# Multiresolution Gaussian Processes

Emily B. Fox, David B. Dunson

NeurIPS 2012

Zheng Wang

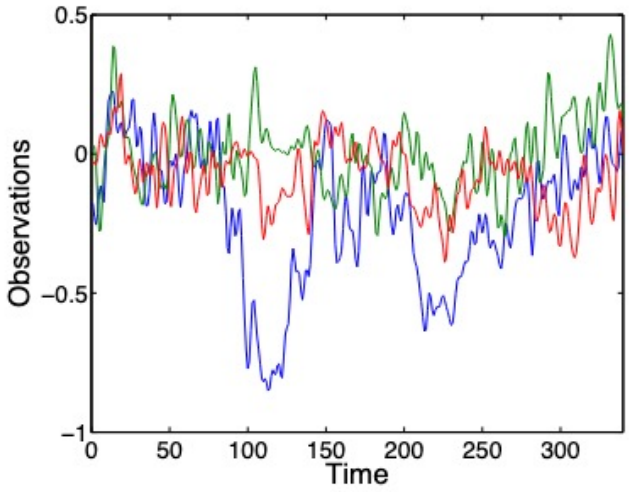
02/18/2022

# Outline

- Motivation
- Method
- Results

# Motivation

- GPs typically assume ***smoothness*** properties of the underlying function being modeled that can blur key elements of the signal if ***abrupt changes occur***
- Motivating application:
  - Magnetoencephalography (MEG) recordings of brain activity in response to some word stimulus



# Contribution

- Direct interpretability
- Local stationarity
- Irregular grids of observations
- Sharing information across related time series

# Model

- Data: $y = \{y_1, \ldots, y_n\},\ y_i \in \Re,$

$\{x_1, \ldots, x_n\},\ x_i \in \mathcal{X} \subset \Re^p$:

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

- Partition over input domain (***multiple resolutions***)

$$\mathcal{A} = \{\mathcal{A}^0, \mathcal{A}^1, \ldots, \mathcal{A}^{L-1}\}$$

$$\mathcal{A}^0 = \mathcal{X} \qquad \mathcal{X} = \bigcup_i \mathcal{A}_i^\ell \text{ and } \mathcal{A}_i^{\ell-1} = \mathcal{A}_{2i-1}^\ell \cup \mathcal{A}_{2i}^\ell$$

- Conditional GPs between resolutions:

$$f^\ell(\mathcal{A}_i^\ell) \sim \mathrm{GP}(f^{\ell-1}(\mathcal{A}_i^\ell), c_i^\ell)$$

$$g = f^{L-1}$$

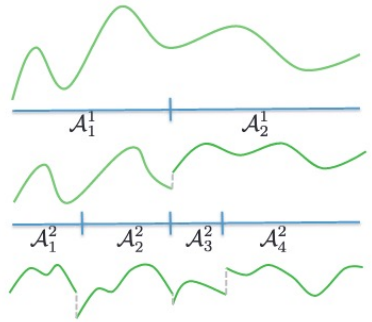



Figure 2: mGP: Parent function is split by $\mathcal{A}^1 = \{\mathcal{A}_1^1, \mathcal{A}_2^1\}$. Recursing down the tree, each partition has a GP with mean given by its parent function restricted to that set.

# Model

- Covariance function:

$$c_i^\ell = d_i^\ell \exp(-\kappa_i^\ell ||x - x'||_2^2),$$

$$d_i^\ell = d^\ell \qquad \sum_{\ell=1}^{\infty} (d^\ell)^2 < 1$$

$$\kappa_i^\ell = \kappa / ||\mathcal{A}_i^\ell||_2^2$$

*Vary less from the parent functions*

- Hyperparameters:

$$\theta = \{d^0, \ldots, d^{L-1}, \kappa\},$$

# Model

- Marginal GP

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

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

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

1. *Nonstationary kernel: distance; location*
2. *Continuous local functions*
3. *Discrete partitions*

# Inference

- Posterior Global Trajectory

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

$$\tilde{\mathbf{y}} = \Sigma^{-1} \textstyle\sum_i \mathbf{y}^{(i)} \qquad \Sigma = \sigma^2 I_n + \sum_{\ell=1}^{L-1} K_\ell$$

- Predictive Distribution

$$p(\mathbf{y}^{(J+1)} \mid \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}, \mathcal{A}) = \int p(\mathbf{y}^{(J+1)} \mid f^0(\mathbf{x}), \mathcal{A}) p(f^0(\mathbf{x}) \mid \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(J)}, \mathcal{A}) df^0$$

$$= N\Big( \left(K_0^{-1} + J\Sigma^{-1}\right)^{-1} \tilde{\mathbf{y}}, \Sigma + \left(K_0^{-1} + J\Sigma^{-1}\right)^{-1} \Big).$$

- Marginal Likelihood

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

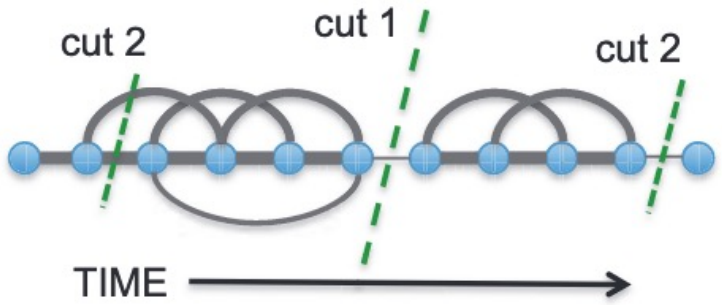
# Hierarchical Partition

- Normalized Cut

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

- Proposal Distribution

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



# Hierarchical Partition

**Algorithm 1** One Iteration of mGP MCMC Sampler - GLOBAL SEARCH

**Input:** Cost matrix $W$, input locations $\mathcal{X}$, hyperparameters $\theta$,
previous partition $\mathcal{A}$ and corresponding $\Sigma$

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

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

**for** $\ell = 1, \dots, L-1$ **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 $K_\ell$ 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 $K_\ell$ submatrix corresponding to $\mathcal{A}'^\ell_{\nu+1}$

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

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

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

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

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

# Hierarchical Partition

**Algorithm 2** One Iteration of mGP MCMC Sampler - LOCAL SEARCH

**Input:** Cost matrix $W$, input locations $\mathcal{X}$, hyperparameters $\theta$,
previous partition $\mathcal{A}$ and corresponding $\Sigma$

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

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

$\mathcal{A}^{\ell *}_{\nu *} \sim$ `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 $\nu$ 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 $K_\ell$ 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 $K_\ell$ submatrix corresponding to $\mathcal{A}'^{\ell}_{\nu+1}$

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

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

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

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

# Experiments

- Setting
  - Synthesized Data
    - 100 replicates of length 200 generated by a 5-level mGP
  - MEG Data
    - 10 words and 102 sensors
    - 15 training trials, 5 testing trials, length 340
- Competing Methods
  - Hierarchical GP
  - Wavelet-based Functional Mixed Model

# Hierarchical GP

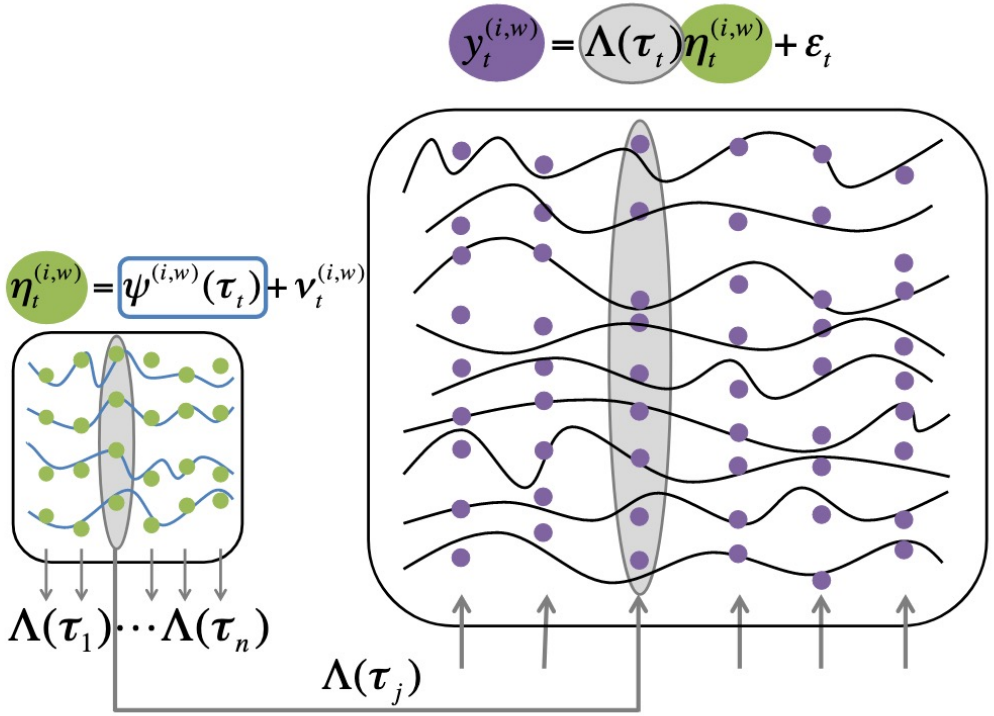



**Figure 3:** The latent model for one trial $i$ of word $w$. Independent noise is added to the child latent process $\psi^{(i,w)}$ (see Figure 2) to produce $\eta_t^{(i,w)}$ (green circles). Then, $\eta_t^{(i,w)}$ is projected to the full dimensional space via the time-varying $\Lambda(\tau)$ (grey).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

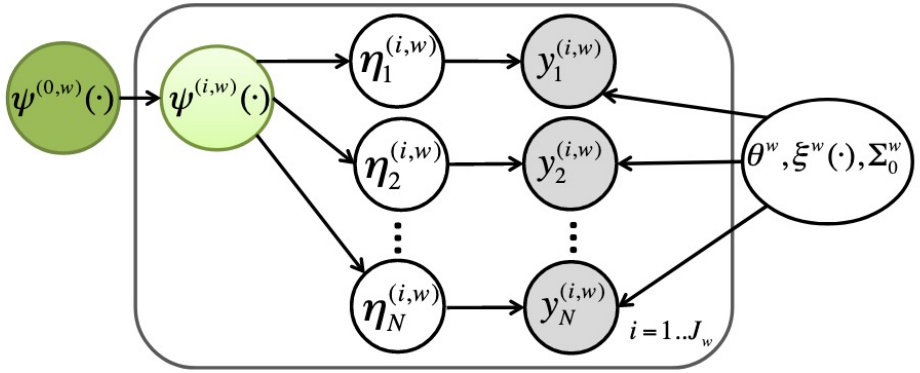



**Figure 1:** A graphical representation of the model outlined in Section 4 for one word $w$, and its trials $i = 1 \ldots J_w$. The mean of the child latent process $\psi^{(i,w)}$ (light green) is given by the parent latent process $\psi^{(0,w)}$ (dark green). The latent factors $\eta_t^{(i,w)}$ are centered about $\psi^{(i,w)}$. The marginal mean of $y^{(i,w)}$ is governed by $\Theta^w, \xi^w$ and $\psi^{(i,w)}$ while the covariance of $y^{(i,w)}$ is governed by $\Theta^w, \xi^w$ and $\Sigma_0^w$ as in Equation 10.

# Wavelet-based Functional Mixed Model

- Wavelet Transform

The Hilbert basis is constructed as the family of functions $\{\psi_{jk} : j, k \in \mathbb{Z}\}$ by means of dyadic translations and dilations of $\psi$,

$$\psi_{jk}(x) = 2^{\frac{j}{2}} \psi\left(2^j x - k\right)$$

for integers $j, k \in \mathbb{Z}$.

The **integral wavelet transform** is the integral transform defined as

$$[W_\psi f](a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \overline{\psi\left(\frac{x-b}{a}\right)} f(x) dx$$

The **wavelet coefficients** $c_{jk}$ are then given by

$$c_{jk} = [W_\psi f]\left(2^{-j}, k2^{-j}\right)$$

Here, $a = 2^{-j}$ is called the **binary dilation** or **dyadic dilation**, and $b = k2^{-j}$ is the **binary** or **dyadic position**.

# Wavelet-based Functional Mixed Model

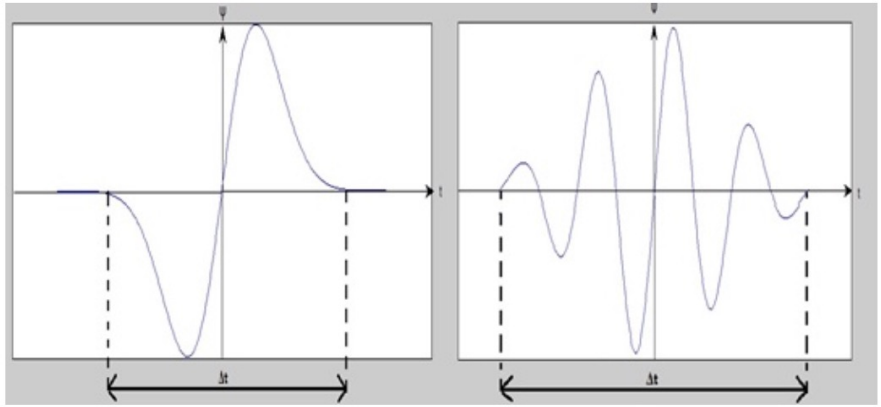
- Wavelet Transform

When $\Delta t$ is large,

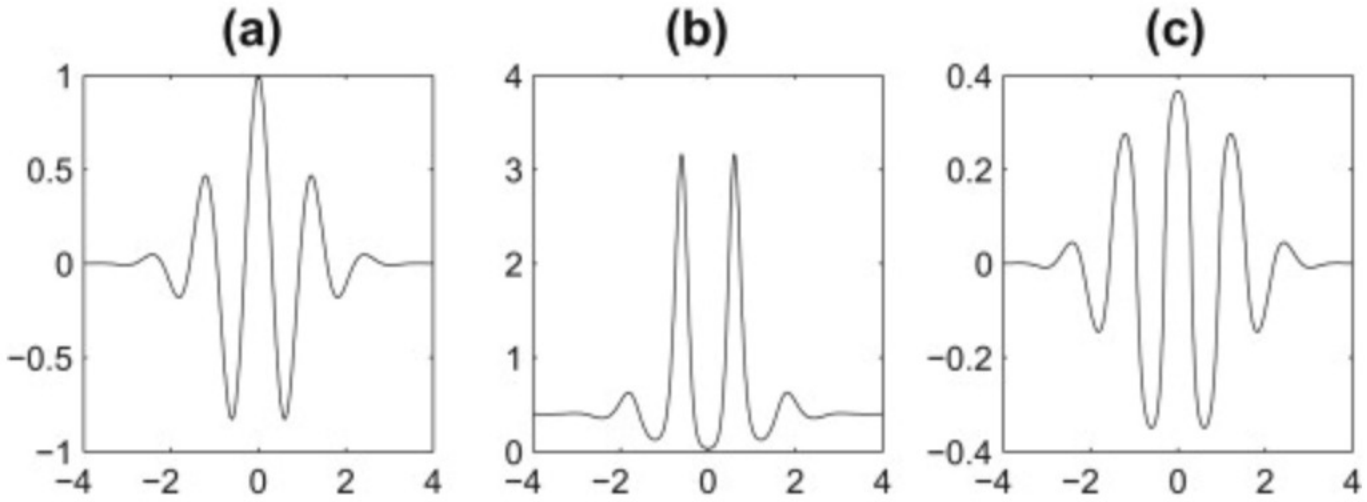
1. Bad time resolution
2. Good frequency resolution
3. Low frequency, large scaling factor

When $\Delta t$ is small

1. Good time resolution
2. Bad frequency resolution
3. High frequency, small scaling factor



| Transform | Representation | Input |
|---|---|---|
| Fourier transform | $\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi}\,dx$ | $\xi$ frequency |
| Time-frequency analysis | $X(t, f)$ | $t$ time; $f$ frequency |
| Wavelet transform | $X(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \overline{\Psi\left(\frac{t-b}{a}\right)} x(t)\,dt$ | $a$ scaling; $b$ time shift factor |



# Wavelet-based Functional Mixed Model

- Discrete Wavelet Transform

Applied the following discretization of frequency and time:

$$c_n = c_0^n$$
$$\tau_m = m \cdot T \cdot c_0^n$$

Leading to wavelets of the form, the discrete formula for the basis wavelet:

$$\Psi(k,n,m) = \frac{1}{\sqrt{c_0^n}} \cdot \Psi\left[\frac{k - mc_0^n}{c_0^n} T\right] = \frac{1}{\sqrt{c_0^n}} \cdot \Psi\left[\left(\frac{k}{c_0^n} - m\right) T\right]$$

Such discrete wavelets can be used for the transformation:

$$Y_{DW}(n,m) = \frac{1}{\sqrt{c_0^n}} \cdot \sum_{k=0}^{K-1} y(k) \cdot \Psi\left[\left(\frac{k}{c_0^n} - m\right) T\right]$$

# Wavelet-Based FMM:
## General Approach

1. **Project** observed functions Y **into wavelet space.**
2. **Fit FMM** in wavelet space
   (Use MCMC to get posterior samples)
3. **Project** wavelet-space estimates (posterior samples) **back to data space.**

# Projecting FMM to Wavelet Space

$$\underset{N\times \mathrm{T}}{Y}\overbrace{\mathbf{W}'}^{\mathrm{T}\times\mathrm{T}} = \overbrace{X}^{N\times p}\underset{p\times \mathrm{T}}{B}\overbrace{\mathbf{W}'}^{\mathrm{T}\times\mathrm{T}} + \overbrace{Z}^{N\times m}\underset{m\times \mathrm{T}}{U}\overbrace{\mathbf{W}'}^{\mathrm{T}\times\mathrm{T}} + \underset{N\times \mathrm{T}}{E}\overbrace{\mathbf{W}'}^{\mathrm{T}\times\mathrm{T}}$$

$$U\mathbf{W}' \sim MN(P, \mathbf{W}Q\mathbf{W}')$$

$$E\mathbf{W}' \sim MN(R, \mathbf{W}S\mathbf{W}')$$

# Experiments

- Synthetic Data

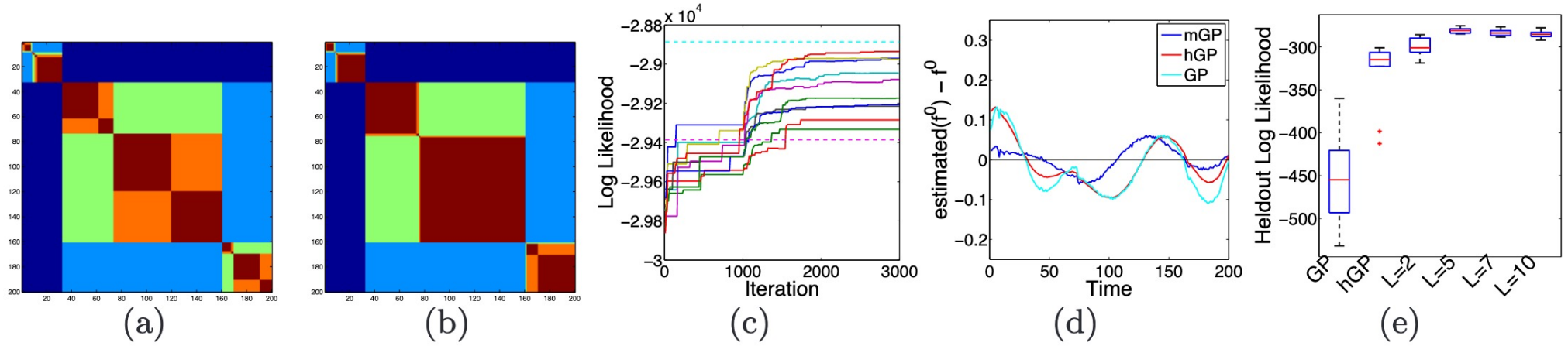



Figure 6: For the data of Fig. 4, (a) true and (b) MAP partitions. (c) Trace plots of log likelihood versus MCMC iteration for 10 chains. Log likelihood under the true partition (*cyan*) and minimized normalized cut partition of Fig. 4 (*magenta*) are also shown. (d) Errors between posterior mean $f^0$ and true $f^0$ for GP, hGP, and mGP. (e) Predictive log likelihood of 10 heldout sequences for GP, hGP, and mGP with $L = 2, 5(true), 7, 10$.

# Experiments

- MEG Data

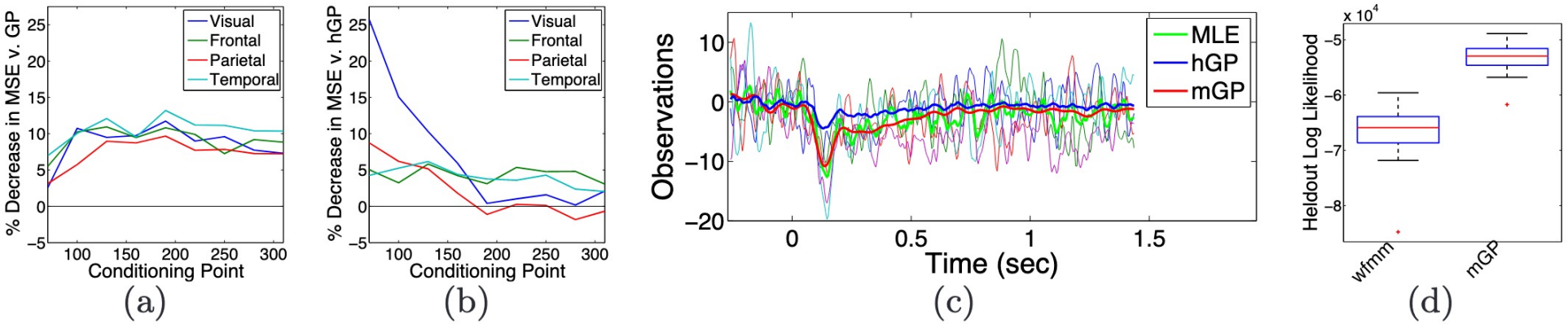



Figure 7: Per-lobe comparison of mGP to (a) GP and (b) hGP: For various values of $\tau$, % decrease in predictive MSE of heldout $y^*_{\tau:\tau+30}$ conditioned on $y^*_{1:\tau-1}$ and 15 training sequences. (c) For a visual cortex sensor and word *hammer*, plots of test data, empirical mean (MLE), and hGP and mGP predictive mean for entire heldout $\mathbf{y}^*$. (d) Boxplots of predictive log likelihood of $\mathbf{y}^*$ for the mGP and wavelet-based method of [18]. Plots aggregate results over 5 heldout sequences $\mathbf{y}^*$ per word.

# Experiments

- MEG Data

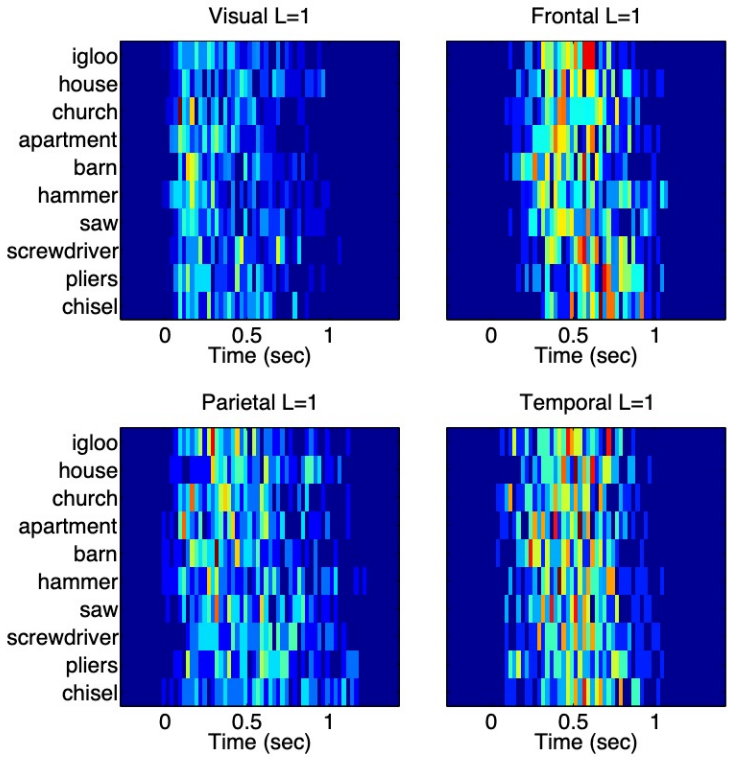



Figure 8: Inferred changepoints at level 1 aggregated over sensors within each lobe: visual (*top-left*), frontal (*top-right*), parietal (*bottom-left*), and temporal (*bottom-right*).

# Multivariate Sparse Coding of Nonstationary Covariances with Gaussian Processes

$$p(z_{ik} = 1|\theta_k, Q) = \frac{\exp(\theta_k^T Q(\cdot, y_i))}{\sum_{k'=1}^{K} \exp(\theta_{k'}^T Q(\cdot, y_i))}$$

$$p(\mathbf{g}|V) = N(\mathbf{g}|0, \Sigma_g) \qquad p(\mathbf{f}|\mathbf{g}, Z) = N(\mathbf{f}|Z\mathbf{g}, \Sigma_f)$$

$$p(\mathbf{f}|Z) = N(\mathbf{f}|0, Z\Sigma_g Z^T + \Sigma_f)$$

# Some Thoughts

- Nonparameteric time domain partition
  - Automatically generate the clusters: Dirichlet process

Thanks!