. In addition to the easier interpretability of the mGP, the predictive performance also exceeds that of the wfmm.

8 Discussion

The mGP provides a flexible framework for characterizing the dependence structure of real data, such as the examined MEG recordings, capturing certain features more accurately than previous models. In particular, the mGP provides a hierarchical functional data analysis framework for modeling (i) strong, locally smooth sharing of information, (ii) global long-range correlations, and (iii) abrupt changes. The simplicity of the mGP formulation enables further theoretical analysis, for example, combining posterior consistency results from changepoint analysis with those for GPs. Although we focused on univariate time series analysis, our formulation is amenable to multivariate functional data analysis extensions: one can naturally accommodate hierarchical dependence structures through partial sharing of parents in the tree, or possibly via mGP factor models.

There are many interesting questions relating to the proposed covariance function. Our fractal specification represents a particular choice to avoid over-parameterization, although alternatives could be considered. For hyperparameter inference, we anticipate that joint sampling with the partition would mix poorly, and consider it a topic for future exploration. Another interesting topic is to explore proposals for more general tree structures. We believe that the proposed mGP represents a powerful, broadly applicable new framework for non-stationary analyses, especially in a functional data analysis setting, and sets the foundation for many interesting possible extensions.

Acknowledgments

The authors thank Alona Fyshe, Gustavo Sudre and Tom Mitchell for their help with data acquisition, preprocessing, and useful suggestions. This work was supported in part by AFOSR Grant FA9550-12-1-0453 and the National Institute of Environmental Health Sciences (NIEHS) of the NIH under Grant R01 ES017240.

References

[1] F. Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical Report 0909.0844v1, arXiv, 2009.

[2] J. Beran and Y. Shumeyko. On asymptotically optimal wavelet estimation of trend functions under long-range dependence. *Bernoulli*, 18(1):137–176, 2012.

[3] C. Blundell, Y. W. Teh, and K. A. Heller. Bayesian rose trees. In *Proc. Uncertainty in Artificial Intelligence*, pages 65–72, 2010.

[4] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B*, 68(3):411–436, 2006.

[5] F. X. Diebold and G. D. Rudebusch. Long memory and persistence in aggregate output. *Journal of Monetary Economics*, 24:189–209, 1989.

[6] D. Duvenaud, H. Nickisch, and C. E. Rasmussen. Additive Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 24, pages 226–234, 2011.

[7] A. Y. Fyshe, E. B. Fox, D. B. Dunson, and T. Mitchell. Hierarchical latent dictionaries for models of brain activation. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 409–421, 2012.

[8] R. .B. Gramacy and H. K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[9] P. Hansen, M. Kringelbach, and R. Salmelin. *MEG: An Introduction to Methods*. Oxford University Press, USA, 2010. ISBN 0195307232.

[10] R. Henao and J. E. Lucas. Efficient hierarchical clustering for continuous data. Technical Report 1204.4708v1, arXiv, 2012.

[11] N. S. Jones and J. Moriarty. Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. Technical Report 1004.4668v2, arXiv, 2011.

[12] H. M. Kim, B. K. Mallick, and C. C. Holmes. Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668, 2005.

[13] P. S. Kokoszka and M. S. Taqqu. Parameter estimation for infinite variance fractional ARIMA. *The Annals of Statistics*, 24(5):1880–1913, 1996.

[14] E. Meeds and S. Osindero. An alternative mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems*, volume 18, pages 883–890, 2006.

[15] J. S. Morris and R. J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68(2):179–199, 2006.

[16] S. Park and S. Choi. Hierarchical Gaussian process regression. In *Proc. Asian Conference on Machine Learning*, pages 95–110, 2010.

[17] C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems*, volume 2, pages 881–888, 2002.

[18] C. E. Rasmussen and C. K. .I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[19] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2005.

[20] D. M. Roy and Y. W. Teh. The Mondrian process. In *Advances in Neural Information Processing Systems*, volume 21, pages 1377–1384, 2009.

[21] Y. Saatci, R. Turner, and C. E. Rasmussen. Gausssian process change point models. In *Proc. International Conference on Machine Learning*, pages 927–934, 2010.

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[23] G. Sudre, D. Pomerleaum, M. Palatucci, L. Wehbe, A. Fyshe, R. Salmelin, and T. Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *Neuroimage*, 62(1):451–463, 2012.

[24] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[25] A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, 2002.