Spatio-temporal spike and slab priors for MMV problems

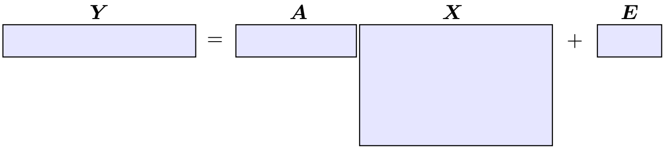
Michael Riis Andersen (Joint work with Ole Winther & Lars Kai Hansen)

SPARS Workshop 2015, Cambridge, UK

DTU Compute
Department of Applied Mathematics and Computer Science

The multiple measurement vector problem

The multiple measurement vector (MMV) is given by



where

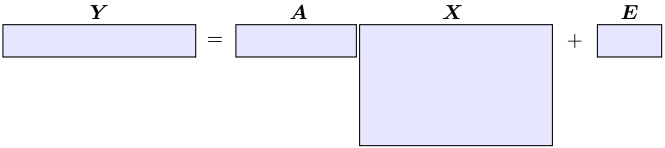
- $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \dots & \boldsymbol{y}_T \end{bmatrix} \in \mathbb{R}^{N \times T}$ is an observation matrix
- $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ is known forward model
- $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_T \end{bmatrix} \in \mathbb{R}^{D \times T}$ is the desired solution

We assume

- Underdetermined regime $N \ll D$

The multiple measurement vector problem

The multiple measurement vector (MMV) is given by

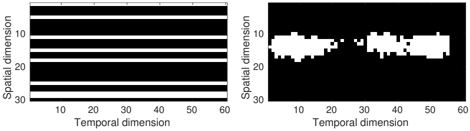


where

- $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \dots & \boldsymbol{y}_T \end{bmatrix} \in \mathbb{R}^{N \times T}$ is an observation matrix
- $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ is known forward model
- $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_T \end{bmatrix} \in \mathbb{R}^{D \times T}$ is the desired solution

We assume

- Underdetermined regime $N \ll D$
- Sparsity pattern of $\boldsymbol{X}$ is assumed to be spatial-temporally correlated



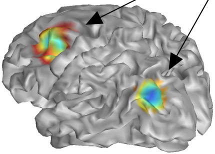
(a) **Joint sparsity** (b) **Spatiotemporal**

Motivation: 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 model $\boldsymbol{A} \in \mathbb{R}^{N \times D}$.

Approximately sparse and spatiotemporal structure

$= \mathrm{A} \cdot$



$+ \boldsymbol{E}$

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

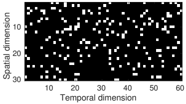
Roadmap

- Our goal is to formulate a probabilistic model for incorporating the prior knowledge of the support and apply Bayes rule for inference

- The so-called spike and slab prior is often used for imposing sparsity in a probabilistic setting

  T.J. Mitchell & J.J. Beauchamp: ***Bayesian Variable Selection in Linear Regression*, *Journal of the American Statistical Association 1988***

(independent) spike and slab prior



Roadmap

- Our goal is to formulate a probabilistic model for incorporating the prior knowledge of the support and apply Bayes rule for inference

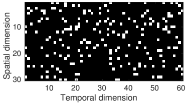
- The so-called spike and slab prior is often used for imposing sparsity in a probabilistic setting

  T.J. Mitchell & J.J. Beauchamp: ***Bayesian Variable Selection in Linear Regression, Journal of the American Statistical Association 1988***

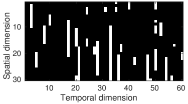
- The *Structured* spike and slab prior is a spatial extension

  M. R. Andersen, O. Winther & L. K. Hansen: ***Bayesian Inference for Structured Spike and Slab Priors, NIPS 2014***

(independent) spike and slab prior



structured spike and slab prior



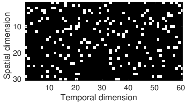
Roadmap

- Our goal is to formulate a probabilistic model for incorporating the prior knowledge of the support and apply Bayes rule for inference

- The so-called spike and slab prior is often used for imposing sparsity in a probabilistic setting

  T.J. Mitchell & J.J. Beauchamp: ***Bayesian Variable Selection in Linear Regression, Journal of the American Statistical Association 1988***

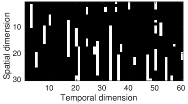
(independent) spike and slab prior



- The *Structured* spike and slab prior is a spatial extension

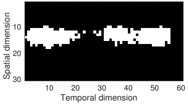
  M. R. Andersen, O. Winther & L. K. Hansen: ***Bayesian Inference for Structured Spike and Slab Priors, NIPS 2014***

structured spike and slab prior



- We generalize it to model both spatial and temporal smoothness

spatio-temporal spike and slab prior



Spike and slab priors for promoting sparsity

- First we consider the (independent) spike and slab prior

- 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

- First we consider the (independent) spike and slab prior

- 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,

$$p(s_i) = \mathsf{Ber}\left(p_0\right)$$
$$p(x_i|s_i) = (1 - s_i)\delta(x_i) + s_i\mathcal{N}\left(x_i \middle| 0, \tau\right)$$

Spike and slab priors for promoting sparsity

- First we consider the (independent) spike and slab prior

- Assume x_i is composed of two variables

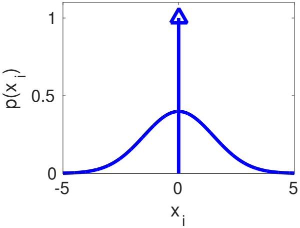
$$x_i = s_i \cdot c_i, \quad s_i \in \{0, 1\}, \quad c_i \in \mathbb{R}$$

- In terms of probability distributions,

$$p(s_i) = \text{Ber}\,(p_0)$$
$$p(x_i|s_i) = (1 - s_i)\delta(x_i) + s_i\mathcal{N}\left(x_i\middle|0, \tau\right)$$

- Marginalizing out s_i

$$p(x_i) = (1 - p_0)\underbrace{\delta(x_i)}_{\text{Spike}} + p_0\underbrace{\mathcal{N}\left(x_i\middle|0, \tau\right)}_{\text{Slab}}$$



The structured spike and slab prior

- Consider first the single measurement vector problem $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$

- 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

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

The structured spike and slab prior

- Consider first the single measurement vector problem $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$
- 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 $\phi : \mathbb{R} \rightarrow (0, 1)$

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

- We choose ϕ to be the standard normal CDF.
- $\boldsymbol{\Sigma}$ now determines the correlation structure of the support of $\boldsymbol{x}$ and

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

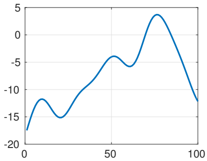
Sampling from the structured spike and slab prior

- We can understand how the prior works by sampling from it
- Consider a signal $\boldsymbol{x} \in \mathbb{R}^{100}$ with support $\boldsymbol{s} \in \{0,1\}^{100}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$.

Sampling from the structured spike and slab prior

- We can understand how the prior works by sampling from it
- Consider a signal $\boldsymbol{x} \in \mathbb{R}^{100}$ with support $\boldsymbol{s} \in \{0, 1\}^{100}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$.

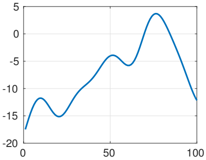
$\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$



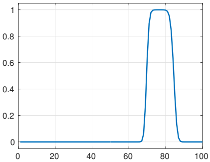
Sampling from the structured spike and slab prior

- We can understand how the prior works by sampling from it
- Consider a signal $\boldsymbol{x} \in \mathbb{R}^{100}$ with support $\boldsymbol{s} \in \{0,1\}^{100}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$.

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

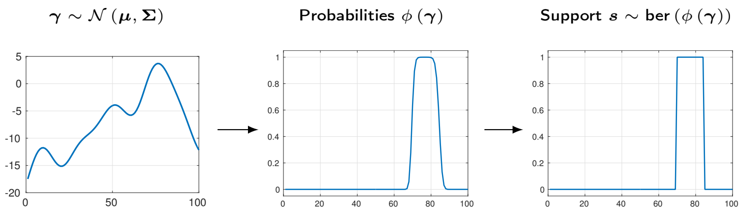


Probabilities $\phi(\boldsymbol{\gamma})$



Sampling from the structured spike and slab prior

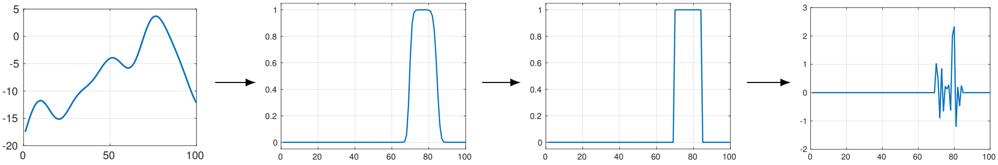
- We can understand how the prior works by sampling from it
- Consider a signal $\boldsymbol{x} \in \mathbb{R}^{100}$ with support $\boldsymbol{s} \in \{0,1\}^{100}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$.



Sampling from the structured spike and slab prior

- We can understand how the prior works by sampling from it
- Consider a signal $\boldsymbol{x} \in \mathbb{R}^{100}$ with support $\boldsymbol{s} \in \{0,1\}^{100}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$.

$\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | Probabilities $\phi(\boldsymbol{\gamma})$ | Support $\boldsymbol{s} \sim \text{ber}(\phi(\boldsymbol{\gamma}))$ | Signal $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{s})$



The spatio-temporal extension

- We model the slowly changing sparsity pattern by imposing temporal smoothness on γ_t using a first order Markov process

$$p\left(\boldsymbol{\gamma}_t \middle| \boldsymbol{\gamma}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \left(1-\alpha\right)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right), \tag{1}$$

- α and β control the temporal correlation and the "innovation" of the process.

The spatio-temporal extension

- We model the slowly changing sparsity pattern by imposing temporal smoothness on γ_t using a first order Markov process

$$p\left(\boldsymbol{\gamma}_t \middle| \boldsymbol{\gamma}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \left(1-\alpha\right)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right), \tag{1}$$

- α and β control the temporal correlation and the "innovation" of the process.

- Assuming a prior distribution on $p(\boldsymbol{\gamma}_1) = \mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$, then the marginal distribution of $\boldsymbol{\gamma}_2$ becomes

$$p(\boldsymbol{\gamma}_2) = \int p\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\gamma}_1\right) p(\boldsymbol{\gamma}_1) \mathrm{d}\boldsymbol{\gamma}_1 = \mathcal{N}\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\mu}_0, \left(\alpha^2 + \beta\right)\boldsymbol{\Sigma}_0\right).$$

The spatio-temporal extension

- We model the slowly changing sparsity pattern by imposing temporal smoothness on γ_t using a first order Markov process

$$p\left(\boldsymbol{\gamma}_t \middle| \boldsymbol{\gamma}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \left(1-\alpha\right)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right), \quad (1)$$

- α and β control the temporal correlation and the "innovation" of the process.

- Assuming a prior distribution on $p(\boldsymbol{\gamma}_1) = \mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$, then the marginal distribution of $\boldsymbol{\gamma}_2$ becomes

$$p(\boldsymbol{\gamma}_2) = \int p\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\gamma}_1\right) p(\boldsymbol{\gamma}_1) \mathrm{d}\boldsymbol{\gamma}_1 = \mathcal{N}\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\mu}_0, \left(\alpha^2+\beta\right)\boldsymbol{\Sigma}_0\right).$$

- Induction argument: if α and β satisfy $\alpha^2 + \beta = 1$, then the marginal distribution of $\boldsymbol{\gamma}_t$ is $p(\boldsymbol{\gamma}_t) = \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ for all t.

The spatio-temporal extension

- We model the slowly changing sparsity pattern by imposing temporal smoothness on $\boldsymbol{\gamma}_t$ using a first order Markov process

$$p\left(\boldsymbol{\gamma}_t \middle| \boldsymbol{\gamma}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \left(1-\alpha\right)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right), \qquad (1)$$

- α and β control the temporal correlation and the "innovation" of the process.

- Assuming a prior distribution on $p(\boldsymbol{\gamma}_1) = \mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$, then the marginal distribution of $\boldsymbol{\gamma}_2$ becomes

$$p(\boldsymbol{\gamma}_2) = \int p\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\gamma}_1\right) p(\boldsymbol{\gamma}_1) \mathrm{d}\boldsymbol{\gamma}_1 = \mathcal{N}\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\mu}_0, \left(\alpha^2 + \beta\right)\boldsymbol{\Sigma}_0\right).$$

- Induction argument: if α and β satisfy $\alpha^2 + \beta = 1$, then the marginal distribution of $\boldsymbol{\gamma}_t$ is $p(\boldsymbol{\gamma}_t) = \mathcal{N}\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ for all t.

- Generalization of joint sparsity and time-independent sparsity
  - Joint sparsity for $\alpha = 1$ and $\beta = 0$
  - Time-independent for $\alpha = 0$ and $\beta = 1$

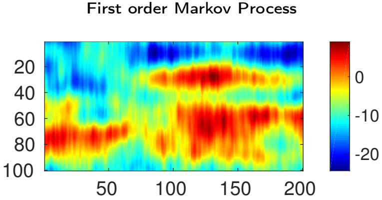
Sampling from the spatio-temporal prior

- Consider a signal $\boldsymbol{X} \in \mathbb{R}^{100\times 200}$ with support $\S \in \{0,1\}^{100\times 200}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma}_1 \sim \mathcal{N}\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$ and temporal correlation $\alpha = 0.99$.

Sampling from the spatio-temporal prior

- Consider a signal $\boldsymbol{X} \in \mathbb{R}^{100\times 200}$ with support $\S \in \{0,1\}^{100\times 200}$

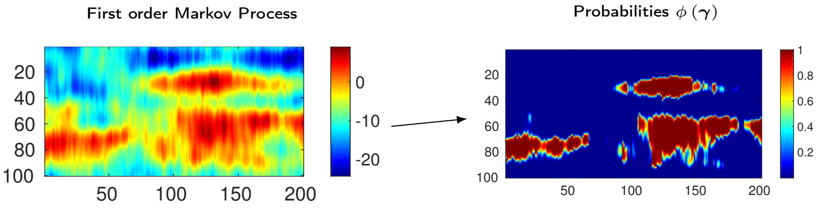
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma}_1 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$ and temporal correlation $\alpha = 0.99$.



Sampling from the spatio-temporal prior

- Consider a signal $\boldsymbol{X} \in \mathbb{R}^{100\times 200}$ with support $\S \in \{0,1\}^{100\times 200}$

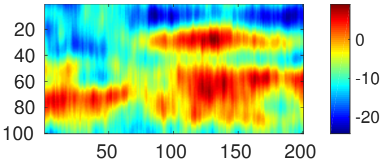
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$ and temporal correlation $\alpha = 0.99$.



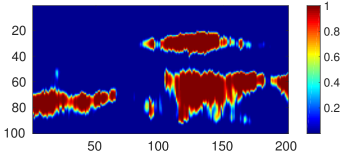
Sampling from the spatio-temporal prior

- Consider a signal $\boldsymbol{X} \in \mathbb{R}^{100\times 200}$ with support $\S \in \{0,1\}^{100\times 200}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$ and temporal correlation $\alpha = 0.99$.

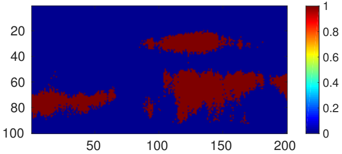
First order Markov Process



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



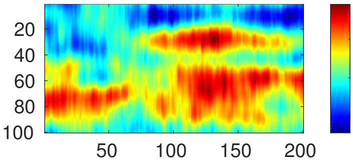
Support $\boldsymbol{s} \sim$ ber $\left(\phi\left(\boldsymbol{\gamma}\right)\right)$



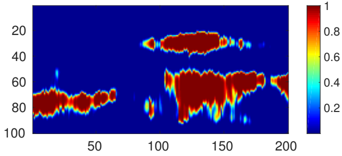
Sampling from the spatio-temporal prior

- Consider a signal $\boldsymbol{X} \in \mathbb{R}^{100\times 200}$ with support $\S \in \{0,1\}^{100\times 200}$
- We choose a smooth kernel for $\boldsymbol{\Sigma}$ in $\boldsymbol{\gamma}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, e.g. $\boldsymbol{\Sigma}_{ij} = K_1 \exp\left(-\frac{D_{ij}^2}{2K_2}\right)$ and temporal correlation $\alpha = 0.99$.

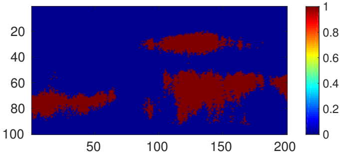
First order Markov Process



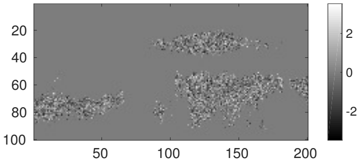
Probabilities $\phi\left(\boldsymbol{\gamma}\right)$



Support $\boldsymbol{s} \sim$ ber $\left(\phi\left(\boldsymbol{\gamma}\right)\right)$



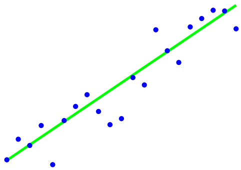
Signal $\boldsymbol{X} \sim p\left(\boldsymbol{X}|\boldsymbol{S}\right)$



Inference using the spatio-temporal spike and slab prior

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

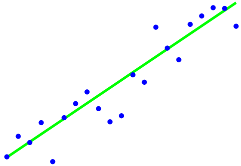
- Assuming isotropic Gaussian noise leads to a posterior distribution of the form



- Joint distribution

$$p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma}) = \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{y}_t \middle| \boldsymbol{A}\boldsymbol{x}_t, \sigma_0^2 \boldsymbol{I}\right) \prod_{t=1}^{T} \prod_{i=1}^{D} \left[(1 - s_{i,t})\delta(x_{i,t}) + s_{i,t}\mathcal{N}\left(x_{i,t} \middle| 0, \tau_0\right)\right]$$
$$\prod_{t=1}^{T} \prod_{i=1}^{D} \mathrm{Ber}\left(s_{i,t} \middle| \phi\left(\gamma_{i,t}\right)\right) \mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \prod_{t=2}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| (1-\alpha)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right)$$

Inference using the spatio-temporal spike and slab prior

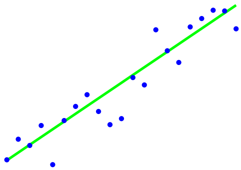


- Recall the model $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}$
- Assuming isotropic Gaussian noise leads to a posterior distribution of the form
- Posterior of interest

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

$$\prod_{t=1}^{T} \prod_{i=1}^{D} \mathrm{Ber}\left(s_{i,t} | \phi\left(\gamma_{i,t}\right)\right) \mathcal{N}\left(\boldsymbol{\gamma}_1 | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \prod_{t=2}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t | (1-\alpha)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right)$$

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
- Posterior of interest

$$p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma} | \boldsymbol{Y}) = \frac{1}{P(\boldsymbol{Y})} \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{y}_t \middle| \boldsymbol{A}\boldsymbol{x}_t, \sigma_0^2 \boldsymbol{I}\right) \prod_{t=1}^{T} \prod_{i=1}^{D} \left[(1 - s_{i,t})\delta(x_{i,t}) + s_{i,t}\mathcal{N}\left(x_{i,t} \middle| 0, \tau_0\right)\right]$$
$$\prod_{t=1}^{T} \prod_{i=1}^{D} \text{Ber}\left(s_{i,t} \middle| \phi\left(\gamma_{i,t}\right)\right) \mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \prod_{t=2}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| (1-\alpha)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right)$$

- Exact inference is intractable since $P(\boldsymbol{Y})$ is intractable due to the product over mixtures
- We resort Expectation Propagation for approximate inference

Approximate inference using EP

- Expectation propagation: deterministic framework for approximating probability distributions

  T. Minka: ***Expectation Propagation for Approximate Bayesian Inference, 2011***

- The resulting EP approximation has the form

$$Q(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}) = \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{x}_t \middle| \tilde{\boldsymbol{m}}_t, \tilde{\boldsymbol{V}}_t\right) \prod_{t=1}^{T} \prod_{i=1}^{D} \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\tilde{\gamma}_{i,t}\right)\right) \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t\right). \quad (2)$$

- We obtain posterior mean values AND the associated uncertainties

- Dense posterior covariance matrices, i.e. $\tilde{\boldsymbol{V}}_t$ & $\tilde{\boldsymbol{\Sigma}}_t$ to capture posterior correlations

- The computational bottleneck of the algorithm is the update of the covariance matrices $\tilde{\boldsymbol{V}}_t$ and $\tilde{\boldsymbol{\Sigma}}_t$, which scale $\mathcal{O}\left(N^2 D\right)$ and $\mathcal{O}\left(TD^3\right)$, respectively.

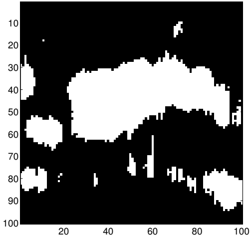
- Since SPARS2015 deadline: We introduced further approximations to get computational complexity down to either $\mathcal{O}\left(KTD\right)$ or $\mathcal{O}\left(TD^2\right)$ for more flexible temporal model

Evaluation of proposed method

- We compared the performance of the proposed method to competing methods based on synthetic data

- We generate linear observations of $\boldsymbol{X}$ using the model $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}$ for

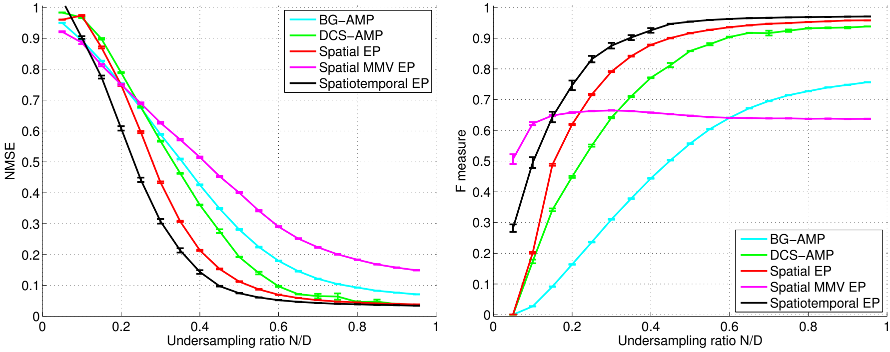
  1. $D = T = 100$ and SNR fixed to 10dB
  2. Undersampling ratio $\frac{N}{D} \in \{0.05, 0.10, \ldots, 0.95\}$
  3. Gaussian i.i.d forward model $A_{ij} \sim \mathcal{N}(0, 1/N)$
  4. Gaussian i.i.d coefficients $x_i | s_i = 1 \sim \mathcal{N}(0, 1)$
  5. Average over 100 realizations of coefficients and noise



- Quantification of performance using NMSE and F-measure

$$\text{NMSE} = \frac{\sum_{i,t} \left(X_{i,t} - \hat{X}_{i,t} \right)^2}{\sum_{i,t} X_{i,t}^2} \qquad F = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{3}$$

Results



(c) (d)

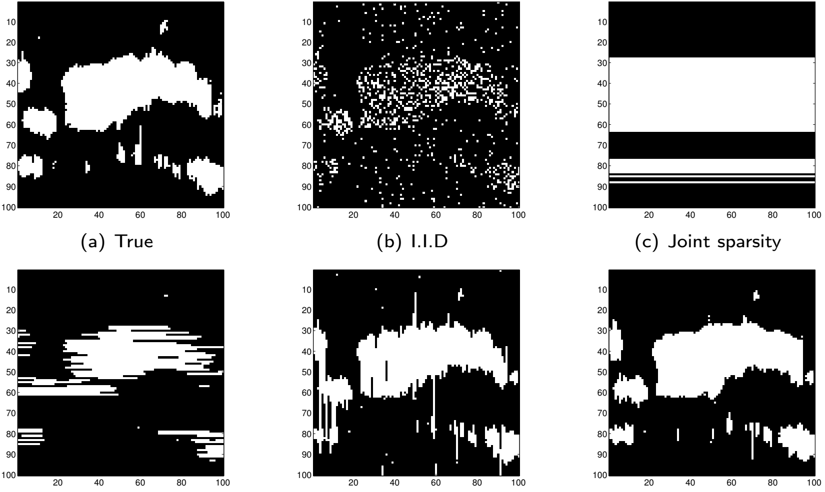
Figure: Normalized mean square error & F-measure error as a function of undersampling ratio. The data are generated from $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{E}$, where $D = 100, T = 100$ and SNR $= 10dB$. The entries in $\boldsymbol{A}$ are Gaussian i.i.d, i.e. $A_{i,j} \sim \mathcal{N}(0, 1/N)$. The results are averaged over 100 realizations.

J. P. Vila & P. Schniter: *Expectation-Maximization Gaussian-Mixture Approximate Message Passing, 2013*

J. Ziniel & P. Schniter: *Dynamic Compressive Sensing of Time-Varying Signals via Approximate Message Passing, 2013*

M. R. Andersen, O. Winther & L. K. Hansen: *Bayesian Inference for Structured Spike and Slab Priors, 2014*

Reconstructions for $N/D = 0.4$



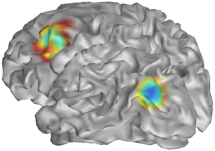
(a) True (b) I.I.D (c) Joint sparsity

(d) Temporal (e) Spatial (f) Spatiotemporal

Figure: True and reconstructed support. The undersampling ratio is $N/D = 0.4$ and $D = 100, T = 100$ and $SNR = 10$dB. a) True support, b) BG-AMP (NMSE = 0.805, F = 0.450), c) Spatial MMV EP (NMSE = 0.833, F = 0.663), d) Spatial EP (NMSE = 0.658, F = 0.902), e) DCS-AMP (NMSE = 0.777, F = 0.763), f) Spatio-temporal EP (NMSE = 0.618, F = 0.935).

Preliminary results for EEG source localization

$= \mathrm{A} \cdot$



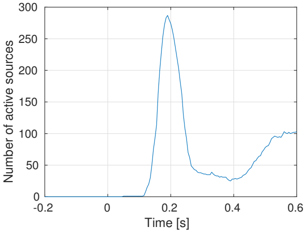
$+ \boldsymbol{E}$

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

- 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 $\boldsymbol{Y} \in \mathbb{R}^{128 \times 161}$.
- Subjects are presented with pictures of faces and "scrambled" faces
- Data is preprocessed and averaged over trials and subjects (difference contrast)
- Results obtained with EP and slightly more flexible temporal model

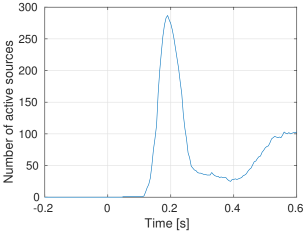
Active sources as a function of time

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

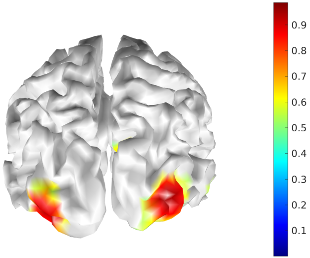


Active sources as a function of time

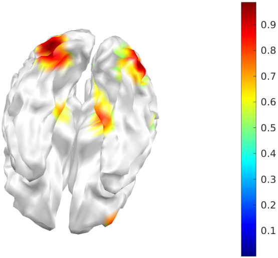
- Number of active sources as a function of time, i.e. $\sum_i p(s_{i,t}|\boldsymbol{Y})$ vs. t.



- Active sources at $t = 0.2s$



(c)



(d)

Conclusion

- We proposed a probabilistic model to incorporate prior knowledge of spatial and temporal smoothness of the support for MMV problems
- We derived an Expectation propagation based algorithm for approximate Bayesian inference
- We demonstrated the performance of our model based on numerical experiments with synthetic and real data

Thank you for listening!

Any questions?

Approximate inference using Expectation Propagation (EP)

- By defining $\boldsymbol{\theta} \triangleq \left(\{\boldsymbol{x}_t\}_t, \{\boldsymbol{s}_t\}_t, \{\boldsymbol{\gamma}_t\}_t\right)$, the desired posterior can be written as

$$p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma} \big| \boldsymbol{Y}) \propto \prod_a f_a\left(\boldsymbol{\theta}_{\boldsymbol{a}}\right), \quad \boldsymbol{\theta}_{\boldsymbol{a}} \subset \boldsymbol{\theta} \tag{4}$$

- EP approximates each factor with a term from the exponential family

$$Q(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma}) = \prod_a \tilde{f}_a\left(\boldsymbol{\theta}_{\boldsymbol{a}}\right) \tag{5}$$

Approximate inference using Expectation Propagation (EP)

- By defining $\boldsymbol{\theta} \triangleq \left(\{\boldsymbol{x}_t\}_t, \{\boldsymbol{s}_t\}_t, \{\boldsymbol{\gamma}_t\}_t\right)$, the desired posterior can be written as

$$p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma} \big| \boldsymbol{Y}) \propto \prod_a f_a\left(\boldsymbol{\theta}_{\boldsymbol{a}}\right), \quad \boldsymbol{\theta}_{\boldsymbol{a}} \subset \boldsymbol{\theta} \tag{4}$$

- EP approximates each factor with a term from the exponential family

$$Q(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma}) = \prod_a \tilde{f}_a\left(\boldsymbol{\theta}_{\boldsymbol{a}}\right) \tag{5}$$

- EP updates each term $\tilde{f}_b$ as follows:
  1. Obtain *cavity distribution* by removing $\tilde{f}_b$ from *global* approximation

$$Q_{-b}\left(\boldsymbol{\theta}\right) \propto \frac{Q(\boldsymbol{\theta})}{\tilde{f}_b\left(\boldsymbol{\theta}_b\right)} \tag{6}$$

Approximate inference using Expectation Propagation (EP)

- By defining $\boldsymbol{\theta} \triangleq \left(\{\boldsymbol{x}_t\}_t, \{\boldsymbol{s}_t\}_t, \{\boldsymbol{\gamma}_t\}_t\right)$, the desired posterior can be written as

$$p(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma} \big| \boldsymbol{Y}) \propto \prod_a f_a\left(\boldsymbol{\theta}_{\boldsymbol{a}}\right), \quad \boldsymbol{\theta}_{\boldsymbol{a}} \subset \boldsymbol{\theta} \tag{4}$$

- EP approximates each factor with a term from the exponential family

$$Q(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Gamma}) = \prod_a \tilde{f}_a\left(\boldsymbol{\theta}_{\boldsymbol{a}}\right) \tag{5}$$

- EP updates each term $\tilde{f}_b$ as follows:
  1. Obtain *cavity distribution* by removing $\tilde{f}_b$ from *global* approximation

$$Q_{-b}\left(\boldsymbol{\theta}\right) \propto \frac{Q(\boldsymbol{\theta})}{\tilde{f}_b\left(\boldsymbol{\theta}_b\right)} \tag{6}$$

  2. Combine with exact term $f_b\left(\boldsymbol{\theta}_b\right) Q_{-b}\left(\boldsymbol{\theta}\right)$ and minimize KL divergence

$$\underset{\tilde{f}_b}{\arg\min}\, \mathrm{KL}\left(f_b\left(\boldsymbol{\theta}_b\right) Q_{-b}\left(\boldsymbol{\theta}\right) || \tilde{f}_b\left(\boldsymbol{\theta}_b\right) Q_{-b}\left(\boldsymbol{\theta}\right)\right) \tag{7}$$