Efficient Hierarchical Bayesian Inference for Spatio-temporal Regression Models in Neuroimaging

Ali Hashemi
Technische Universität Berlin

joint work with Yijing Gao, Chang Cai, Sanjay Ghosh,
Klaus-Robert Müller, Srikantan S. Nagarajan, and Stefan Haufe

35th Conference on Neural Information Processing Systems (NeurIPS 2021)

Multi-task Linear Regression

$\mathbf{Y}_g = \mathbf{L}\mathbf{X}_g + \mathbf{E}_g$ **Spatio-temporal generative model**

$\mathbf{Y}_g \in \mathbb{R}^{M \times T}$ for $g = 1, \ldots, G$, G:#sample blocks or tasks

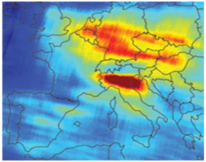
$\mathbf{X}_g \in \mathbb{R}^{N \times T}$ M:#measurements or observations, T:#Samples,

$\mathbf{E}_g \in \mathbb{R}^{N \times T}$ N:#coefficients or source components,

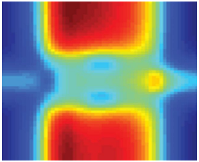
$\mathbf{L} \in \mathbb{R}^{M \times N}$ forward matrix (known): maps $\mathbf{X}_g$ to $\mathbf{Y}_g$

Goal: Estimate $\{\mathbf{X}_g\}_{g=1}^{G}$ given $\mathbf{L}$ and $\{\mathbf{Y}_g\}_{g=1}^{G}$:

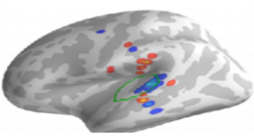
- Inverse problem in physics
- Multiple measurement vector (MMV) recovery problem in signal processing



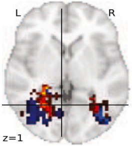
Temperature monitoring of climate [S. Beirle et al. 2003]



Temperature monitoring of CPU/GPU [J. Ranieri et al. 2012]



EEG/MEG Source Localization [H. Janati et al. 2020]

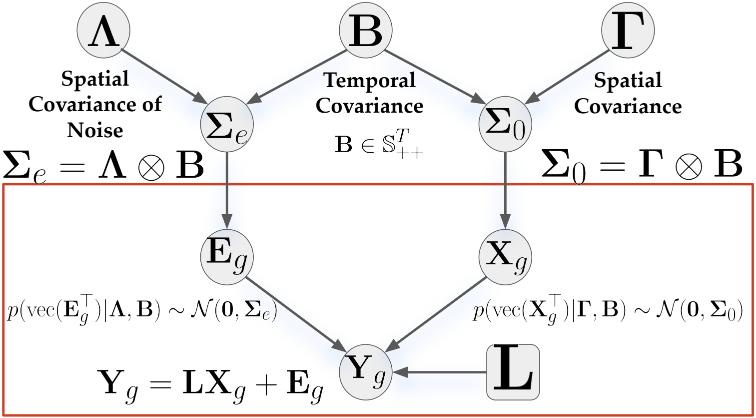


fMRI data analysis [M. B. Cai, et al. 2020]

Hierarchical Bayesian Learning

Spatio-temporal dynamics of model parameters and noise are modeled to have Kronecker product covariance structure.

Probabilistic graphical model:



Hierarchical Bayesian Inference and Type-II Loss

Posterior source distribution: $p(\mathrm{vec}(\mathbf{X}_g^\top)|\mathrm{vec}(\mathbf{Y}_g^\top), \mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) \sim \mathcal{N}(\bar{\mathbf{x}}_g, \mathbf{\Sigma}_\mathbf{x})$
with

$$\begin{aligned}
\bar{\mathbf{x}}_g &= \mathrm{vec}(\bar{\mathbf{X}}_g^\top) = \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{y}_g \\
\mathbf{\Sigma}_\mathbf{x} &= \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{D} \mathbf{\Sigma}_0 \\
\tilde{\mathbf{\Sigma}}_\mathbf{y} &= \mathbf{\Sigma}_\mathbf{y} \otimes \mathbf{B} \\
\mathbf{\Sigma}_\mathbf{y} &= \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top + \mathbf{\Lambda} \ ,
\end{aligned}$$

where $\mathbf{D} = \mathbf{L} \otimes \mathbf{I}_T$.

$\mathbf{\Gamma}$, $\mathbf{\Lambda}$, $\mathbf{B}$ are learned by minimizing the negative log marginal likelihood (Type-II) loss, $-\log p(\mathbf{Y}|\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$.

$$\text{Type} - \text{II Loss} : \mathcal{L}_{\mathrm{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T \log |\mathbf{\Sigma}_\mathbf{y}| + M \log |\mathbf{B}| + \frac{1}{G} \sum_{g=1}^{G} \mathrm{tr}(\mathbf{\Sigma}_\mathbf{y}^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^\top)$$

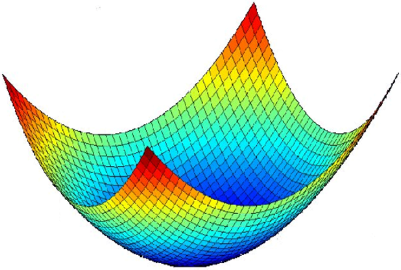
Challenges

$$\text{Type} - \text{II Loss}: \mathcal{L}_{\text{kron}}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}) = T \log |\boldsymbol{\Sigma}_{\mathbf{y}}| + M \log |\mathbf{B}| + \frac{1}{G} \sum_{g=1}^{G} \text{tr}(\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^{\top})$$

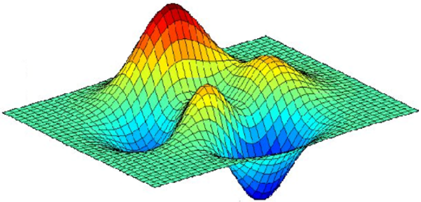
Challenges

$$\text{Type} - \text{II Loss}: \mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T \log |\mathbf{\Sigma}_{\mathbf{y}}| + M \log |\mathbf{B}| + \frac{1}{G} \sum_{g=1}^{G} \text{tr}(\mathbf{\Sigma}_{\mathbf{y}}^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^{\top})$$

1. Non-convex Type-II ML loss function: non-trivial to solve.



convex function



non-convex function

Challenges

$$\text{Type} - \text{II Loss} : \mathcal{L}_{\mathsf{kron}}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}) = T \log |\boldsymbol{\Sigma}_{\mathbf{y}}| + M \log |\mathbf{B}| + \frac{1}{G} \sum_{g=1}^{G} \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^{\top})$$

1. Non-convex Type-II ML loss function: non-trivial to solve.
2. Most contributions in the literature *neglect the temporal structure* and are based on MAP (Type-I) estimation.
3. A few works that model the temporal dynamics often involve a *computationally demanding inference* scheme mostly based on expectation-maximization (EM).

Our Contributions

- Derive novel Type-II algorithms that automatically learn the temporal structure
  1. Exploit the intrinsic Riemannian geometry of temporal autocovariance matrices.
  2. For stationary dynamics described by Toeplitz matrices, we employ the theory of circulant embeddings.
- Devise an efficient inference based on majorization-minimization optimization with guaranteed convergence properties.

To this end, we present a series of theorems resulting in a novel and efficient hierarchical Bayesian inference for spatio-temporal multi-task regression models.

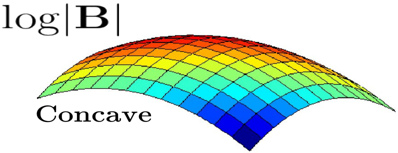
Convex Majorizing Functions

Theorem (Majorizing function for temporal covariance update)

Optimizing $\mathcal{L}_{kron}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ *with respect to* $\mathbf{B}$ *is equivalent to optimizing the following convex surrogate function, which majorizes* $\mathcal{L}_{kron}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$*:*

$$\mathcal{L}_{\mathrm{conv}}^{\mathrm{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B}) = \mathrm{tr}\left((\mathbf{B}^k)^{-1}\mathbf{B}\right) + \mathrm{tr}(\mathbf{M}_{\mathrm{time}}^k \mathbf{B}^{-1}),$$

where $\mathbf{M}_{\mathrm{time}}^k := \frac{1}{MG}\sum_{g=1}^{G} \mathbf{Y}_g^\top \left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{Y}_g$.



$$\boxed{\mathcal{L}_{\mathrm{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T\log|\mathbf{\Sigma}_{\mathbf{y}}| + M\log|\mathbf{B}| + \frac{1}{G}\sum_{g=1}^{G}\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{Y}_g\mathbf{B}^{-1}\mathbf{Y}_g^\top)}$$

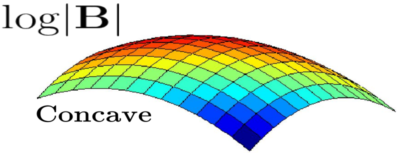
Convex Majorizing Functions

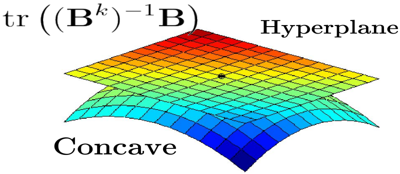
Theorem (Majorizing function for temporal covariance update)

Optimizing $\mathcal{L}_{kron}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ *with respect to* $\mathbf{B}$ *is equivalent to optimizing the following convex surrogate function, which majorizes* $\mathcal{L}_{kron}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$*:*

$$\mathcal{L}_{\mathrm{conv}}^{\mathrm{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B}) = \mathrm{tr}\left((\mathbf{B}^k)^{-1}\mathbf{B}\right) + \mathrm{tr}(\mathbf{M}_{\mathrm{time}}^k \mathbf{B}^{-1}),$$

where $\mathbf{M}_{\mathrm{time}}^k := \frac{1}{MG}\sum_{g=1}^{G} \mathbf{Y}_g^\top \left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{Y}_g$.





$$\mathcal{L}_{\mathrm{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T\log|\mathbf{\Sigma}_{\mathbf{y}}| + M\log|\mathbf{B}| + \frac{1}{G}\sum_{g=1}^{G}\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{Y}_g\mathbf{B}^{-1}\mathbf{Y}_g^\top)$$

Convex Majorizing Functions

Theorem (Majorizing function for temporal covariance update)

Optimizing $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$ *with respect to* $\mathbf{B}$ *is equivalent to optimizing the following convex surrogate function, which majorizes* $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$*:*

$$\mathcal{L}_{\mathrm{conv}}^{\mathrm{time}}(\boldsymbol{\Gamma}^k, \boldsymbol{\Lambda}^k, \mathbf{B}) = \mathrm{tr}\left((\mathbf{B}^k)^{-1}\mathbf{B}\right) + \mathrm{tr}(\mathbf{M}_{\mathrm{time}}^k \mathbf{B}^{-1}),$$

where $\mathbf{M}_{\mathrm{time}}^k := \frac{1}{MG} \sum_{g=1}^{G} \mathbf{Y}_g^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{Y}_g$.

Theorem (Majorizing function for spatial covariance update)

Let $\mathbf{H} = \mathrm{diag}(\mathbf{h})$*,* $\mathbf{h} = [\gamma_1, \ldots, \gamma_N, \sigma_1^2, \ldots, \sigma_M^2]^\top$*,* $\boldsymbol{\Phi} := [\mathbf{L}, \mathbf{I}]$*, and* $\boldsymbol{\Sigma}_{\mathbf{y}} = \boldsymbol{\Phi}\mathbf{H}\boldsymbol{\Phi}^\top$*. Then, optimizing* $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$ *with respect to* $\mathbf{H}$ *is equivalent to minimizing the following convex surrogate function, which majorizes* $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$*:*

$$\mathcal{L}_{\mathrm{conv}}^{\mathrm{space}}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}^k) = \mathcal{L}_{\mathrm{conv}}^{\mathrm{space}}(\mathbf{H}, \mathbf{B}^k) = \mathrm{tr}\left(\boldsymbol{\Phi}^\top (\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1} \boldsymbol{\Phi}\mathbf{H}\right) + \mathrm{tr}\left(\mathbf{M}_{\mathrm{SN}}^k \mathbf{H}^{-1}\right) \ ,$$

where $\mathbf{M}_{\mathrm{SN}}^k := \mathbf{H}^k \boldsymbol{\Phi}^\top (\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1} \mathbf{M}_{\mathrm{space}}^k (\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1} \boldsymbol{\Phi}\mathbf{H}^k$*,*
$\mathbf{M}_{\mathrm{space}}^k := \frac{1}{TG} \sum_{g=1}^{G} \mathbf{Y}_g (\mathbf{B}^k)^{-1} \mathbf{Y}_g^\top$.

Convex Majorizing Functions

Theorem (Majorizing function for temporal covariance update)

Optimizing $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$ with respect to $\mathbf{B}$ is equivalent to optimizing the following convex surrogate function, which majorizes $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$:

$$\mathcal{L}_{\text{conv}}^{\text{time}}(\boldsymbol{\Gamma}^k, \boldsymbol{\Lambda}^k, \mathbf{B}) = \operatorname{tr}\left((\mathbf{B}^k)^{-1}\mathbf{B}\right) + \operatorname{tr}(\mathbf{M}_{\text{time}}^k \mathbf{B}^{-1}),$$

where $\mathbf{M}_{\text{time}}^k := \frac{1}{MG}\sum_{g=1}^{G} \mathbf{Y}_g^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{Y}_g$.

Theorem (Majorizing function for spatial covariance update)

Let $\mathbf{H} = \operatorname{diag}(\mathbf{h})$, $\mathbf{h} = [\gamma_1, \dots, \gamma_N, \sigma_1^2, \dots, \sigma_M^2]^\top$, $\boldsymbol{\Phi} := [\mathbf{L}, \mathbf{I}]$, *and* $\boldsymbol{\Sigma}_{\mathbf{y}} = \boldsymbol{\Phi}\mathbf{H}\boldsymbol{\Phi}^\top$. *Then, optimizing $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$ with respect to $\mathbf{H}$ is equivalent to minimizing the following convex surrogate function, which majorizes $\mathcal{L}_{kron}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$:*

$$\mathcal{L}_{\text{conv}}^{\text{space}}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}^k) = \mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k) = \operatorname{tr}\left(\boldsymbol{\Phi}^\top(\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1}\boldsymbol{\Phi}\mathbf{H}\right) + \operatorname{tr}\left(\mathbf{M}_{\text{SN}}^k \mathbf{H}^{-1}\right) ,$$

where $\mathbf{M}_{\text{SN}}^k := \mathbf{H}^k\boldsymbol{\Phi}^\top(\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1}\mathbf{M}_{\text{space}}^k(\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1}\boldsymbol{\Phi}\mathbf{H}^k$,
$\mathbf{M}_{\text{space}}^k := \frac{1}{TG}\sum_{g=1}^{G} \mathbf{Y}_g(\mathbf{B}^k)^{-1}\mathbf{Y}_g^\top$.

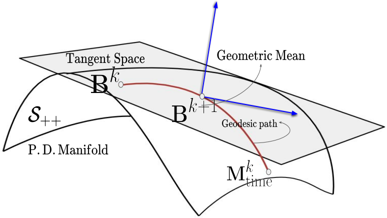
Riemannian Update on the Manifold of P.D. Matrices

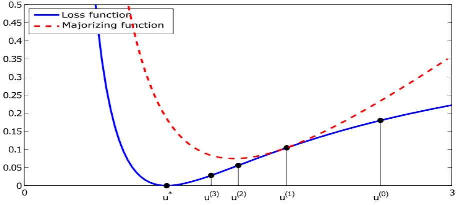
Theorem (Geometric mean)

The cost function $\mathcal{L}_{\text{conv}}^{\text{time}}(\boldsymbol{\Gamma}^k, \boldsymbol{\Lambda}^k, \mathbf{B})$ *is strictly geodesically convex with respect to the P.D. manifold and its minimum with respect to* $\mathbf{B}$ *can be attained by iterating the following update rule until convergence:*

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{1/2} \left((\mathbf{B}^k)^{-1/2} \mathbf{M}_{\text{time}}^k (\mathbf{B}^k)^{-1/2} \right)^{1/2} (\mathbf{B}^k)^{1/2} ,$$

which leads to a majorization-minimization (MM) algorithm with convergence guarantees $\rightsquigarrow$ **Full Dugh**





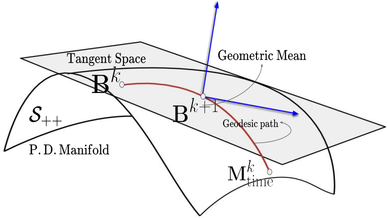
Riemannian Update on the Manifold of P.D. Matrices

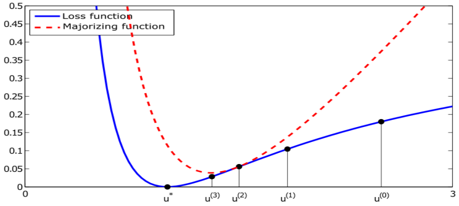
Theorem (Geometric mean)

The cost function $\mathcal{L}_{\mathrm{conv}}^{\mathrm{time}}(\boldsymbol{\Gamma}^k, \boldsymbol{\Lambda}^k, \mathbf{B})$ *is strictly geodesically convex with respect to the P.D. manifold and its minimum with respect to* $\mathbf{B}$ *can be attained by iterating the following update rule until convergence:*

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{1/2} \left((\mathbf{B}^k)^{-1/2} \mathbf{M}_{\mathrm{time}}^k (\mathbf{B}^k)^{-1/2} \right)^{1/2} (\mathbf{B}^k)^{1/2} \, ,$$

which leads to a majorization-minimization (MM) algorithm with convergence guarantees $\rightsquigarrow$ **Full Dugh**



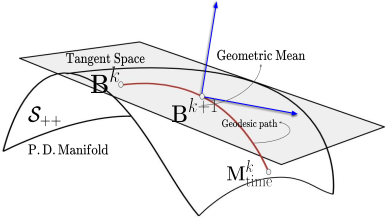


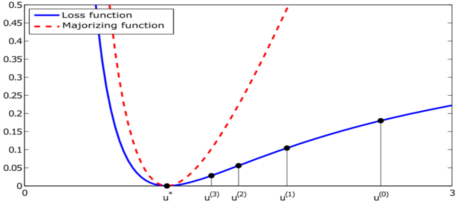
Theorem (Geometric mean)

The cost function $\mathcal{L}_{\text{conv}}^{\text{time}}(\boldsymbol{\Gamma}^k, \boldsymbol{\Lambda}^k, \mathbf{B})$ *is strictly geodesically convex with respect to the P.D. manifold and its minimum with respect to* $\mathbf{B}$ *can be attained by iterating the following update rule until convergence:*

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{1/2} \left((\mathbf{B}^k)^{-1/2} \mathbf{M}_{\text{time}}^k (\mathbf{B}^k)^{-1/2} \right)^{1/2} (\mathbf{B}^k)^{1/2} ,$$

which leads to a majorization-minimization (MM) algorithm with convergence guarantees $\rightsquigarrow$ **Full Dugh**





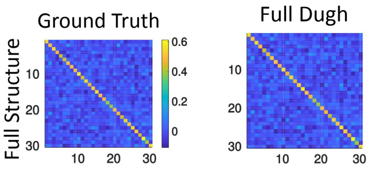
Riemannian Update on the Manifold of P.D. Matrices

Theorem (Geometric mean)

The cost function $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$ *is strictly geodesically convex with respect to the P.D. manifold and its minimum with respect to* $\mathbf{B}$ *can be attained by iterating the following update rule until convergence:*

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{1/2} \left((\mathbf{B}^k)^{-1/2} \mathbf{M}_{\text{time}}^k (\mathbf{B}^k)^{-1/2} \right)^{1/2} (\mathbf{B}^k)^{1/2} \, ,$$

which leads to a majorization-minimization (MM) algorithm with convergence guarantees $\rightsquigarrow$ **Full Dugh**



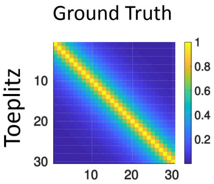
Riemannian Update on the Manifold of P.D. Matrices

Theorem (Geometric mean)

The cost function $\mathcal{L}^{\text{time}}_{\text{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$ *is strictly geodesically convex with respect to the P.D. manifold and its minimum with respect to* $\mathbf{B}$ *can be attained by iterating the following update rule until convergence:*

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{1/2} \left((\mathbf{B}^k)^{-1/2}\mathbf{M}^k_{\text{time}}(\mathbf{B}^k)^{-1/2}\right)^{1/2} (\mathbf{B}^k)^{1/2} ,$$

which leads to a majorization-minimization (MM) algorithm with convergence guarantees $\rightsquigarrow$ **Full Dugh**



Riemannian Update for Toeplitz Matrices

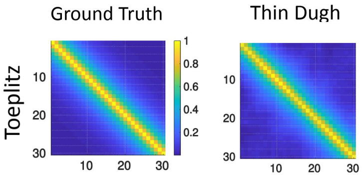
Theorem (Temporal covariance update using circulant embedding)

Let $\mathcal{L}_{\mathrm{conv}}^{\mathrm{time}}(\boldsymbol{\Gamma}^k, \boldsymbol{\Lambda}^k, \mathbf{B})$ *is constrained to the set of real-valued positive-definite Toeplitz matrices,* $\mathbf{B} \in \mathcal{B}^L : \mathbf{B} = \mathbf{Q}\mathbf{P}\mathbf{Q}^H$, *where* $\mathbf{P} = \mathrm{diag}(\mathbf{p}) \in \mathbb{R}^{L \times L}$ *with* $L > T$ *be the circulant embedding of* $\mathbf{B}$*. Then the resulting constrained loss function is convex in* $\mathbf{p}$*, and its minimum with respect to* $\mathbf{p}$ *can be obtained by iterating the following closed-form update rule until convergence:*

$$p_l^{k+1} \leftarrow \sqrt{\frac{\hat{g}_l^k}{\hat{z}_l^k}} \text{ for } l = 1, \dots, L \text{ , where}$$

$$\hat{\mathbf{g}} := \mathrm{diag}(\mathbf{P}^k \mathbf{Q}^H (\mathbf{B}^k)^{-1} \mathbf{M}_{\mathrm{time}}^k (\mathbf{B}^k)^{-1} \mathbf{Q}\mathbf{P}^k)$$

$$\hat{\mathbf{z}} := \mathrm{diag}(\mathbf{Q}^H (\mathbf{B}^k)^{-1} \mathbf{Q})$$



Riemannian Update for Toeplitz Matrices

Theorem (Temporal covariance update using circulant embedding)

$$p_l^{k+1} \leftarrow \sqrt{\frac{\hat{g}_l^k}{\hat{z}_l^k}} \text{ for } l = 1, \ldots, L \text{ , where}$$

$$\hat{\mathbf{g}} := \mathrm{diag}(\mathbf{P}^k \mathbf{Q}^H (\mathbf{B}^k)^{-1} \mathbf{M}_{\mathrm{time}}^k (\mathbf{B}^k)^{-1} \mathbf{Q} \mathbf{P}^k)$$

$$\hat{\mathbf{z}} := \mathrm{diag}(\mathbf{Q}^H (\mathbf{B}^k)^{-1} \mathbf{Q})$$

Theorem (Spatial covariance with diagonal structure)

The cost function $\mathcal{L}_{\mathrm{conv}}^{\mathrm{space}}(\mathbf{H}, \mathbf{B}^k)$ *is convex in* $\mathbf{h}$*, and its minimum with respect to* $\mathbf{h}$ *can be obtained according to the following closed-form update rule:*

$$h_i^{k+1} \leftarrow \sqrt{\frac{g_i^k}{z_i^k}} \quad \text{for } i = 1, \ldots, N + M \text{ , where}$$

$$\mathbf{g} := \mathrm{diag}(\mathbf{M}_{\mathrm{SN}}^k)$$

$$\mathbf{z} := \mathrm{diag}(\mathbf{\Phi}^\top (\mathbf{\Sigma}_{\mathbf{y}}^k)^{-1} \mathbf{\Phi})$$

Full and Thin Dugh

Combining this theoretical work, we developed a novel algorithm called “Dugh” for joint estimation of **spatial and temporal** covariances of **source and noise**.

Algorithm 1: Full Dugh

Input : The lead field matrix $\mathbf{L} \in \mathbb{R}^{M \times N}$ and G trials of measurement vectors $\{\mathbf{Y}_g\}_{g=1}^{G}$, where $\mathbf{Y}_g \in \mathbb{R}^{M \times T}$.

Result: Estimates of the source and noise variances $\mathbf{h} = [\gamma_1, \dots, \gamma_N, \sigma_1^2, \dots, \sigma_M^2]^\top$, the temporal covariance $\mathbf{B}$, and the posterior mean $\{\bar{\mathbf{x}}_g\}_{g=1}^{G}$ and covariance $\mathbf{\Sigma_x}$ of the sources.

Choose a random initial value for $\mathbf{B}$ as well as $\mathbf{h} = [\gamma_1, \dots, \gamma_N, \sigma_1^2, \dots, \sigma_M^2]^\top$, and construct $\mathbf{H} = \mathrm{diag}(\mathbf{h})$ and $\mathbf{\Gamma} = \mathrm{diag}([\gamma_1, \dots, \gamma_N]^\top)$.
Construct the augmented lead field $\mathbf{\Phi} = [\mathbf{L}, \mathbf{I}_M]$.
Calculate the lead field $\mathbf{D} = \mathbf{L} \otimes \mathbf{I}_T$ for vectorized sources.
Calculate the prior spatio-temporal covariance for the sources as $\mathbf{\Sigma}_0 = \mathbf{\Gamma} \otimes \mathbf{B}$.
Calculate the spatial statistical covariance $\mathbf{\Sigma_y} = \mathbf{\Phi H \Phi}^\top$.
Calculate the spatio-temporal statistical covariance $\tilde{\mathbf{\Sigma}}_\mathbf{y} = \mathbf{\Sigma_y} \otimes \mathbf{B}$.
Initialize $k \leftarrow 1$
repeat
  Calculate the posterior mean as $\bar{\mathbf{x}}_g = \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{y}_g$, for $g = 1, \dots, G$, where $\mathbf{y}_g = \mathrm{vec}\left(\mathbf{Y}_g^\top\right) \in \mathbb{R}^{MT \times 1}$.
  Calculate $\mathbf{M}_{\mathrm{time}}^k$, and update $\mathbf{B}$ based on Riemannian update on the manifold of P.D. matrices.
  Calculate $\mathbf{M}_{\mathrm{SN}}^k$, and update $\mathbf{H}$.
  $k \leftarrow k + 1$
until *stopping condition is satisfied:* $\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|_2^2 \leq \epsilon$ *or* $k = k_{max}$;
Calculate the posterior covariance as $\mathbf{\Sigma_x} = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{D} \mathbf{\Sigma}_0$.

Algorithm 2: Thin Dugh

Input : The lead field matrix $\mathbf{L} \in \mathbb{R}^{M \times N}$, and G trials of measurement vectors $\{\mathbf{Y}_g\}_{g=1}^{G}$, where $\mathbf{Y}_g \in \mathbb{R}^{M \times T}$.

Result: Estimates of the source and noise variances $\mathbf{h} = [\gamma_1, \dots, \gamma_N, \sigma_1^2, \dots, \sigma_M^2]^\top$, the temporal covariance $\mathbf{B}$, and the posterior mean $\{\bar{\mathbf{x}}_g\}_{g=1}^{G}$.

Choose a random initial value for $\mathbf{p}$ as well as $\mathbf{h}$, and construct $\mathbf{H} = \mathrm{diag}(\mathbf{h})$ and $\mathbf{P} = \mathrm{diag}(\mathbf{p})$.
Construct $\mathbf{B} = \mathbf{QPQ}^H$, where $\mathbf{Q} = [\mathbf{I}_M, \mathbf{0}]\mathbf{F}_L$ with $L = 2T + 1$ and $\mathbf{F}_L$ as DFT.
Construct the augmented lead field $\mathbf{\Phi} := [\mathbf{L}, \mathbf{I}_M]$.
Calculate the prior spatio-temporal covariance for the sources as $\mathbf{\Sigma}_0 = \mathbf{\Gamma} \otimes \mathbf{B}$.
Calculate the statistical covariance $\mathbf{\Sigma_y} = \mathbf{\Phi H \Phi}^\top$.
Calculate the spatio-temporal statistical covariance $\tilde{\mathbf{\Sigma}}_\mathbf{y} = \mathbf{\Sigma_y} \otimes \mathbf{B}$.
Initialize $k \leftarrow 1$
repeat
  Calculate the posterior mean efficiently as $\bar{\mathbf{x}}_g = \mathrm{tr}\left(\mathbf{QP}\left(\mathbf{\Pi} \odot \mathbf{Q}^H \mathbf{Y}_g^\top \mathbf{U_x}\right)\left(\mathbf{U_x}^\top \mathbf{L\Gamma}^\top\right)\right)$, where $\mathbf{L\Gamma L}^\top = \mathbf{U_x D_x U_x}^\top$ and $[\mathbf{\Pi}]_{l,m} = \frac{1}{\sigma_m^2 + p_l d_m}$ for $l = 1, \dots, L$ and $m = 1, \dots, M$.
  Calculate $\mathbf{M}_{\mathrm{time}}^k$, and update $\mathbf{B}$ based on Riemannian update for Toeplitz matrices using circulant embedding.
  Calculate $\mathbf{M}_{\mathrm{SN}}^k$, and update $\mathbf{H}$.
  $k \leftarrow k + 1$
until *stopping condition is satisfied:* $\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|_2^2 \leq \epsilon$ *or* $k = k_{max}$;
Calculate the posterior covariance as $\mathbf{\Sigma_x} = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{D} \mathbf{\Sigma}_0$.

Full Dugh: Temporal Covariance Update

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{1/2}\left((\mathbf{B}^k)^{-1/2}\mathbf{M}_{\mathrm{time}}^k(\mathbf{B}^k)^{-1/2}\right)^{1/2}(\mathbf{B}^k)^{1/2}$$

$$\mathbf{M}_{\mathrm{time}}^k := \frac{1}{MG}\sum_{g=1}^{G}\mathbf{Y}_g^\top\left(\mathbf{\Sigma}_\mathbf{y}^k\right)^{-1}\mathbf{Y}_g$$

Thin Dugh: Temporal Covariance Update

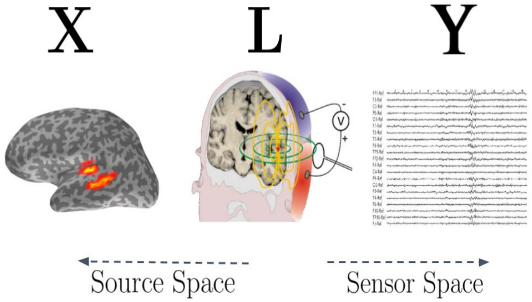
$$\mathbf{B} = \mathbf{QPQ}^H, \; p_l^{k+1} \leftarrow \sqrt{\frac{\hat{g}_l^k}{\hat{z}_l^k}} \text{ for } l = 1, \dots, L$$

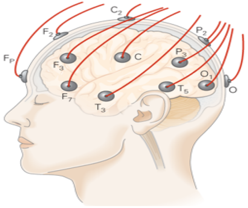
$$\hat{\mathbf{g}} := \mathrm{diag}(\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}_{\mathrm{time}}^k(\mathbf{B}^k)^{-1}\mathbf{QP}^k)$$

$$\hat{\mathbf{z}} := \mathrm{diag}(\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{Q})$$

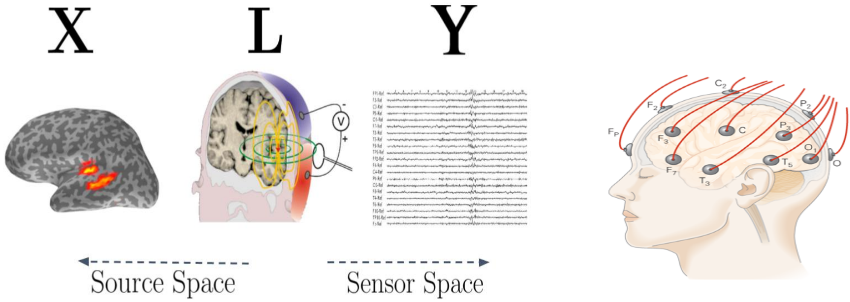
Electromagnetic Brain Source Imaging (BSI)

Electro-/Magnetoencephalography (E/MEG): A non-invasive brain imaging technique with high temporal resolution (order of ms).





Electromagnetic Brain Source Imaging (BSI)

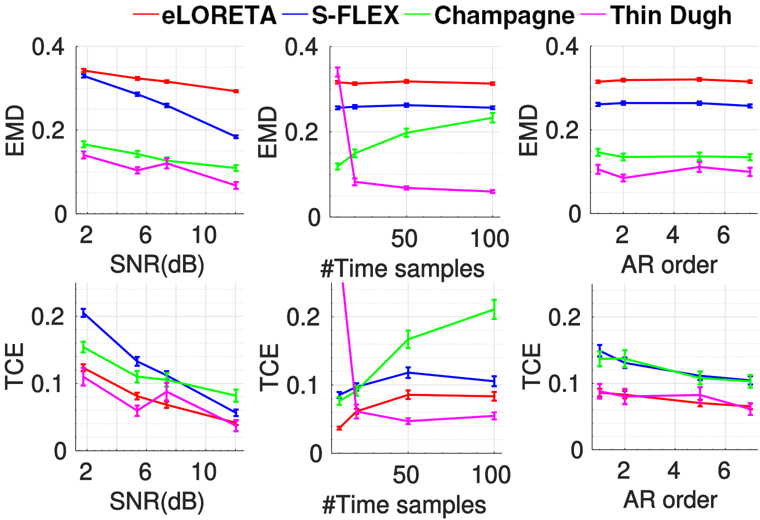


Ill-posed inverse problem: (#Sensors$= 32 \sim 256$ vs #Sources$= 10^3 \sim 10^4$)

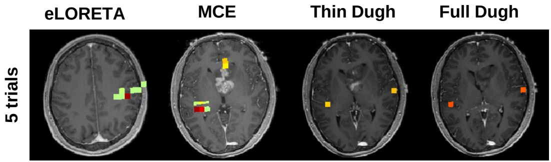
$$\mathbf{X}^* = \underset{\mathbf{X}}{\operatorname{argmin}} \underbrace{\|\mathbf{Y} - \mathbf{L}\mathbf{X}\|_F^2}_{\text{Likelihood:}p(\mathbf{Y}|\mathbf{X})} + \lambda \underbrace{\mathcal{R}(\mathbf{X})}_{\text{Prior:}p(\mathbf{X})}$$

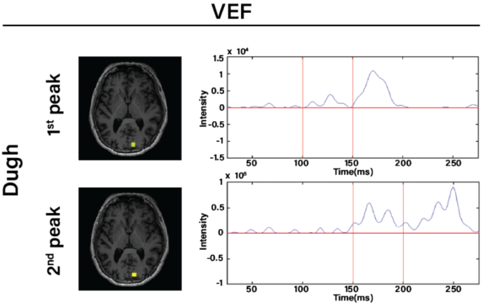
1. Type-I MAP methods: ℓ_1, ℓ_2, $\ell_{1,2}$-norms, sparsity in transformed domains (Gabor).
   [Pascual-Marqui et al., '07][Haufe et al, '08, '11][Gramfort et al., '12, '13][Castaño-Candamil et al., '15]
2. Type-II ML approaches: different sparse Bayesian learning (SBL) variants ignoring the temporal dynamics.
   [Wipf et al., '09, '10, '11][Owen et al, '12][Cai et al., '17, '21]

Conclusion I: Dugh consistently outperforms benchmark methods in the BSI literature according to all evaluation metrics.



Conclusion II: Dugh can provide accurate reconstruction even under extreme SNR conditions - superior to benchmarks.





Thank you for your attention!

References

A. Hashemi, C. Cai, K.-R. Müller, S. S. Nagarajan and S. Haufe
Joint Hierarchical Bayesian Learning of Full-structure Noise for Brain Source Imaging.
Medical Imaging meets NeurIPS (Med-NeurIPS) Workshop, 2020.

A. Hashemi, Y. Gao, C. Cai, S. Ghosh, K.-R. Müller, S. S. Nagarajan and S. Haufe
Joint Learning of Full-structure Noise in Hierarchical Bayesian Regression Models.
Preprint, 2021. Draft is available on bioRxiv.

A. Hashemi, C. Cai, G. Kutyniok, K.-R. Müller, S. S. Nagarajan and S. Haufe
Unification of sparse Bayesian learning algorithms for electromagnetic brain imaging with the majorization minimization framework.
NeuroImage 239, 2021.

C. Cai, A. Hashemi, M. Diwakar, S. Haufe, K. Sekihara, S. S. Nagarajan
Robust estimation of noise for electromagnetic brain imaging with the Champagne algorithm.
NeuroImage 225, 2021.

A. Hashemi and S. Haufe
Improving EEG Source Localization Through Spatio-Temporal Sparse Bayesian Learning.
26th IEEE European Signal Processing Conference (EUSIPCO), 2018.

K. Sekihara and S. S. Nagarajan
Electromagnetic Brain Imaging: A Bayesian Perspective.
Springer, 2015.