Multiresolution Gaussian Processes Supplementary Material

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

1 Derivation of Marginal Likelihood

We derive the marginal likelihood as follows. Throughout, we use $f^0(\mathbf{x})$ to compactly denote $[f^0(x_1) \cdots f^0(x_n)]'$. As discussed in the paper, each trial can be described as an independent draw

$$\mathbf{y}^{(j)} \mid f^0(\mathbf{x}), \mathcal{A} \sim N\Big(\mathbf{y}^{(j)}; f^0(\mathbf{x}), \Sigma\Big), \tag{1}$$

where $\Sigma = \sigma^2 I_n + \sum_{\ell=1}^{L-1} K_\ell$ and $f^0(\mathbf{x}) \sim N(0, K_0)$. Therefore, the joint distribution of $Y = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}\}$ is given by:

$$\begin{aligned}
&p(Y \mid f^0(\mathbf{x}), \mathcal{A}) p(f^0(\mathbf{x})) \\
&\quad = c_1^J c_2 \exp\left(-\frac{1}{2}\sum_i (\mathbf{y}^{(i)} - f^0(\mathbf{x}))'\Sigma^{-1}(\mathbf{y}^{(i)} - f^0(\mathbf{x})) - \frac{1}{2} f^0(\mathbf{x})' K_0^{-1} f^0(\mathbf{x})\right) &(2)\\
&\quad = c_1^J c_2 \exp\left(-\frac{1}{2}\sum_i \mathbf{y}^{(i)'}\Sigma^{-1}\mathbf{y}^{(i)} + f^0(\mathbf{x})'\Sigma^{-1}\sum_i \mathbf{y}^{(i)} - \frac{1}{2} f^0(\mathbf{x})'(J\Sigma^{-1} + K_0^{-1}) f^0(\mathbf{x})\right) &(3)\\
&\quad = c_1^J c_2 \exp\Big(-\frac{1}{2}\sum_i \mathbf{y}^{(i)'}\Sigma^{-1}\mathbf{y}^{(i)} - \frac{1}{2}(f^0(\mathbf{x}) - \phi)'(J\Sigma^{-1} + K_0^{-1})(f^0(\mathbf{x}) - \phi) \\
&\qquad\qquad + \frac{1}{2}\phi'(J\Sigma^{-1} + K_0^{-1})\phi\Big), &(4)
\end{aligned}$$

where $c_1 = (2\pi)^{-n/2}|\Sigma|^{-1/2}$, $c_2 = (2\pi)^{-n/2}|K_0|^{-1/2}$, and $\phi = (J\Sigma^{-1} + K_0^{-1})^{-1}\Sigma^{-1}\sum_i \mathbf{y}^{(i)}$. Then, $p(Y \mid \mathcal{A}) = \int p(Y \mid f^0(\mathbf{x}), \mathcal{A}) p(f^0(\mathbf{x})) df^0(\mathbf{x})$ is derived as

$$\begin{aligned}
p(Y \mid \mathcal{A}) &= c_1^J c_2 \exp\left(-\frac{1}{2}\sum_i \mathbf{y}^{(i)'}\Sigma^{-1}\mathbf{y}^{(i)} + \frac{1}{2}\phi'(J\Sigma^{-1} + K_0^{-1})\phi\right) \\
&\qquad \int \exp\left(-\frac{1}{2}(f^0(\mathbf{x}) - \phi)'(J\Sigma^{-1} + K_0^{-1})(f^0(\mathbf{x}) - \phi)\right) df^0(\mathbf{x}) &(5)\\
&= c_1^J c_2 \exp\left(-\frac{1}{2}\sum_i \mathbf{y}^{(i)'}\Sigma^{-1}\mathbf{y}^{(i)} + \frac{1}{2}\phi'(J\Sigma^{-1} + K_0^{-1})\phi\right)(2\pi)^{n/2}|J\Sigma^{-1} + K_0^{-1}|^{-1/2} &(6)\\
&= (2\pi)^{-nJ/2}|K_0|^{-1/2}|\Sigma|^{-J/2}|J\Sigma^{-1} + K_0^{-1}|^{-1/2} \\
&\qquad \exp\left(-\frac{1}{2}\sum_i \mathbf{y}^{(i)'}\Sigma^{-1}\mathbf{y}^{(i)} + \frac{1}{2}\phi'(J\Sigma^{-1} + K_0^{-1})\phi\right). &(7)
\end{aligned}$$

2 MCMC Sampler Pseudocode

We assume (i) a cost matrix W taken to be the absolute value of the empirical correlation matrix of a set of trials, (ii) a prior F on the partition points, and (iii) hyperparameters $\theta = \{\kappa, d^0, \ldots, d^{L-1}, \sigma^2\}$ defining the mGP kernel bandwith, variances, and nugget noise. The sampler is initialized with a hierarchical partition $\mathcal{A}$ drawn from the normalized cuts proposal q.

The covariance matrix $\Sigma = \sigma^2 I_n + \sum_{\ell=1}^{L-1} K_\ell$ (as defined in Sec. 4 of the main paper) is a deterministic function of the hierarchical partition $\mathcal{A}$ and the hyperparameters θ. In what follows, we use `kernel` to define a function that provides this mapping for a given partition set at a given tree level. The likelihood $p(Y \mid \Sigma, \theta) = p(Y \mid \mathcal{A}, \theta)$ is computed exactly as in Eq. (13) of the main paper.

Algorithm 1 details the global search iterations and Algorithm 2 the local (which can also produce global searches if the root note is selected). The local search algorithm additionally assumes a node-proposal distribution indicated by `nodeproposal`.

Algorithm 1 One Iteration of mGP MCMC Sampler - GLOBAL SEARCH

Input: Cost matrix W, hyperparameters θ, previous partition $\mathcal{A}$ and corresponding Σ

$\{z_1, \ldots z_{2^{L-1}-1}\} \leftarrow$ partition points of $\mathcal{A}$

$\mathcal{A}'^0 = \mathcal{X}$, $\Sigma' = 0_{n \times n}$ initialize structures for proposal

for $\ell = 1, \ldots, L-1$ **do**

  for $\nu = 1 : 2 : 2^\ell$ **do**

    $\{\mathcal{A}'^\ell_\nu, \mathcal{A}'^\ell_{\nu+1}\} \sim q(\cdot \mid \mathcal{A}'^{\ell-1}_{(\nu+1)/2}, W)$ normalized cut proposal

    $\Sigma'(\mathcal{A}'^\ell_\nu) = \Sigma'(\mathcal{A}'^\ell_\nu) + \texttt{kernel}(\mathcal{A}'^\ell_\nu, \theta, \ell)$ add K_ℓ submatrix corresponding to $\mathcal{A}'^\ell_\nu$

    $\Sigma'(\mathcal{A}'^\ell_{\nu+1}) = \Sigma'(\mathcal{A}'^\ell_{\nu+1}) + \texttt{kernel}(\mathcal{A}'^\ell_{\nu+1}, \theta, \ell)$ add K_ℓ submatrix corresponding to $\mathcal{A}'^\ell_{\nu+1}$

$\Sigma' = \Sigma' + \sigma^2 I_n$

$\{z'_1, \ldots z'_{2^{L-1}-1}\} \leftarrow$ partition points of $\mathcal{A}'$

$\rho \sim \text{Ber}(\min(r(\mathcal{A}' \mid \mathcal{A}), 1)), \quad r(\mathcal{A}' \mid \mathcal{A}) = \frac{p(Y|\Sigma',\theta) \prod_i F(z'_i) \prod_{\nu_{odd},\ell} q(\{\mathcal{A}'^\ell_\nu, \mathcal{A}'^\ell_{\nu+1}\} | \mathcal{A}'^{\ell-1}_{(\nu+1)/2}, W)}{p(Y|\Sigma,\theta) \prod_i F(z_i) \prod_{\nu_{odd},\ell} q(\{\mathcal{A}'^\ell_\nu, \mathcal{A}'^\ell_{\nu+1}\} | \mathcal{A}'^{\ell-1}_{(\nu+1)/2}, W)}$

$\mathcal{A} \leftarrow \rho\mathcal{A}' + (1-\rho)\mathcal{A}, \quad \Sigma \leftarrow \rho\Sigma' + (1-\rho)\Sigma$ accept or reject proposal

Output: $\mathcal{A}, \Sigma$

3 Alternative Covariance and Tree Specifications

The ideas underlying the mGP readily extend to other covariance hyperparameter specifications beyond the one presented in the main paper (see Sec. 3). For example, instead of assuming a single bandwidth parameter κ scaled by the partition size, one could instantiate a pool of bandwidths that are selected between by each partition set. Likewise, we could introduce a set of partition-specific scale parameters d^ℓ_i. Of course, these changes come at the cost of performing inference over the assignments as well as issues associated with identifiability.

For our hierarchical partition prior and proposal, we assume balanced, binary trees as a simplifying assumption. Working within such a framework, one could allow for unbalanced trees as follows. First, force siblings in the tree to share scale parameters: $d^\ell_i = d^\ell_{i+1}$ for i odd. Then, place a spike and slab prior on d^ℓ_i. If $d^\ell_i = d^\ell_{i+1} = 0$, then the functions defined on the associated partition sets will *exactly* equal the parent function, making it as if that split were not there.

For unbalanced tree priors, one could employ a Mondrian process [2]; splits of the input space continue along branches as long as a budget has not been exhausted. Likewise, one could consider schemes similar to those employed in the randomized and unbalanced partitions of the optional Polya tree [7] (tailored there to the density estimation task). For unbalanced tree proposals, we could introduce randomized stopping criterion to the normalized cuts proposals based on the extent of the drop in correlation at each stage. Just like in image segmentation, we could also use normalized cuts to produce a variable number of splits at every level leading to non-binary tree proposals.

Algorithm 2 One Iteration of mGP MCMC Sampler - LOCAL SEARCH

Input: Cost matrix W, hyperparameters θ, previous partition $\mathcal{A}$ and corresponding Σ

$\{z_1, \dots z_{2^{L-1}-1}\} \leftarrow$ partition points of $\mathcal{A}$

$\Sigma' \leftarrow \Sigma, \quad \mathcal{A}' \leftarrow \mathcal{A}$ — initialize proposals to previous values

$\mathcal{A}^{\ell*}_{\nu*} \sim \texttt{nodeproposal}$ — select a set (tree node) to repartition

$S \leftarrow \{(\nu, \ell) \mid \mathcal{A}'^{\ell}_{\nu} \subset \mathcal{A}^{\ell*}_{\nu*}\}$ — node descendants

for $(\nu, \ell) \in S$ **do**

$\Sigma'(\mathcal{A}'^{\ell}_{\nu}) = \Sigma'(\mathcal{A}'^{\ell}_{\nu}) - \texttt{kernel}(\mathcal{A}'^{\ell}_{\nu}, \theta, \ell)$ — remove contributions from node descendants

for $(\nu, \ell) \in S$ such that ν is odd **do**

$\{\mathcal{A}'^{\ell}_{\nu}, \mathcal{A}'^{\ell}_{\nu+1}\} \sim q(\cdot \mid \mathcal{A}'^{\ell-1}_{(\nu+1)/2}, W)$ — normalized cut proposal

$\Sigma'(\mathcal{A}'^{\ell}_{\nu}) = \Sigma'(\mathcal{A}'^{\ell}_{\nu}) + \texttt{kernel}(\mathcal{A}'^{\ell}_{\nu}, \theta, \ell)$ — add K_ℓ submatrix corresponding to $\mathcal{A}'^{\ell}_{\nu}$

$\Sigma'(\mathcal{A}'^{\ell}_{\nu+1}) = \Sigma'(\mathcal{A}'^{\ell}_{\nu+1}) + \texttt{kernel}(\mathcal{A}'^{\ell}_{\nu+1}, \theta, \ell)$ — add K_ℓ submatrix corresponding to $\mathcal{A}'^{\ell}_{\nu+1}$

$\{z'_1, \dots z'_{2^{L-1}-1}\} \leftarrow$ partition points of $\mathcal{A}'$

$$\rho \sim \text{Ber}(\min(r(\mathcal{A}' \mid \mathcal{A}), 1)), \;\; r(\mathcal{A}' \mid \mathcal{A}) = \frac{p(Y|\Sigma',\theta)\prod_i F(z'_i)\prod_{(\nu_{odd},\ell)\in S} q(\{\mathcal{A}'^{\ell}_{\nu}, \mathcal{A}'^{\ell}_{\nu+1}\}|\mathcal{A}'^{\ell-1}_{(\nu+1)/2}, W)}{p(Y|\Sigma,\theta)\prod_i F(z_i)\prod_{(\nu_{odd},\ell)\in S} q(\{\mathcal{A}'^{\ell}_{\nu}, \mathcal{A}'^{\ell}_{\nu+1}\}|\mathcal{A}'^{\ell-1}_{(\nu+1)/2}, W)}$$

$\mathcal{A} \leftarrow \rho\mathcal{A}' + (1-\rho)\mathcal{A}, \quad \Sigma \leftarrow \rho\Sigma' + (1-\rho)\Sigma$ — accept or reject proposal

Output: $\mathcal{A}, \Sigma$

Another possible extension is to utilize GPs pinned to 0 at the endpoints to define the ϕ^ℓ of the additive mGP specification outlined in Sec. 3 of the main paper. This would enable changes in correlation (i.e., non-stationarity) without introducing discontinuities in the resulting function g. Recall, however, that our inferences over the hierarchical partition allow for blurring of discontinuities, producing functions which can appear smooth if discontinuities are not present in the data.

4 Details on MEG Experiments

In what follows, we expound upon the data collection, prior settings, and hyperparameter optimization for our MEG experiments. We also provide additional figures to supplement the results of the main paper. Recall that for the MEG experiments, the data from each word w and sensor p were treated independently. That is, each assumed a unique partition structure for the mGP. Additionally, for all models the hyperparameters were set in a training-data-driven fashion, as outlined below.

4.1 MEG Data Acquisition

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. 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 at the beginning of each session we recorded the position of the participants head with four head position indicator (HPI) coils placed on the subject's scalp. The HPI coils, along with three cardinal points (nasian, left and right pre-auricular), were digitized into the system.

The data were preprocessed using the Signal Space Separation method (SSS) [4, 5] and temporal extension of SSS (tSSS) [3] 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 down-sampled to 200 Hz. The Signal Space Projection method (SSP) [6] 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 correct 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 multiply the signal by 10^{12} to avoid numerical precision problems.

4.2 MEG Prior Settings

The hierarchical partition prior $p(\mathcal{A})$, determined by F on $\mathcal{X}$ as described in Sec. 5 of the main paper, was set as follows for a given word w and sensor p. All of the training trials associated with sensor p except for those of the considered word were used to produce a recursively minimized normalized cut partition for a 4-level tree. The associated strength of each cut (i.e., amount of empirical correlation cut) was also recorded. F was then defined as a kernel-smoothed version of the cut points and associated cut strengths, along with baseline mass at all points. This prior was used to mimic the information that might be garnered from a domain expert. Experiments were also run under a uniform prior and produced nearly identical results after burn-in. The aggregated posterior changepoints samples, depicted in Fig. 7 only for level 1, were clearly different from the prior setting, demonstrating learning of the partition points (not dominated by the prior).

4.3 Hyperparameter Optimization

The following describes how the hyperparameter optimization is performed for the MEG comparisons. In all scenarios, the input space $\mathcal{X} = [1 : 340]$ was first normalized to take values in $[0 : 1]$, as was the case for the simulated study.

Gaussian Process The GP was specified as follows. The covariance function was taken to be a squared exponential, which for word w and sensor p took the form $c_{w,p} = d_{w,p} \exp(-\kappa||x - x'||_2^2)$. The scale parameter was constrained to be a fixed linear function of the average time-specific sample variance $\hat{\sigma}^2_{w,p}$ of the training data: $d_{w,p} = \alpha^0 \hat{\sigma}^2_{w,p}$. Likewise, the nugget noise was of the form $\sigma^2_{w,p} = \beta \hat{\sigma}^2_{w,p}$. The parameters κ, α^0, and β were optimized on a grid to maximize the marginal likelihood of the training data over all words and sensors.

Hierarchical GP For the 2-level hierarchical GP (hGP), a squared exponential kernel was also assumed for both levels. As in [1], a single bandwidth parameter was assumed. In particular, for the shared top level GP, $c^0_{w,p} = d^0_{w,p} \exp(-\kappa||x - x'||_2^2)$, and for the trial-specific level, $c^1_{w,p} = d^1_{w,p} \exp(-\kappa||x - x'||_2^2)$. As in the GP, the hyperparameters were constrained as a function of $\hat{\sigma}^2_{w,p}$: $d^0_{w,p} = \alpha^0 \hat{\sigma}^2_{w,p}$, $d^1_{w,p} = \alpha^1 \hat{\sigma}^2_{w,p}$, and $\sigma^2_{w,p} = \beta \hat{\sigma}^2_{w,p}$. Here, κ, α^0, α^1, and β were jointly optimized to maximize marginal likelihood. For numerical reasons, the minimum allowable nugget noise was set to 1% of $\hat{\sigma}^2_{w,p}$ (i.e., $\beta = 0.01$).

Multiresolution GP For the multiresolution GP (mGP), a squared exponential kernel with a shared bandwidth κ is used for each GP in the hierarchy, as specified in Sec. 3 of the main paper. The scale parameters $\{d^0, d^1, \ldots, d^{L-1}\}$ were constrained as follows. The global parent GP was assigned scale $d^0 = \alpha^0 \hat{\sigma}^2_{w,p}$. The scale parameters of the $L - 1$ trial-specific levels of the mGP hierarchy were constrained by a fixed functional form as in the simulated data setup, determined by two parameters. In particular, $d^\ell = [\alpha^1 \exp(-\rho * \ell)] \hat{\sigma}^2_{w,p}$. Finally, the nugget noise followed $\sigma^2_{w,p} = \beta \hat{\sigma}^2_{w,p}$. In this scenario, κ, α^0, α^1, ρ, and β were jointly optimized to maximize marginal likelihood based on initial samples of tree partitions $\mathcal{A}^{(m)}$ using the hyperparameter settings of the simulated data example. Again, for numerical reasons, the minimum allowable nugget noise was set to 1% of $\hat{\sigma}^2_{w,p}$.

The optimized hyperparameter values were as follows:

	κ	α^0	α^1	β	ρ
GP	1350	0.15	–	1	–
hGP	13000	0.033	1	0.01	–
mGP	900	0.1	1.67	0.01	1.1

For the hGP, a large bandwidth (low temporal correlation) is taken to account for the abrupt changes. Also note that the parent GP was given little variance and instead the variance was attributed to the lower level GP to account for the significant trial-to-trial variation. The GP accounts for both the large trial-to-trial variability and the abrupt changes through a large nugget noise. The mGP variance

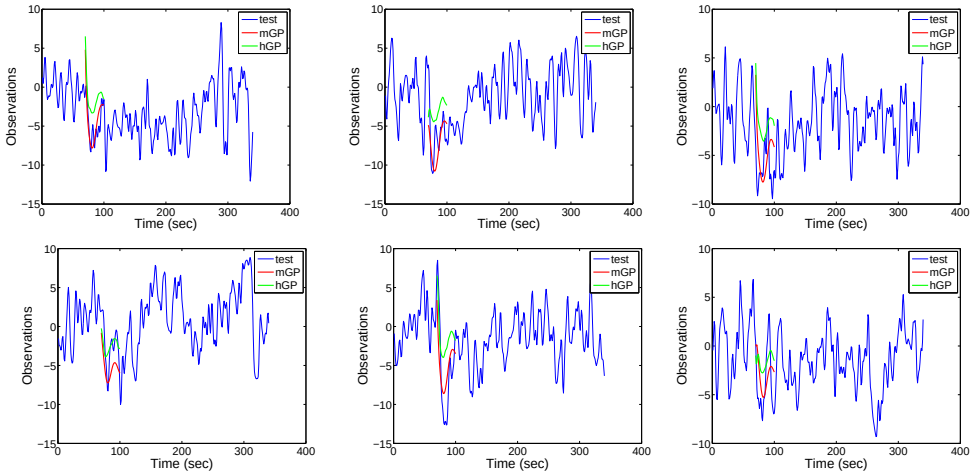

Figure 1: Heldout test data for the visual cortex sensor #77 and 6 different words. For $\tau = 70$, we show the predictive mean $y^*_{\tau:\tau+30}$ under an hGP and mGP conditioned on $y^*_{1:\tau-1}$ and 15 training sequences.

dropped off fairly rapidly with tree level, as indicated by ρ. Note that both the hGP and mGP are able to account for trial-to-trial variability in the tree hierarchy instead of through the nugget noise, as indicated by low values of β.

In our optimization procedure, we found that the performance of the GP was the most sensitive to changes in the hyperparameter specification. Both the hGP and GP were fairly robust over a reasonably large range of settings (partially indicated by a flat marginal likelihood of the training data over the range.)

4.4 Additional Figures

We provide some additional figures related to the MEG results presented in Fig. 6 of the main paper. In Fig. 1, we display examples of the heldout test data for the visual cortex sensor #77 and 6 different words. Each plot only shows the first heldout trial of each word; the results of Fig. 6 in the main paper perform a full analysis on each of the 5 heldout trials. For $\tau = 70$, we show the predictive mean $y^*_{\tau:\tau+30}$ conditioned on $y^*_{1:\tau-1}$ and 15 training sequences. We compare the performance of the mGP to that of the hGP. Only the predictive mean is displayed for clarity. The predictive variances associated with the hGP were similar to, but slightly larger than those of the mGP (7% larger on average). The 95% predictive intervals included the heldout observations in all 6 cases for the mGP and in 5 cases for the hGP. However, note the significantly better mean predictions for the mGP.

References

[1] 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.

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

[3] S. Taulu and R. Hari. Removal of magnetoencephalographic artifacts with temporal signal-space separation: Demonstration with single-trial auditory-evoked responses. *Human brain mapping*, 30(5):1524–34, 2009.

[4] 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.

[5] S. Taulu, M. Kajola, and J. Simola. The Signal Space Separation method. *ArXiv Physics*, 2004. URL http://arxiv.org/abs/physics/0401166.

[6] 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, 1997.

[7] W. H. Wong and L. Ma. Optional Pólya tree and Bayesian inference. *The Annals of Statistics*, 38(3): 1433–1459, 2010.