Sparse linear regression with spatially and temporally correlated support

Michael Riis Andersen

Technical University of Denmark

Joint work with

Aki Vehtari, Ole Winther (supervisor) & Lars K. Hansen (supervisor)

October 29, 2014

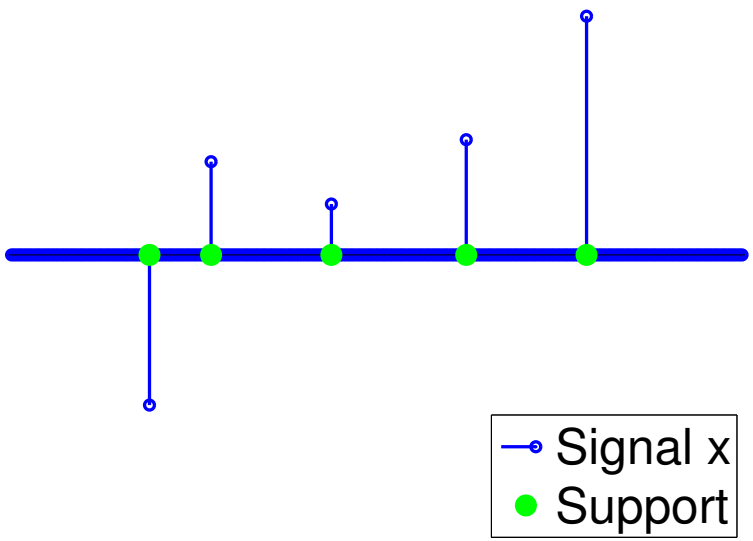
Sparse Linear Regression and Structured Sparsity

Goal

Given $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ for $N < D$, find *structured* sparse $\boldsymbol{x} \in \mathbb{R}^D$ s.t.

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \text{noise}$$

Sparse signal $\boldsymbol{x}$



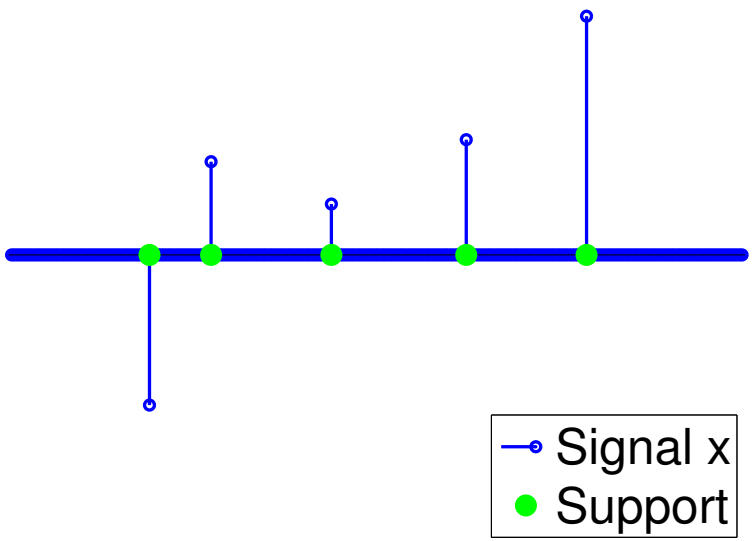
Sparse Linear Regression and Structured Sparsity

Goal

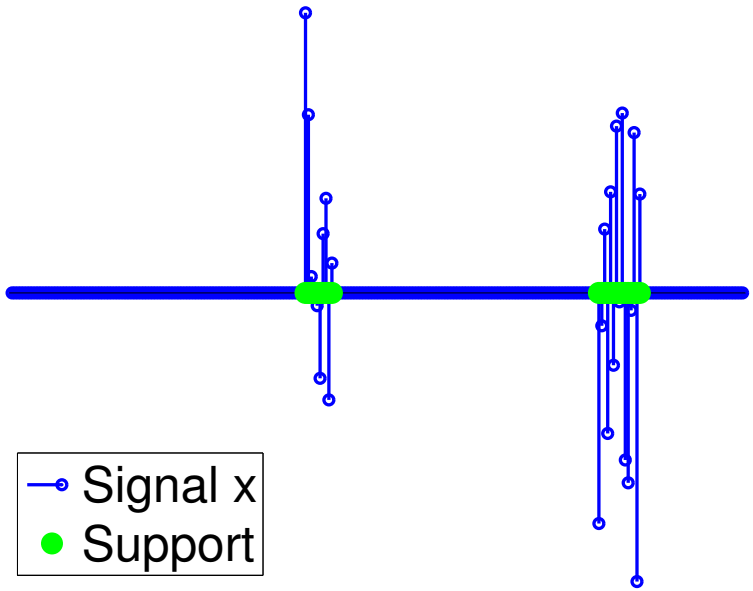
Given $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ for $N < D$, find *structured* sparse $\boldsymbol{x} \in \mathbb{R}^D$ s.t.

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \text{noise}$$

Sparse signal $\boldsymbol{x}$



Structured sparse signal $\boldsymbol{x}$



Multiple sparse regression problems

- In many applications we observe a series of problems of the $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{e}_t$ for different points in time $t = 1, 2.., T$
- By defining as $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_T \end{bmatrix} \in \mathbb{R}^{D \times T}$ and similar for $\boldsymbol{Y}, \boldsymbol{E}$, we can write

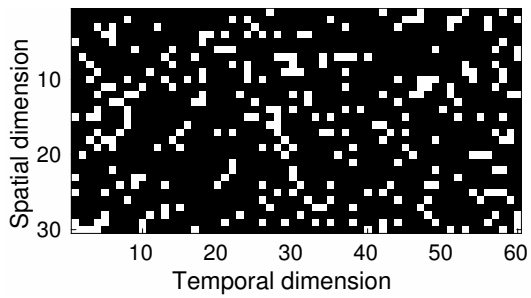
$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}.$$

Multiple sparse regression problems

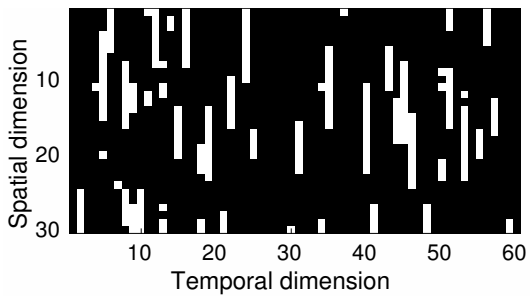
- In many applications we observe a series of problems of the $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{e}_t$ for different points in time $t = 1, 2.., T$
- By defining as $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_T \end{bmatrix} \in \mathbb{R}^{D \times T}$ and similar for $\boldsymbol{Y}, \boldsymbol{E}$, we can write

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}.$$

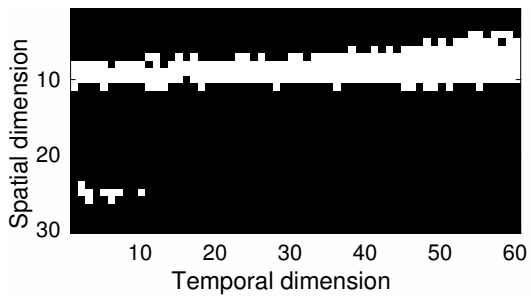
- I.I.D. vs. spatial vs. spatio-temporal sparsity patterns for $\boldsymbol{X}$



(d) I.I.D.



(e) Spatial

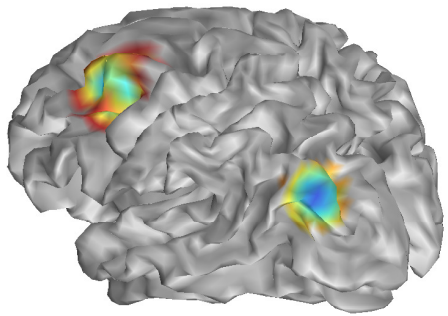


(f) Spatiotemporal

Application: EEG Source Localization

- We observe a multivariate time series $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$ and want to infer the underlying sources $\boldsymbol{X} \in \mathbb{R}^{D \times T}$ given a forward $\boldsymbol{A} \in \mathbb{R}^{N \times D}$.

$$= \boldsymbol{A} \cdot$$



$$+ \boldsymbol{E}$$

EEG measurement		Forward model	Source magnitudes		Noise
$\boldsymbol{Y}$	$=$	$\boldsymbol{A}$	$\boldsymbol{X}$	$+$	$\boldsymbol{E}$

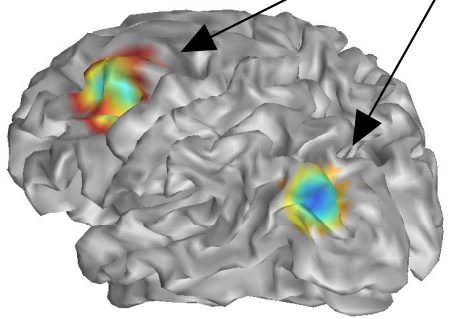
- The brain is modelled using a set of discrete current dipoles

Application: EEG Source Localization

- We observe a multivariate time series $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$ and want to infer the underlying sources $\boldsymbol{X} \in \mathbb{R}^{D \times T}$ given a forward $\boldsymbol{A} \in \mathbb{R}^{N \times D}$.

Sparse and
spatiotemporal structure

$$= \boldsymbol{A} \cdot$$



$$+ \boldsymbol{E}$$

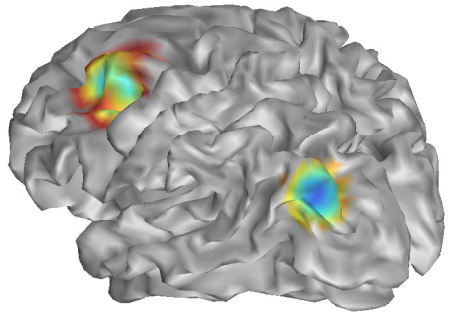
EEG measurement $\boldsymbol{Y}$ $=$ Forward model $\boldsymbol{A}$ Source magnitudes $\boldsymbol{X}$ $+$ Noise $\boldsymbol{E}$

- The brain is modelled using a set of discrete current dipoles

Spike and slab priors for promoting sparsity

- Assume x_i is composed of two variables

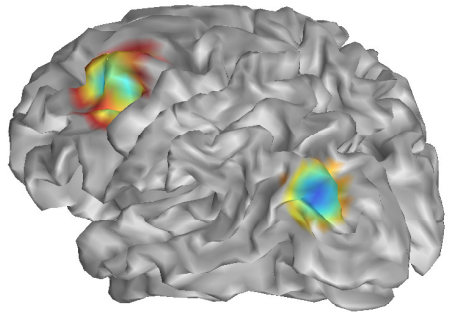
$$x_i = s_i \cdot c_i, \quad s_i \in \{0, 1\}, \quad c_i \in \mathbb{R}$$



Spike and slab priors for promoting sparsity

- Assume x_i is composed of two variables

$$x_i = s_i \cdot c_i, \quad s_i \in \{0, 1\}, \quad c_i \in \mathbb{R}$$



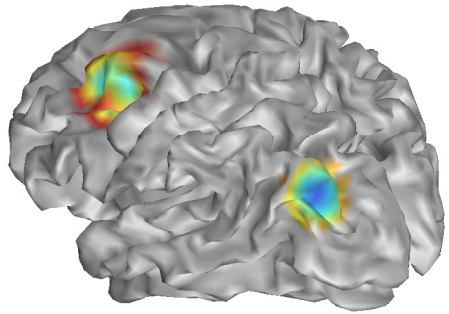
- In terms of probability distributions,

$$p(s_i) = \text{Ber}(p)$$
$$p(x_i|s_i) = (1 - s_i)\delta(x_i) + s_i\mathcal{N}\left(x_i\middle|0, \tau\right)$$

Spike and slab priors for promoting sparsity

- Assume x_i is composed of two variables

$$x_i = s_i \cdot c_i, \quad s_i \in \{0, 1\}, \quad c_i \in \mathbb{R}$$



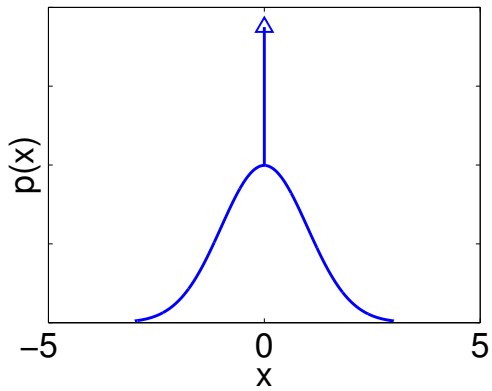
- In terms of probability distributions,

$$p(s_i) = \text{Ber}(p)$$
$$p(x_i|s_i) = (1 - s_i)\delta(x_i) + s_i\mathcal{N}(x_i|0, \tau)$$

- Marginalizing out s_i

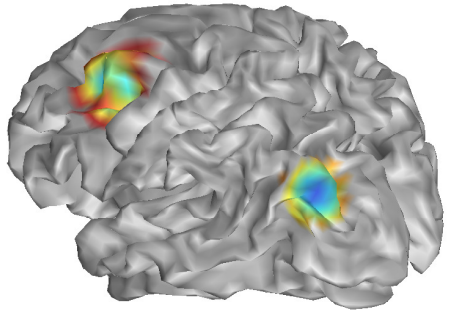
$$p(x_i) = (1 - p)\delta(x_i) + p\mathcal{N}(x_i|0, \tau)$$

Spike Slab



The structured spike and slab prior

- We want to build a prior distribution for $\boldsymbol{x}$ s.t. the binary variables $\boldsymbol{s}$ are spatially correlated



- Idea: Generate a set of correlated random variables

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma} \middle| \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

- and transform them into probabilities using a map $\phi : \mathbb{R} \to (0, 1)$

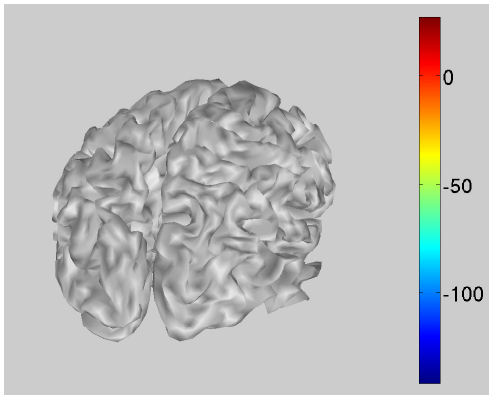
$$p(s_i \middle| \gamma_i) = \text{Ber}\left(s_i \middle| \phi\left(\gamma_i\right)\right)$$
$$p(x_i \middle| s_i) = (1 - s_i)\,\delta\left(x_i\right) + s_i \mathcal{N}\left(x_i \middle| 0, \tau\right)$$

- $\boldsymbol{\Sigma}$ now determines the correlation structure of the support of $\boldsymbol{x}$ and

$$p(s_i = 1) = \phi\left(\frac{\mu_i}{\sqrt{1 + \Sigma_{ii}}}\right)$$

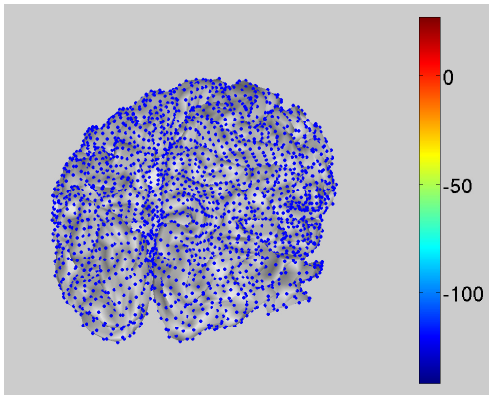
Sampling from the prior

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(s_i\middle|\gamma_i) = \text{Ber}\left(s_i\middle|\phi\left(\gamma_i\right)\right)$$



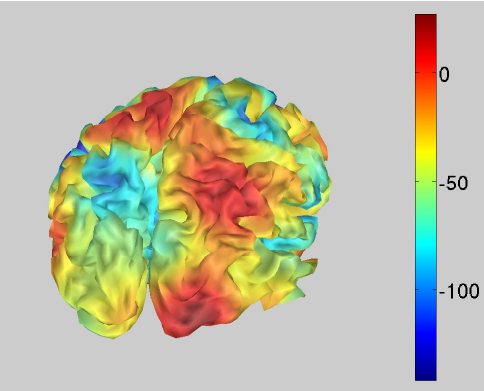
Sampling from the prior

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(s_i\middle|\gamma_i) = \text{Ber}\left(s_i\middle|\phi\left(\gamma_i\right)\right)$$



Sampling from the prior

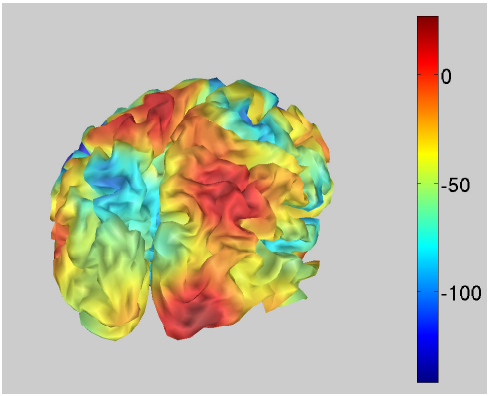
$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(s_i\middle|\gamma_i) = \text{Ber}\left(s_i\middle|\phi\left(\gamma_i\right)\right)$$



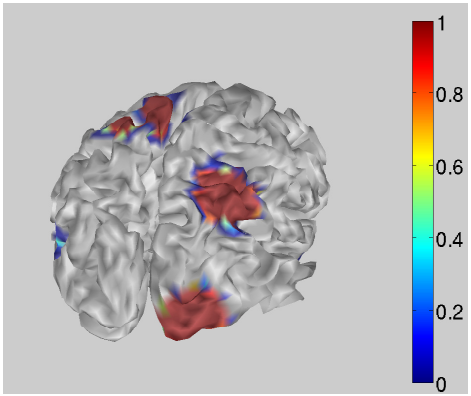
1) $\boldsymbol{\gamma} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

Sampling from the prior

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(s_i\middle|\gamma_i) = \text{Ber}\left(s_i\middle|\phi\left(\gamma_i\right)\right)$$



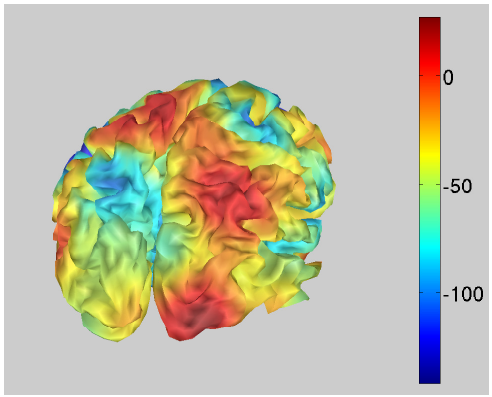
1) $\boldsymbol{\gamma} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$



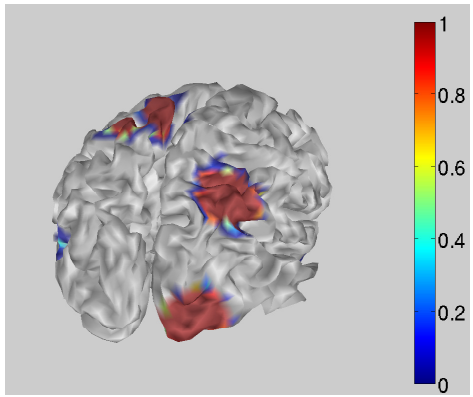
2) Probabilities $\phi\left(\boldsymbol{\gamma}\right)$

Sampling from the prior

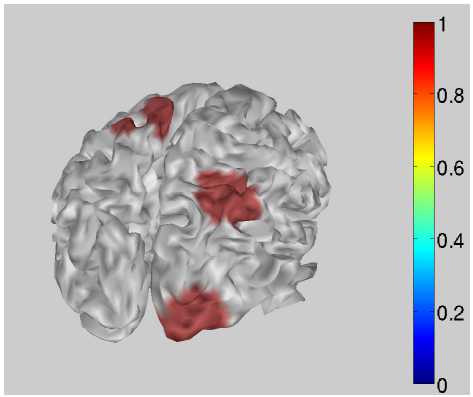
$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(s_i\middle|\gamma_i) = \text{Ber}\left(s_i\middle|\phi\left(\gamma_i\right)\right)$$



1) $\boldsymbol{\gamma} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$



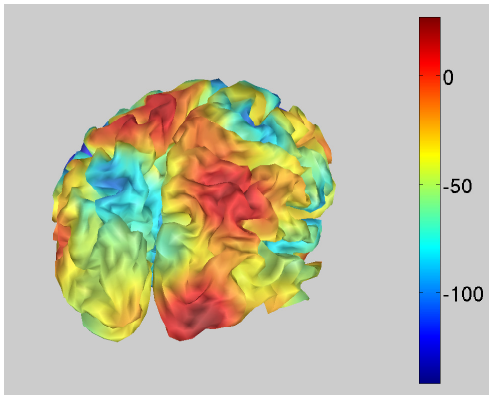
2) Probabilities $\phi\left(\boldsymbol{\gamma}\right)$



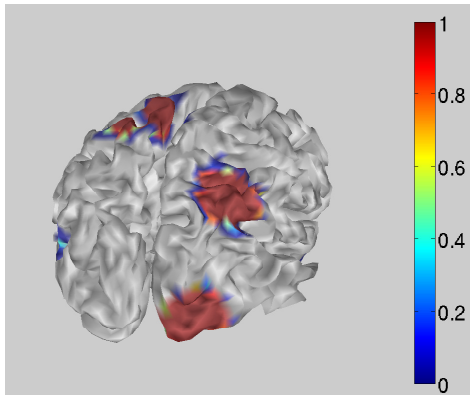
3) $\boldsymbol{s} \sim \text{Ber}\left(\phi\left(\boldsymbol{\gamma}\right)\right)$

Sampling from the prior

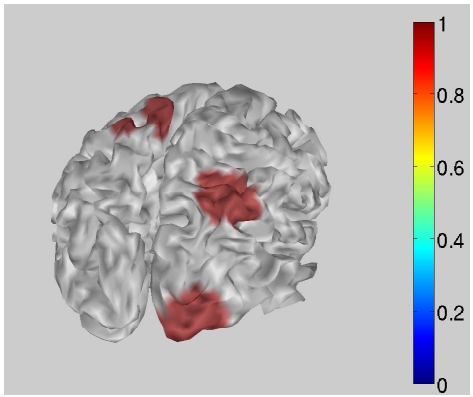
$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(s_i\middle|\gamma_i) = \text{Ber}\left(s_i\middle|\phi\left(\gamma_i\right)\right)$$



1) $\boldsymbol{\gamma} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$

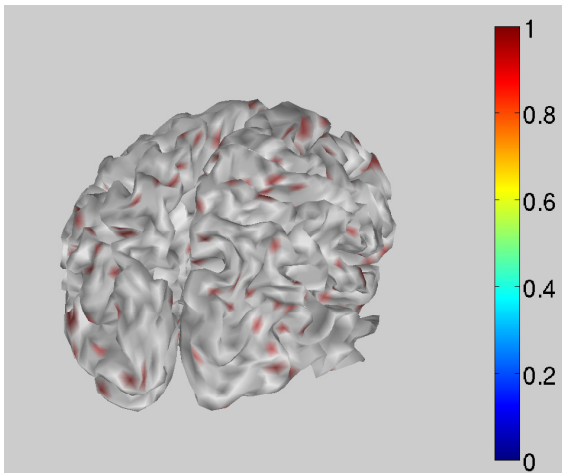


2) Probabilities $\phi\left(\boldsymbol{\gamma}\right)$



3) $\boldsymbol{s} \sim \text{Ber}\left(\phi\left(\boldsymbol{\gamma}\right)\right)$

Compare to sample from standard i.i.d. spike and slab prior



Spatiotemporal spike and slab priors

- Extending the prior to the spatio-temporal case is straightforward
- Let $\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_1 & \gamma_2 \dots \gamma_T \end{bmatrix}$ and $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_1 & \boldsymbol{S}_2 & \dots & \boldsymbol{S}_T \end{bmatrix}$
- Imposing a GP on $\boldsymbol{\Gamma}$

$$p(\boldsymbol{\Gamma}) = \mathcal{N}\left(\boldsymbol{\Gamma} | \boldsymbol{\mu}, \boldsymbol{\Sigma}_0\right)$$

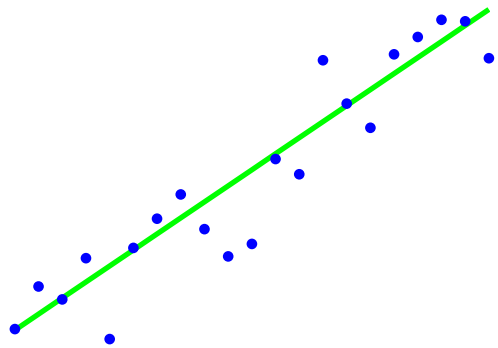
$$p(\boldsymbol{X}|\boldsymbol{S}) = \prod_{t=1}^{T} \prod_{i=1}^{D} \left[(1 - s_{t,i})\delta(x_{t,i}) + s_{t,i}\mathcal{N}\left(x_{t,i}|0, \tau\right)\right]$$

$$p(\boldsymbol{S}|\boldsymbol{\Gamma}) = \prod_{t=1}^{T} \prod_{i=1}^{D} \text{Ber}\left(z_{t,i}|\phi\left(\gamma_{t,i}\right)\right)$$

- Assuming regular sampling in time: $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\text{temporal}} \otimes \boldsymbol{\Sigma}_{\text{spatial}}$

Inference using the spatio-temporal spike and slab prior

- Recall the model $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}$



- Assuming isotropic Gaussian noise leads to a posterior distribution of the form

$$p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma} | \boldsymbol{Y}) \propto \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{y}_t | \boldsymbol{A}\boldsymbol{x}_t, \sigma_0^2 \boldsymbol{I}\right) \prod_{t=1}^{T} \prod_{i=1}^{D} \left[(1 - s_{t,i})\delta\left(x_{t,i}\right) + s_{t,i}\mathcal{N}\left(x_i | 0, \tau_0\right)\right]$$
$$\prod_{t=1}^{T} \prod_{i=1}^{D} \text{Ber}\left(s_{t,i} | \phi\left(\gamma_{t,i}\right)\right) \mathcal{N}\left(\boldsymbol{\Gamma} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$$

- Intractable due to the product over mixtures

- We use (parallel) expectation propagation for approximate inference

Approximate inference using EP

- Factorization of the true posterior

$$p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma} | \boldsymbol{Y}) \propto \prod_{t=1}^{T} \underbrace{\mathcal{N}\left(\boldsymbol{y}_t | \boldsymbol{A}\boldsymbol{x}_t, \sigma_0^2 \boldsymbol{I}\right)}_{f_{1,t}(\boldsymbol{x}_t)} \prod_{t=1}^{T} \prod_{i=1}^{D} \underbrace{\left[(1 - s_{i,t})\delta\left(x_{i,t}\right) + s_{i,t}\mathcal{N}\left(x_i | 0, \tau_0\right)\right]}_{f_{2,t,i}\left(x_{i,t}, s_{i,t}\right)}$$

$$\prod_{t=1}^{T} \prod_{i=1}^{D} \underbrace{\text{Ber}\left(s_{i,t} | \phi\left(\gamma_{i,t}\right)\right)}_{f_{3,t,i}\left(s_{i,t}, \gamma_{i,t}\right)} \underbrace{\mathcal{N}\left(\boldsymbol{\Gamma} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)}_{f_4(\boldsymbol{\Gamma})}$$

- Site approximations

$$f_{2,t,i}(x_{i,t}, s_{i,t}) \approx \tilde{f}_{2,t,i}(x_{i,t}, s_{i,t}) = \mathcal{N}\left(x_{i,t} | \tilde{m}_{2,t,i}, \tilde{v}_{2,t,i}\right) \text{Ber}\left(s_{i,t} | \phi\left(\tilde{\gamma}_{2,t,i}\right)\right)$$

$$f_{3,t,i}(s_{i,t}, \gamma_{i,t}) \approx \tilde{f}_{3,t,i}(s_{i,t}, \gamma_{i,t}) = \text{Ber}\left(s_{i,t} | \phi\left(\tilde{\gamma}_{3,t,i}\right)\right) \mathcal{N}\left(\gamma_{i,t} | \tilde{\mu}_{3,t,i}, \tilde{\Sigma}_{3,t,i}\right)$$

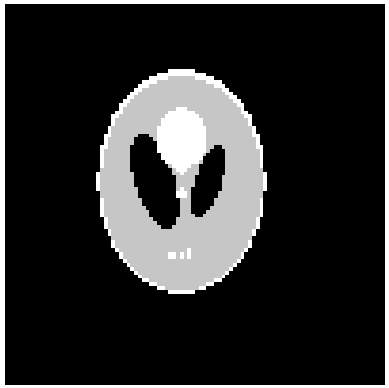
- Approximate posterior distribution has the form

$$Q\left(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma}\right) = \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{x}_t | \hat{\boldsymbol{m}}_t, \hat{\boldsymbol{V}}_t\right) \prod_{t=1}^{T} \text{Ber}\left(\boldsymbol{s}_t | \phi\left(\hat{\boldsymbol{\gamma}}_t\right)\right) \mathcal{N}\left(\boldsymbol{\Gamma} | \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right).$$

Toy example: Recovery of Shepp-Logan Phantom I

Synthetic data generated using

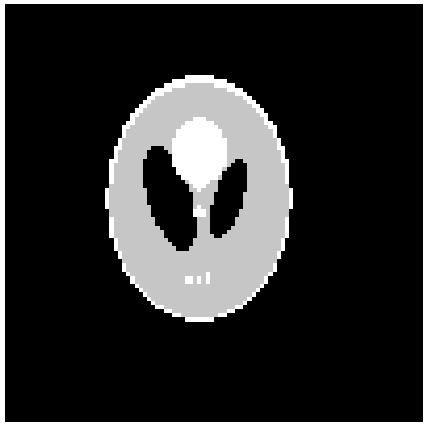
$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e},$$

$\boldsymbol{x}_0 =$ 

- Spatial example, i.e. $T = 1$
- Image size is 100×100, i.e. $\boldsymbol{x}_0 \in \mathbb{R}^{10^4}$
- Number of non-zero pixels is $K = 1723$
- Number of measurements is $1.5K \approx 2585$, i.e. $\boldsymbol{y} \in \mathbb{R}^{2585}$
- SNR is 20dB
- Matrix $\boldsymbol{A}$ is Gaussian I.I.D., i.e. $A_{ij} \sim \mathcal{N}(0, 1)$
- $\boldsymbol{\Sigma}$ = squared exponential covariance defined on the image grid

Toy example: Recovery of Shepp-Logan Phantom II

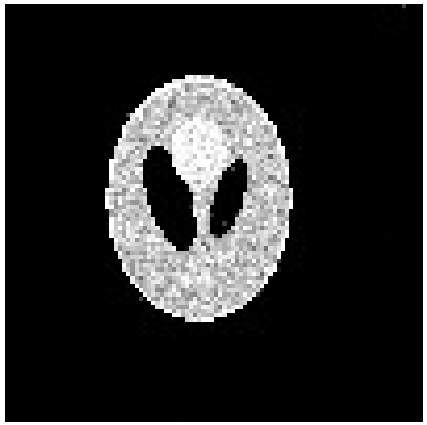
Weights



$\boldsymbol{x}_0$

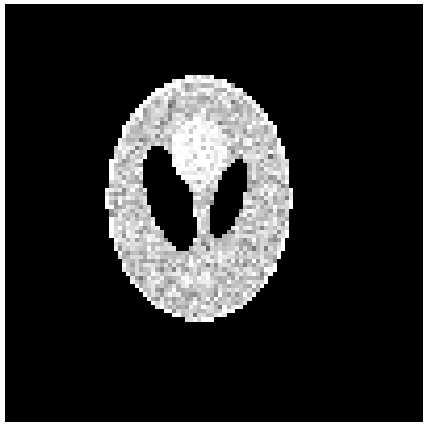
NMSE $= 0.7996$

$\hat{\boldsymbol{x}}_{\text{spike and slab}}$



NMSE $= 0.1444$

$\hat{\boldsymbol{x}}_{\text{structured}}$

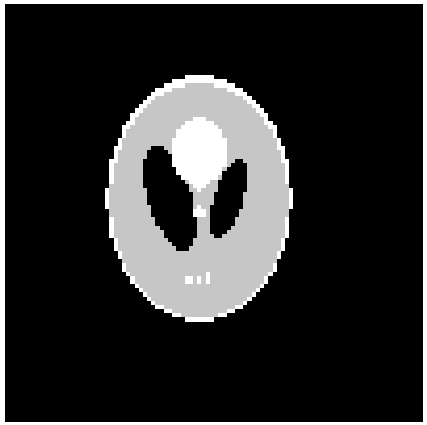


NMSE $= 0.1401$

$\hat{\boldsymbol{x}}_{\text{oracle}}$

Toy example: Recovery of Shepp-Logan Phantom II

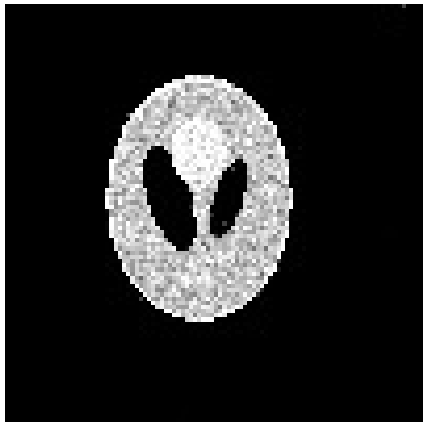
Weights



$\boldsymbol{x}_0$

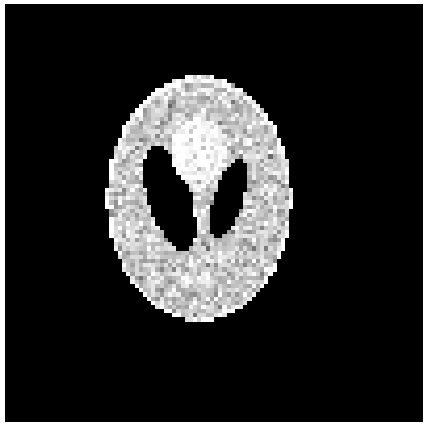
NMSE $= 0.7996$

$\hat{\boldsymbol{x}}_{\text{spike and slab}}$



NMSE $= 0.1444$

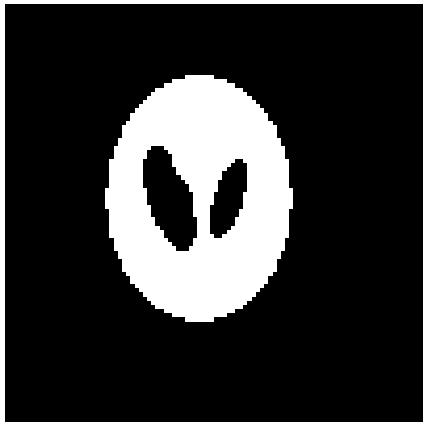
$\hat{\boldsymbol{x}}_{\text{structured}}$



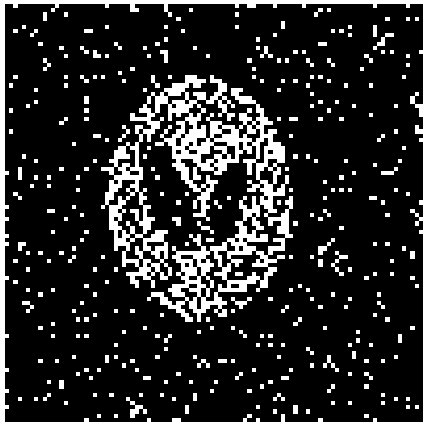
NMSE $= 0.1401$

$\hat{\boldsymbol{x}}_{\text{oracle}}$

Support

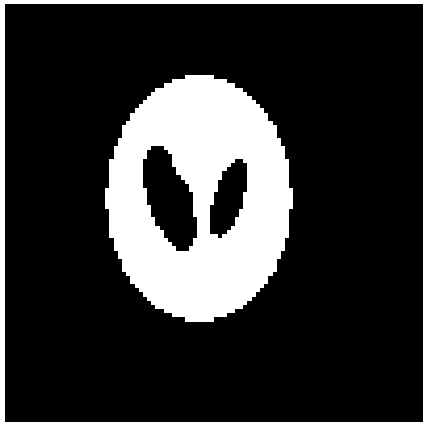


$\boldsymbol{s}_0$



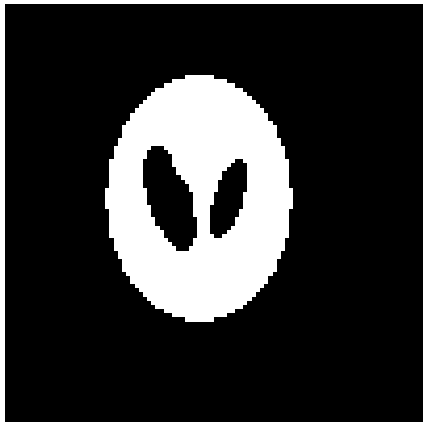
F $= 0.2436$

$\hat{\boldsymbol{s}}_{\text{spike and slab}}$



F $= 0.9997$

$\hat{\boldsymbol{s}}_{\text{structured}}$



F $= 1.000$

$\boldsymbol{s}_{\text{oracle}}$

The computational bottlenecks

- The computational bottlenecks are the updates of the posterior covariance matrices for $\boldsymbol{X} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{DT \times DT}$

- For $\boldsymbol{x}_t$ for $t = 1, .., T$:

$$\hat{\boldsymbol{V}}_t^* = \left[\frac{1}{\sigma^2} \boldsymbol{A}^T \boldsymbol{A} + \boldsymbol{V}_{2,t}^{-1} \right]^{-1}$$

- For $\boldsymbol{\Gamma}$:

$$\hat{\boldsymbol{\Sigma}}^* = \left[\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_{3,t}^{-1} \right]^{-1}$$

- Direct implementation $\mathcal{O}\left(TD^3\right)$ and $\mathcal{O}\left(T^3D^3\right)$, respectively.

Updating posterior covariance for $\boldsymbol{X}$

- Direct implementation: $\mathcal{O}\left(T^3 D^3\right)$
- For $\boldsymbol{X}$ we can use the assumptions $N \ll D$ and $\boldsymbol{V}_{1,t}^{-1} = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{A}$ and apply the matrix inversion lemma, so we can update each $\boldsymbol{V}_{1,t}$ as

$$\tilde{\boldsymbol{V}}_t = \tilde{\boldsymbol{V}}_{2,t} - \tilde{\boldsymbol{V}}_{2,t}\boldsymbol{A}^T\left(\sigma_0^2\boldsymbol{I} + \boldsymbol{A}\tilde{\boldsymbol{V}}_{2,t}\boldsymbol{A}^T\right)^{-1}\boldsymbol{A}\tilde{\boldsymbol{V}}_{2,t}.$$

- We only need diagonal of $\boldsymbol{V}_t$, i.e. $\mathcal{O}\left(N^2 D\right)$.
- Total cost per iteration: $\mathcal{O}\left(TN^2 D\right)$.
- Unfortunately, we cannot do the same for the posterior covariance matrix for $\boldsymbol{\Gamma} \in \mathbb{R}^{DT \times DT}$

Approximations

To reduce the computational complexity, we consider three different approximation schemes for the posterior covariance matrix for $\mathbf{\Gamma}$

Approximation		**Complexity**	**Storage**
Standard EP	(EP)	$\mathcal{O}\left(T^3D^3\right)$	$\mathcal{O}\left(T^2D^2\right)$
Low rank	(LR)	$\mathcal{O}\left(TDK^2\right)$	$\mathcal{O}\left(TDK\right)$
Grouping/subsampling	(G)	$\mathcal{O}\left(T_g^3D_g^3\right)$	$\mathcal{O}\left(T_g^2D_g^2\right)$
Common site precision	(CP)	$\mathcal{O}\left(T^2D + TD^2\right)$	$\mathcal{O}\left(T^2 + D^2 + TD\right)$

- The costs above is only for updating posterior for $\mathbf{\Gamma}$
- The cost for updating $\boldsymbol{X}$ is $\mathcal{O}\left(N^2D\right)$

The low rank approximation

- Use low rank approximation for prior covariance matrix for $\mathbf{\Gamma}$

$$\mathbf{\Sigma}_0 = \mathbf{\Sigma}_{\text{spatial}} \otimes \mathbf{\Sigma}_{\text{temporal}} \approx \boldsymbol{USU}^T + \mathbf{\Lambda}$$

  where $\boldsymbol{U} \in \mathbb{R}^{DT \times K}$, $\boldsymbol{S} \in \mathbb{R}^{K \times K}$ and $\mathbf{\Lambda} \in \mathbb{R}^{DT \times DT}$ is a diagonal matrix

- We can obtain eigenvectors of $\mathbf{\Sigma}_0$ from eigendecompositions of $\mathbf{\Sigma}_{\text{spatial}}$ and $\mathbf{\Sigma}_{\text{temporal}}$

- We can now apply matrix inversion lemma as before to get $\mathcal{O}\left(TDK^2\right)$

- In practice: K can increases with both D and T!

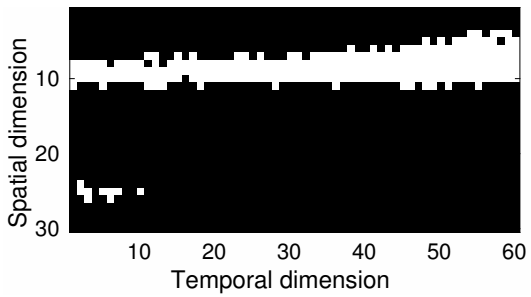
Grouping/subsampling

- If the length-scales are large relative to the grid size, we can group neighbouring spatial locations and assign a single $\overline{\gamma}$ variable for each group and thus reducing the dimension of $\mathbf{\Gamma}$

$$p(\boldsymbol{S}|\overline{\mathbf{\Gamma}}) = \prod_{i,t} \text{Ber}\left(s_{i,t}|\phi\left(\overline{\gamma}_{g(t,i)}\right)\right), \qquad p(\overline{\mathbf{\Gamma}}) = \mathcal{N}\left(\mathbf{\Gamma}|\boldsymbol{\mu}_0, \mathbf{\Sigma}_0\right)$$

  where $g : \mathcal{N} \times \mathcal{N} \rightarrow \mathcal{N}$ is function that assign each pair (t, i) to a group.

- Two variables in the same group do *not* share the same binary $z_{i,t}$ variable, but only the same probability of being active.



The common site precision approximation

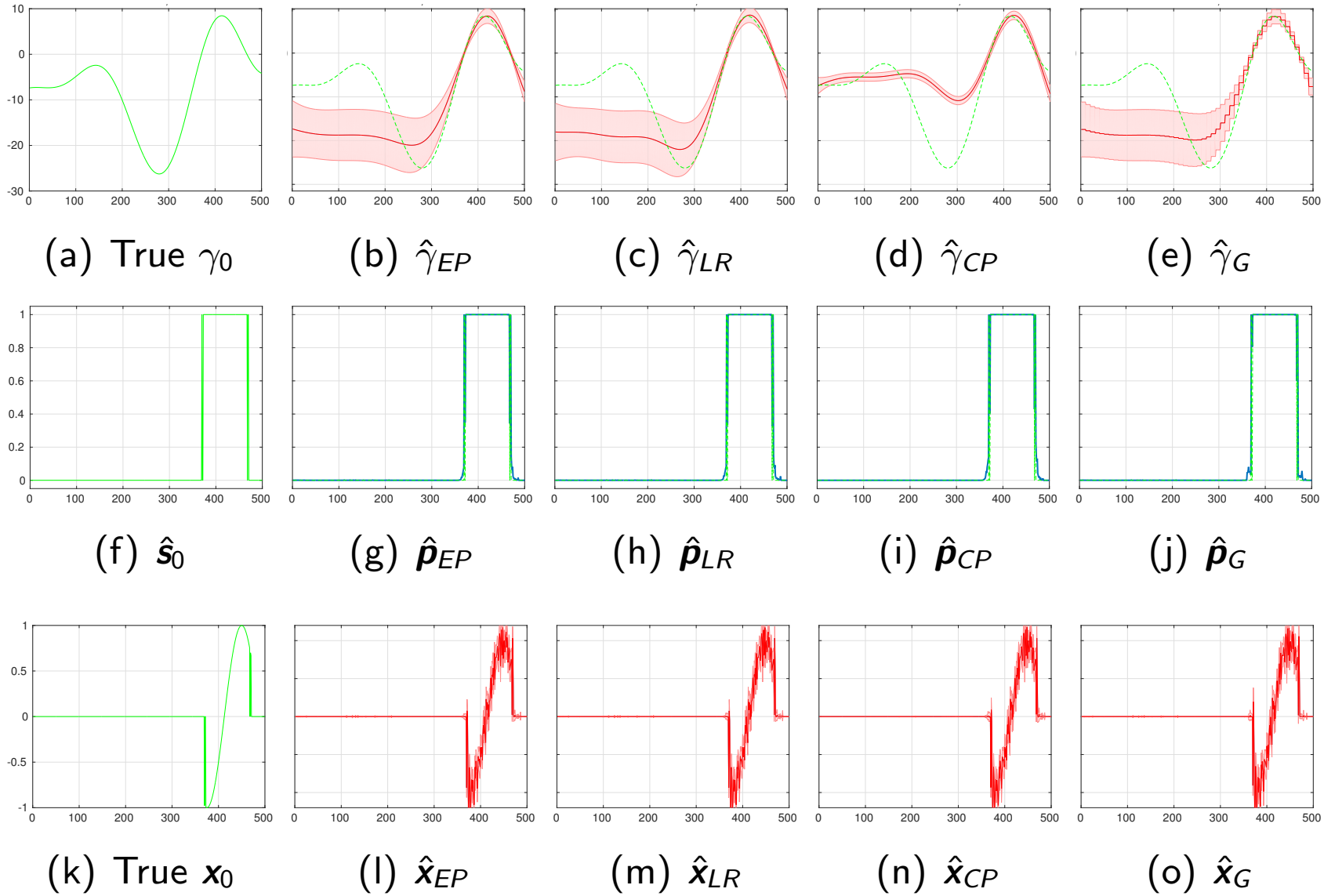
$$f_{3,t,i}(s_{i,t}, \gamma_{i,t}) \approx \tilde{f}_{3,t,i}(s_{i,t}, \gamma_{i,t}) = \text{Ber}\left(s_{i,t} | \phi\left(\tilde{\gamma}_{3,t,i}\right)\right) \mathcal{N}\left(\gamma_{i,t} | \tilde{\mu}_{3,t,i}, \tilde{\Sigma}_{3,t,i}\right)$$

- The idea is to approximate the site precisions/variances for $\boldsymbol{\Gamma}$ with their common mean precision/variance

$$\begin{aligned}
\left(\boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_t + \tilde{\boldsymbol{\Sigma}}_3\right)^{-1} &\approx \left(\boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_t + \bar{\Sigma}_3 \boldsymbol{I}\right)^{-1} \\
&= \left[\left(\boldsymbol{U}_s \otimes \boldsymbol{U}_t\right)\left(\boldsymbol{S}_s \otimes \boldsymbol{S}_t\right)\left(\boldsymbol{U}_s^T \otimes \boldsymbol{U}_t^T\right) + \bar{\Sigma}_3 \boldsymbol{I}\right]^{-1} \\
&= \left(\boldsymbol{U}_s \otimes \boldsymbol{U}_t\right)\left(\boldsymbol{S}_s \otimes \boldsymbol{S}_t + \bar{\Sigma}_3 \boldsymbol{I}\right)^{-1}\left(\boldsymbol{U}_s^T \otimes \boldsymbol{U}_t^T\right),
\end{aligned}$$

- Using properties of Kronecker products, we can compute the diagonal of the covariance matrix in $\mathcal{O}\left(T^2 D + D^2 T\right)$
- If the spatial positions are on a grid: $\boldsymbol{\Sigma}_{\boldsymbol{s}} = \boldsymbol{\Sigma}_x \otimes \boldsymbol{\Sigma}_y \otimes \boldsymbol{\Sigma}_z$
- This approximation is very suitable for grid data

Comparing the approximation schemes for 1D problem



(a) True γ_0 (b) $\hat{\gamma}_{EP}$ (c) $\hat{\gamma}_{LR}$ (d) $\hat{\gamma}_{CP}$ (e) $\hat{\gamma}_{G}$

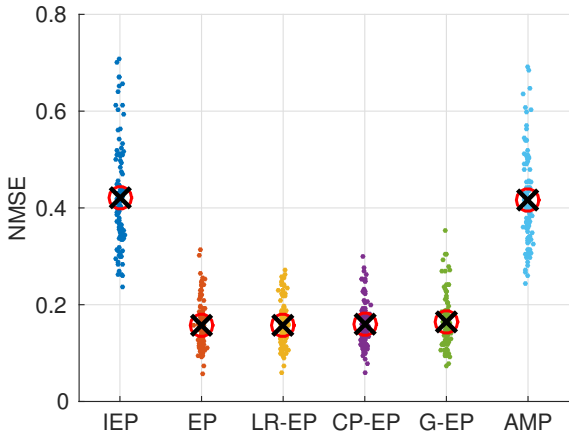
(f) $\hat{\boldsymbol{s}}_0$ (g) $\hat{\boldsymbol{p}}_{EP}$ (h) $\hat{\boldsymbol{p}}_{LR}$ (i) $\hat{\boldsymbol{p}}_{CP}$ (j) $\hat{\boldsymbol{p}}_{G}$

(k) True $\boldsymbol{x}_0$ (l) $\hat{\boldsymbol{x}}_{EP}$ (m) $\hat{\boldsymbol{x}}_{LR}$ (n) $\hat{\boldsymbol{x}}_{CP}$ (o) $\hat{\boldsymbol{x}}_{G}$

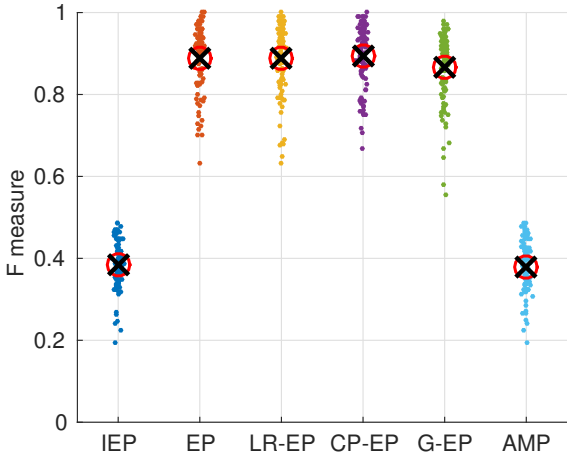
Figure: Data is generated from model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$.

Evaluation of methods

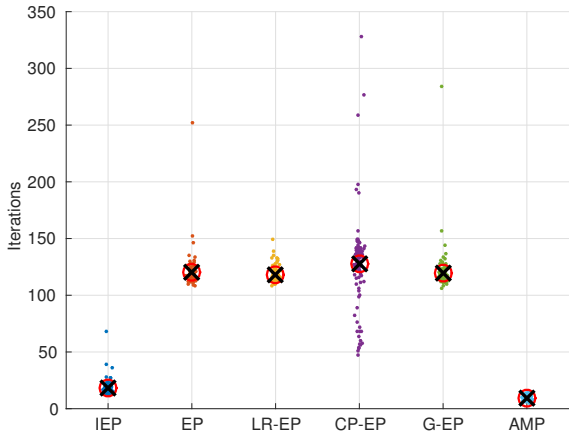
- Repeat for 100 different samples



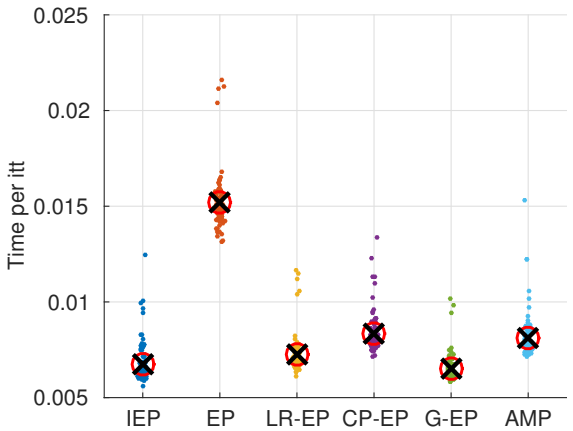
(a) NMSE



(b) F_1 score



(c) Iterations



(d) Time pr. itt

Spatiotemporal example

- We illustrate the spatiotemporal prior with a small simulated example
- We create a test signal $\boldsymbol{X} \in \mathbb{R}^{D \times T}$ with non-stationary sparsity pattern

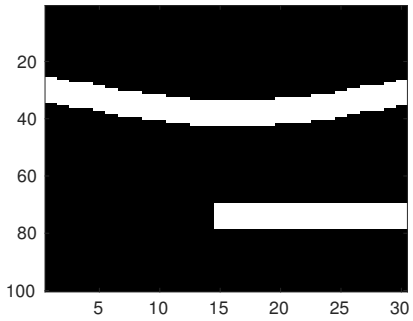
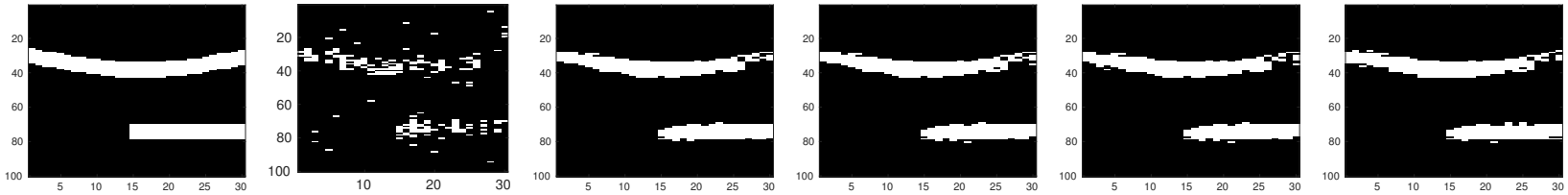
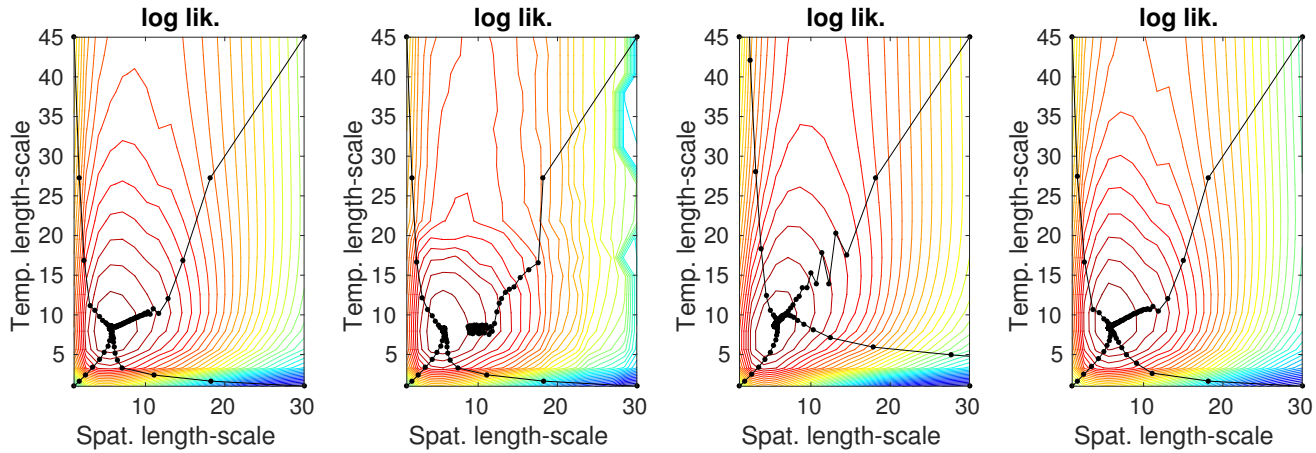


Figure: Sparsity pattern for test signal

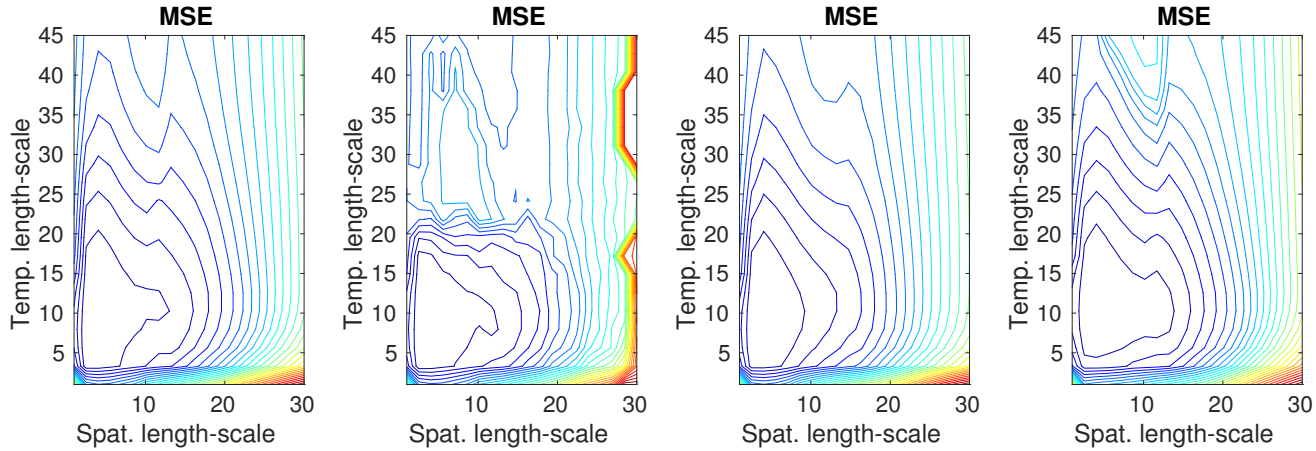
- We generate observations from the model $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}$ with $D = 100,\ T = 30,\ N = \frac{1}{3}D$ and the EP scheme to reconstruct to true signal
- $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{spatial} \otimes \boldsymbol{\Sigma}_{temporal}$ with sq. exponential covariance functions
- Hyperparameters are estimated by maximizing the approximate marginal likelihood $p(\boldsymbol{Y})$



(a) True $\boldsymbol{s}_0$ (b) $\hat{\boldsymbol{s}}_{IEP}$ (c) $\hat{\boldsymbol{s}}_{EP}$ (d) $\hat{\boldsymbol{s}}_{LR}$ (e) $\hat{\boldsymbol{s}}_{CP}$ (f) $\hat{\boldsymbol{s}}_{G}$



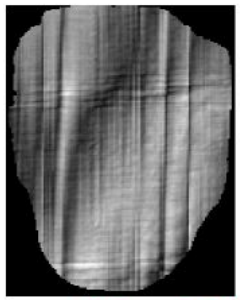
(g) *EP* (h) *LR* (i) *CP* (j) *G*



EEG example: Face vs. scrambled face visual stimuli

- Subjects are presented with pictures of faces and "scrambled" faces
- Data is preprocessed and averaged over trials and subjects
- Number of sources $D = 5124$, number of EEG sensors $N = 128$ and number of time points $T = 161$, i.e. $\boldsymbol{X} \in \mathbb{R}^{5124 \times 161}$ and $\boldsymbol{Y} \in \mathbb{R}^{128 \times 161}$.

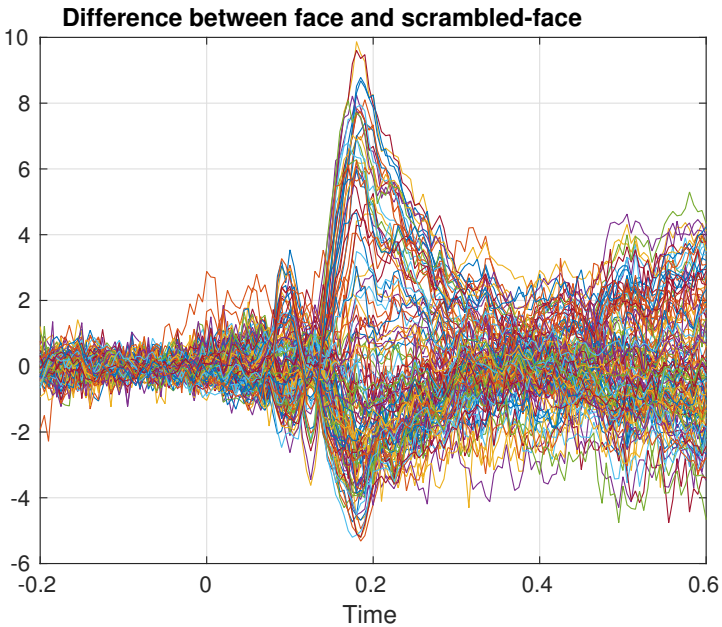
Face



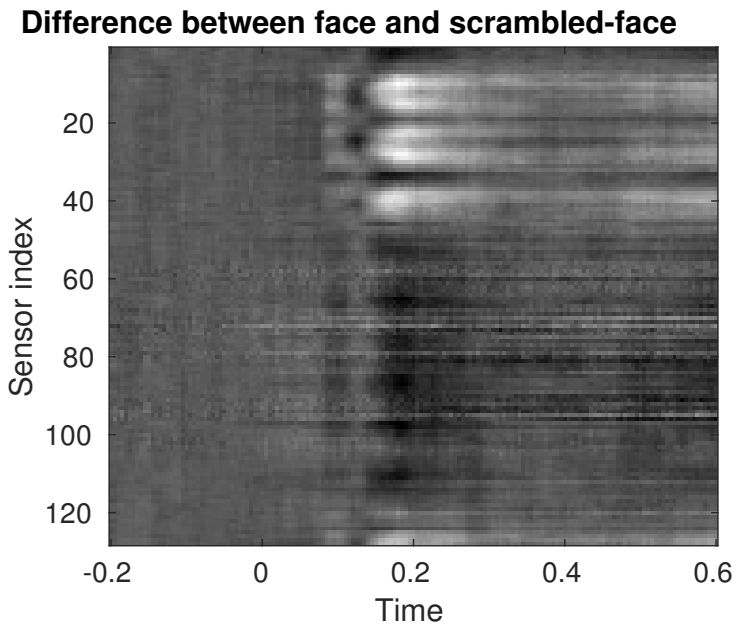
Scrambled Face

EEG example: Face vs. scrambled face visual stimuli

- Subjects are presented with pictures of faces and "scrambled" faces
- Data is preprocessed and averaged over trials and subjects
- Number of sources $D = 5124$, number of EEG sensors $N = 128$ and number of time points $T = 161$, i.e. $\boldsymbol{X} \in \mathbb{R}^{5124 \times 161}$ and $\boldsymbol{Y} \in \mathbb{R}^{128 \times 161}$.



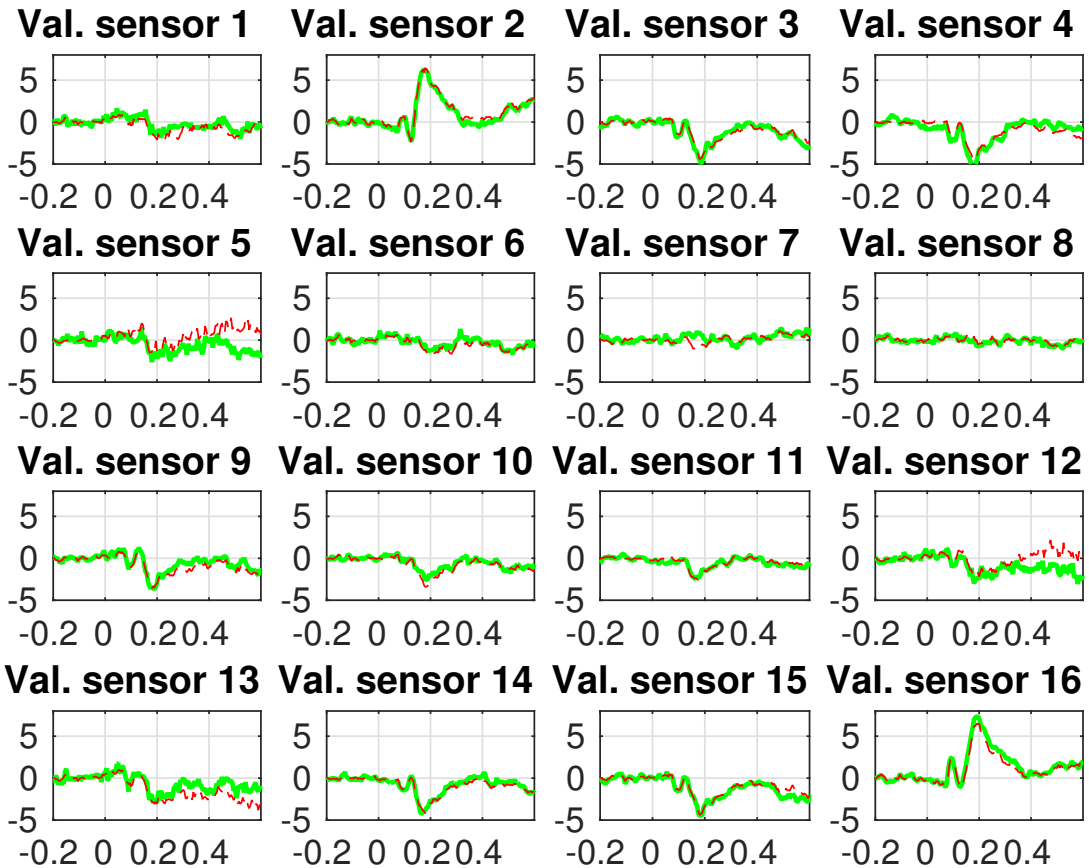
(a)



(b)

EEG example: Face vs. scrambled face visual stimuli

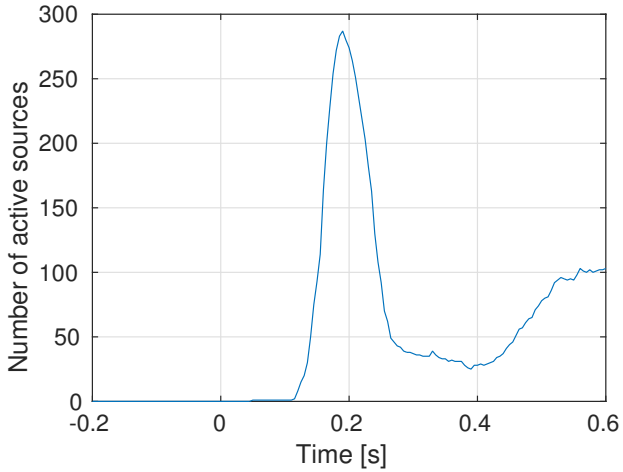
- Common precision approximation
- Sq. exp. kernel for $\mathbf{\Gamma}$, 1 temporal length scale and 1 spatial length scale
- $N_{\text{train}} = 110$ for training and $N_{\text{test}} = 18$ for validation



- Sensors are correlated, so this is only a sanity check at best

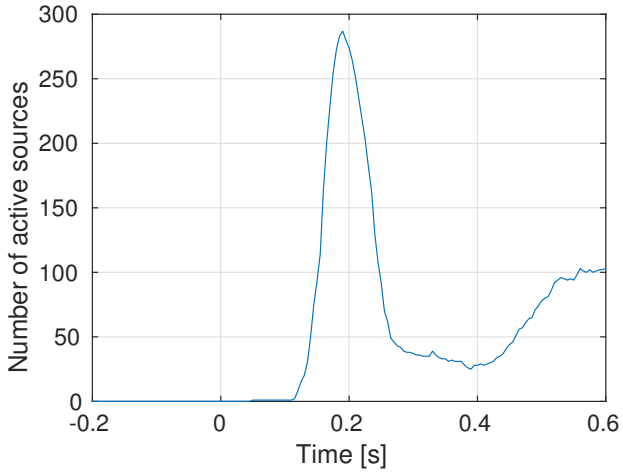
Active sources

- Number of active sources as a fnc. of time, i.e. $\sum_i p(s_{i,t}|\boldsymbol{Y})$ vs. t.

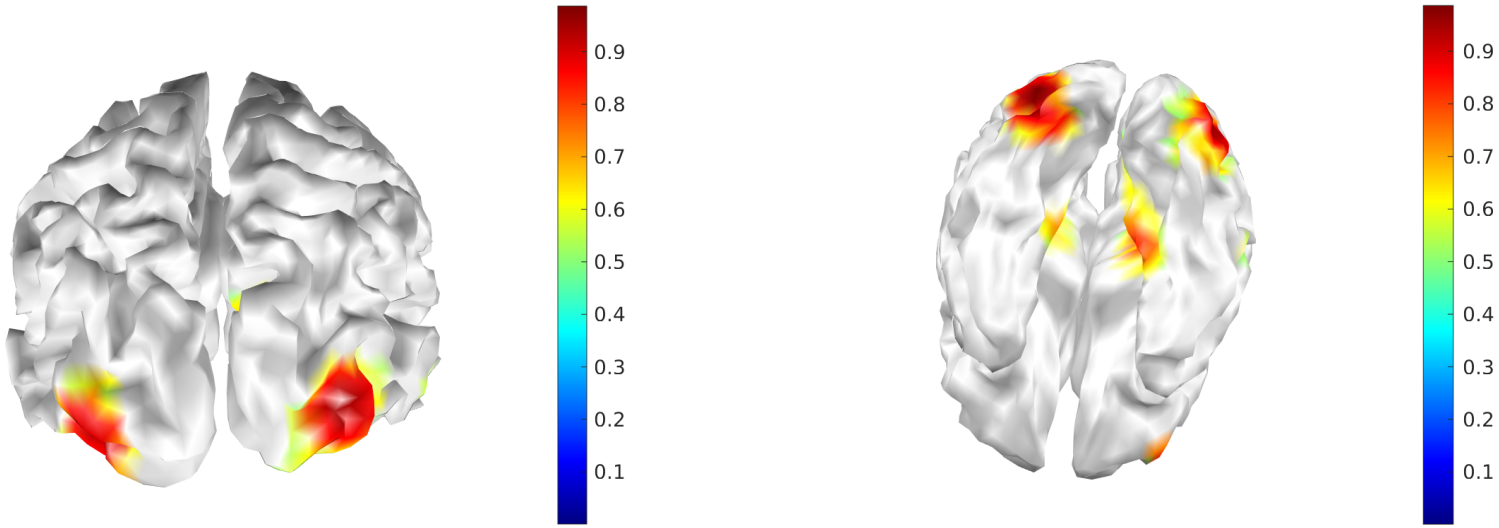


Active sources

- Number of active sources as a fnc. of time, i.e. $\sum_i p(s_{i,t}|\boldsymbol{Y})$ vs. t.



- Active sources at $t = 0.2s$



Thank you for listening!

Any questions?

Low rank approximation for EEG example

Number of variables proportional to

$$D \cdot T = 5124 \cdot 161 \approx 0.8M$$

% of variance explained in $\boldsymbol{\Sigma}_0$	K	Size of $\boldsymbol{U} \in \mathbb{R}^{DT \times K}$
0.95	3961	24.3GB
0.90	2962	18.2GB

Table: Low rank approximation for EEG example