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

Ali Hashemi[1,*], Yijing Gao[2], Chang Cai[2], Sanjay Ghosh[2], Klaus-Robert Müller[1,3,4,*], Srikantan S. Nagarajan[2,*], and Stefan Haufe[1,5,6,*]

[1]Technische Universität Berlin, Germany [2]University of California, San Francisco, USA [3]Max Planck Institute for Informatics, Saarbrücken, Germany [4]Department of Artificial Intelligence, Korea University, Seoul, South Korea [5]Physikalisch-Technische Bundesanstalt Braunschweig und Berlin, Germany [6]Charité – Universitätsmedizin Berlin, Germany

*Emails:{hashemi,haufe,klaus-robert.mueller}@tu-berlin.de & sri@ucsf.edu

Introduction

Multi-task Linear Regression Problem:

$\mathbf{Y}_g = \mathbf{L}\mathbf{X}_g + \mathbf{E}_g$ 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,

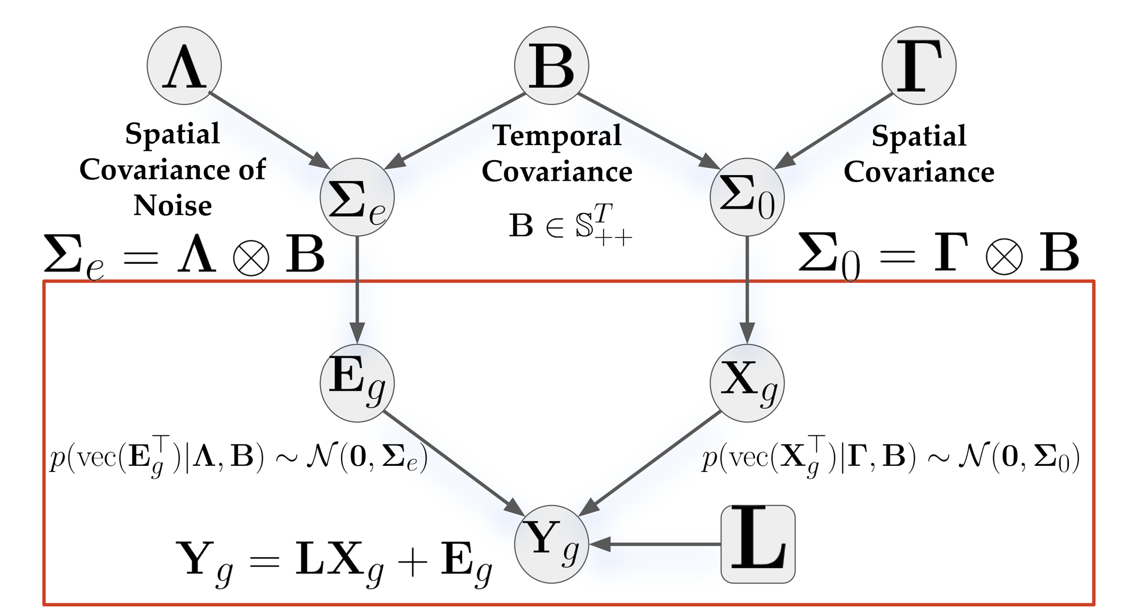
$\mathbf{L} \in \mathbb{R}^{M\times N}$ forward matrix (known): maps $\mathbf{X}_g$ to $\mathbf{Y}_g$

Goal: Estimate $\{\mathbf{X}_g\}_{g=1}^{G}$ given $\mathbf{L}$ and $\{\mathbf{Y}_g\}_{g=1}^{G}$, with a wide range of applications including inverse problem in physics or sparse signal recovery problem in signal processing.

Hierarchical Bayesian Learning

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

Probabilistic graphical model:



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

$$\bar{\mathbf{x}}_g = \mathrm{vec}(\bar{\mathbf{X}}_g^\top) = \Sigma_0 \mathbf{D}^\top \tilde{\Sigma}_{\mathbf{y}}^{-1} \mathbf{y}_g \qquad \tilde{\Sigma}_{\mathbf{y}} = \Sigma_{\mathbf{y}} \otimes \mathbf{B}$$
$$\Sigma_{\mathbf{x}} = \Sigma_0 - \Sigma_0 \mathbf{D}^\top \tilde{\Sigma}_{\mathbf{y}}^{-1} \mathbf{D} \Sigma_0 \qquad \Sigma_{\mathbf{y}} = \mathbf{L}\Gamma\mathbf{L}^\top + \Lambda\,,$$

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

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

Type-II Loss

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

Challenges

1. Non-convex Type-II loss: 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 (MM) optimization [Sun et al., '17][Hashemi et al., '21] with guaranteed convergence properties.

Convex Majorizing Functions and Riemannian Update on the Manifold of P.D. Matrices

Theorem 1: Optimizing $\mathcal{L}_{\text{kron}}(\Gamma, \Lambda, \mathbf{B})$ with respect to $\mathbf{B}$ is equivalent to optimizing the following convex surrogate function, which *majorizes* $\mathcal{L}_{\text{kron}}(\Gamma, \Lambda, \mathbf{B})$:

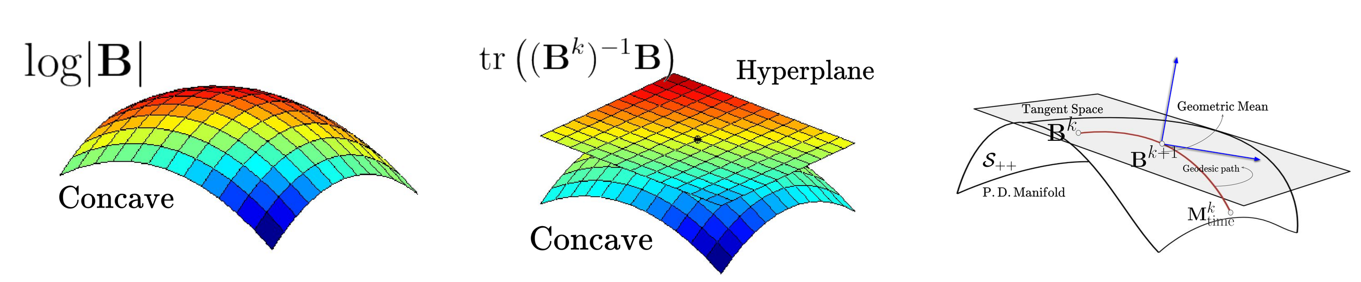
$$\mathcal{L}_{\text{conv}}^{\text{time}}(\Gamma^k, \Lambda^k, \mathbf{B}) = \mathrm{tr}\left((\mathbf{B}^k)^{-1}\mathbf{B}\right) + \mathrm{tr}(\mathbf{M}_{\text{time}}^k \mathbf{B}^{-1}),$$

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

Theorem 2: The cost function $\mathcal{L}_{\text{conv}}^{\text{time}}(\Gamma^k, \Lambda^k, \mathbf{B})$ 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 an MM algorithm with convergence guarantees $\rightsquigarrow$ **Full Dugh**

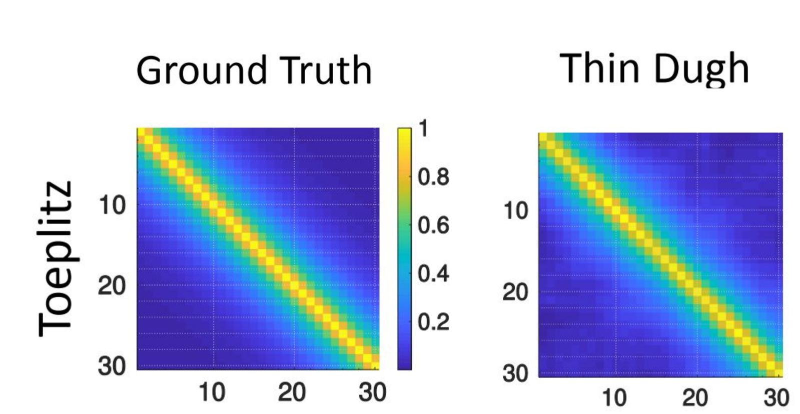


Circulant Embedding for Toeplitz Matrices

Theorem 3: Let $\mathcal{L}_{\text{conv}}^{\text{time}}(\Gamma^k, \Lambda^k, \mathbf{B})$ 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:

$$p_l^{k+1} \leftarrow \sqrt{\frac{\hat{g}_l^k}{\hat{z}_l^k}} \text{ for } l = 1, \ldots, L\,, \text{ where}$$
$$\hat{\mathbf{g}} := \mathrm{diag}(\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}_{\text{time}}^k(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k)$$
$$\hat{\mathbf{z}} := \mathrm{diag}(\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{Q})\,,$$

which leads to an MM algorithm with convergence guarantees $\rightsquigarrow$ **Thin Dugh**



Results

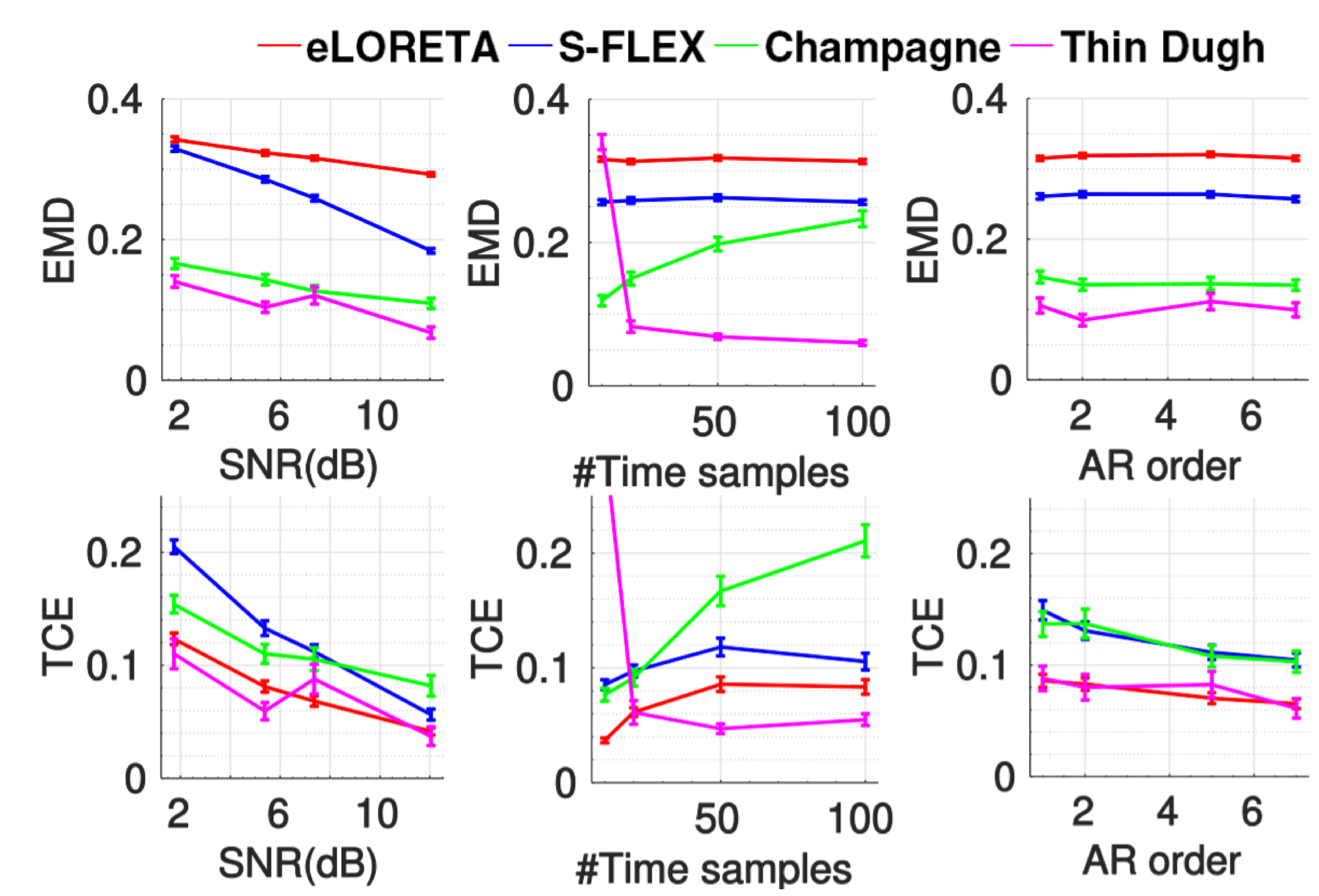
Benchmark methods:

- eLORETA: Represents a smooth inverse solution based on ℓ_2^2-norm minimization [Pascual-Marqui et al., '07]
- MCE: Sparse Type-I method based on ℓ_1-norm minimization [Matsuura et al, '95]
- S-FLEX: Sparse Type-I method based on $\ell_{1,2}$-norm minimization [Haufe et al, '08]
- Champagne or Sparse Bayesian Learning (SBL): A Type-II method ignoring the temporal dynamics [Wipf et al., '10]

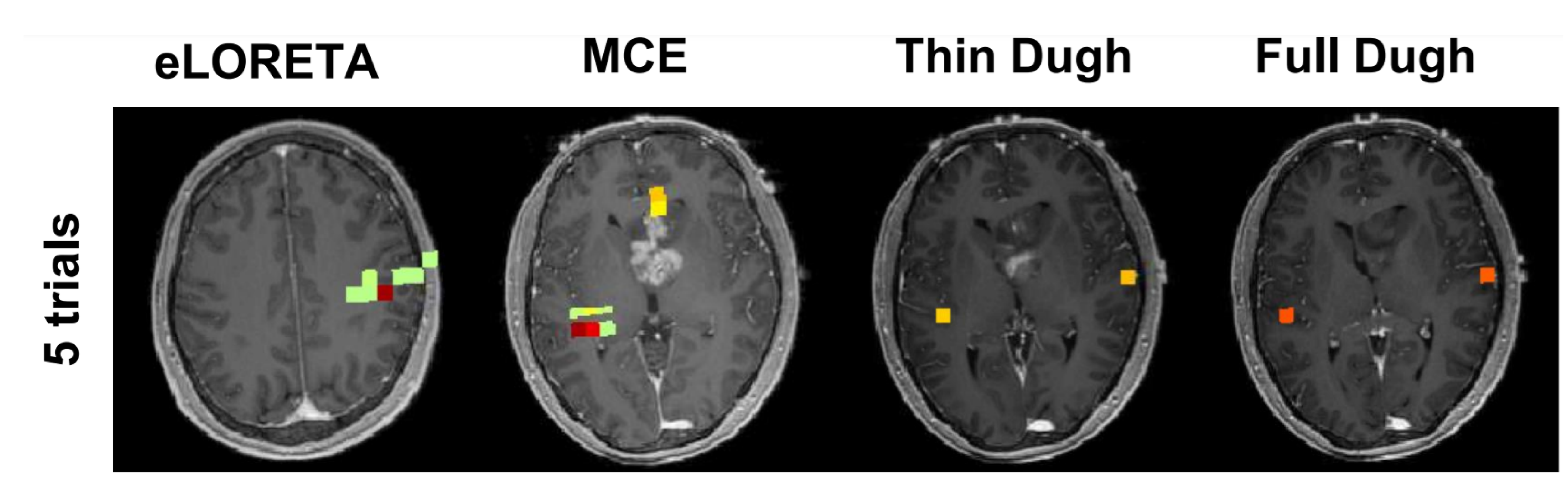
Evaluation metrics:

- Earth Mover's Distance (EMD): Quantifies the spatial localization accuracy
- Time-course Correlation Error (TCE): Measures the similarity in temporal domain using Pearson correlation metric

Numerical simulation: Thin Dugh consistently outperforms benchmark methods in the brain source imaging (BSI) literature according to all evaluation metrics. Note that since thin Dugh incorporates the temporal structure of the sources into the inference scheme, its performance with respect to both evaluation metrics can be significantly improved by increasing the number of time samples.



Real data analysis: We demonstrate the performance of our proposed methods on real auditory evoked fields (AEF) MEG dataset. Both thin and full Dugh were able to accurately reconstruct bilateral auditory cortical activity from only five trials. So, as we can see here, limiting the number of trials to as few as 5 does not negatively influence the reconstruction result of Dugh methods, while this extreme SNR conditions severely affects the reconstruction performance of competing methods, like eLORETA and MCE.



Acknowledgements

This result is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 758985).