Downloaded from orbit.dtu.dk on: Apr 26, 2024

DTU Library

Probabilistic models for structured sparsity

Andersen, Michael Riis

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Andersen, M. R. (2017). *Probabilistic models for structured sparsity*. Technical University of Denmark. DTU Compute PHD-2017 Vol. 452

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Probabilistic models for structured sparsity

Michael Riis Andersen

Kongens Lyngby 2017
PhD-2017-452

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk
PhD-2017-452

Summary (English)

Sparsity has become an increasingly popular choice of regularization in machine learning and statistics. The sparsity assumption for a matrix $\boldsymbol{X}$ means that most of the entries in $\boldsymbol{X}$ are equal to exactly zero. Structured sparsity is generalization of sparsity and assumes that the set of locations of the non-zero coefficients in $\boldsymbol{X}$ contains structure that can be exploited. This thesis deals with probabilistic models for structured sparsity for regularization of ill-posed problems. The aim of the thesis is two-fold; to construct sparsity promoting prior distributions for structured sparsity and to derive efficient inference algorithms for these distributions. The work explores a class of models that uses Gaussian processes (Rasmussen and Williams, 2006) as a latent representation of the structure of sparsity patterns. This representation allows prior knowledge of the structure of the sparsity patterns to be encoded using generic covariance functions through the Gaussian process. This thesis focuses on two specific instances of ill-posed problems: linear inverse problems and time-varying covariance estimation.

The first part of the thesis deals with probabilistic methods for finding structured sparse solutions to linear inverse problems. In this part, the sparsity promoting prior known as the spike-and-slab prior (Mitchell and Beauchamp, 1988) is generalized to the structured sparsity setting. An expectation propagation algorithm is derived for approximate posterior inference. The proposed model and the associated inference algorithm are studied and evaluated using a set of numerical experiments, which include phase transition experiments, compressed sensing, phoneme classification and electroencephalography (EEG) source localization.

The second part of the thesis deals with the problem of time-varying covari-

ance estimation. A hierarchical model for a set of non-stationary time series with time-varying covariance matrices is proposed. The model is tailored to address the problem of dynamic functional connectivity in neuroimaging and it assumes that the instantaneous covariance matrix of each time series is decomposed into a non-negative linear combination of elements from a dictionary of shared covariance matrix components. A variational Bayes algorithm is derived for approximate posterior inference. The proposed model is validated using a functional magnetic resonance imaging (fMRI) dataset.

Resumé (Danish)

Sparsity er blevet mere og mere populært i forbindelse med regularisering i maskinlæring og statistik. Antagelsen om sparsity for en matrix $\boldsymbol{X}$ betyder, at størstedelen af elementerne i $\boldsymbol{X}$ er lig med nul. Struktureret sparsity er en generalisering af sparsity. Her antages det, at positionerne for de koefficienter, der er forskellig fra nul, ikke er tilfældigt fordelt, men at de har en struktur, som kan udnyttes. Denne afhandling omhandler sandsynlighedsteoretiske modeller for struktureret sparsity i forbindelse med regularisering af underbestemte matematiske problemer. Formålet med afhandlingen er to-delt. Først og fremmest er formålet at konstruere sandsynlighedsfordelinger for signaler med strukturerede sparsity-mønstre. Dernæst er formålet at udlede en effektiv inferens-algoritme for disse fordelinger. Afhandlingen undersøger en klasse af modeller, der bruger Gaussiske processer (Rasmussen and Williams, 2006) som en latent repræsentation af strukturen for sparsity-mønstrene. Denne repræsentation gør det muligt at indkode forhåndsviden om strukturen af sparsity-mønstret i modellen via generiske kovariansfunktioner. Denne afhandling fokuserer på to forskellige typer af underbestemte problemer: linear inverse problemer og estimation af tidsvariende kovariansmatricer.

Den første del af afhandlingen omhandler sandsynlighedsbaserede metoder til at finde løsninger med strukturede sparsity-mønstre for lineære underbestemte problemer. I denne del generaliseres den såkaldte spike-and-slab fordeling (Mitchell and Beauchamp, 1988) til vektorer og matricer med strukturede sparsity-mønstre. Via metoden *expectation propagation* udledes en algoritme for approksimativ posterior inferens. Modellen og den tilhørende inferens-algoritme undersøges og evalueres via en række numeriske eksperimenter, blandt andet faseovergangseksperimenter, fonem klassifikation og elektroencefalografi (EEG)

kildelokalisering.

Den anden del af afhandlingen omhandler problemstillingen med at estimere tidsvarierende kovariansstrukturer. Denne del beskriver en hierarkisk model for ikke-stationære tidsrækker med tidsvarierende kovariansmatricer . Modellen er målrettet problemstillingen om *dynamic functional connectivity*. Modellen antager, at den instantane kovariansmatrix for hver tidsrække er sammensat af en linear kombination af kovariansmatrix elementer. En variationel inferens-algoritme er udledt for approksimativ posterior inferens. Modellen og algoritmen er valideret på et datasæt fra et studie baseret på funktionel magnetisk resonans-billeddannelse (fMRI).

Preface

This thesis was prepared at the Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis consists of a summary report and a collection of four papers: two conference papers, a paper submitted to a journal and a paper submitted to a conference. The work has been carried out between March 2014 and April 2017

Lyngby, 31-January-2017

Michael Riis Andersen

List of Publications

Papers included in this thesis

Peer Reviewed

A Andersen, M. R., Winther, O., and Hansen, L. K. (2014), 'Bayesian inference for structured spike and slab priors'. Advances in Neural Information Processing Systems (NIPS) 2014, 9 pages

B Andersen, M. R., Winther, O. and Hansen, L. K. (2015), 'Spatio-temporal spike and slab priors for MMV problems'. Signal Processing with Adaptive Sparse Structured Representations (SPARS) 2015, 6 pages

Submitted

C Andersen, M. R., Winther, O. and Hansen, L. K. (2015), 'Bayesian inference for spatio-temporal spike and slab priors'. *Re-submitted to the Journal of Machine Learning Research (JMLR) (8/3-2017), 57 pages*

D Andersen, M. R., Hansen, L. K., Winther, O., Koyejo, S. and Poldrack, R. (2017), 'A hierarchical model for time-varying functional connectivity'. *Submitted to the International Conference on Machine Learning (ICML) (24/2-2017), 10 pages*

Papers not included in this thesis

Peer Reviewed

Andersen, R. S., Eliasen, A. U., Pedersen, N., Andersen, M. R., Hansen, S. T. , Hansen, L. K. (2017), 'EEG source imaging assists decoding in a face recognition task'. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2017, 5 pages

Abstract

Andersen, M. R., Koyejo, S. Poldrack, R. (2016), 'Model-based dynamic resting state functional connectivity'. Organization for Human Brain Mapping (OHBM) 2016, 2 pages

Acknowledgements

There are many people that I would like to thank for their help and support during my PhD studies. First of all, I would like to thank my supervisor Professor Lars Kai Hansen. I am deeply grateful for all his help, support and guidance as well as his optimism and enthusiasm. Similarly, I would also like to thank my co-supervisor Professor Ole Winther for all his guidance and support.

Furthermore, I would like to thank Professor Aki Vehtari for many insightful discussions and for acting as an unofficial supervisor for part of my PhD.

During my studies I spent five months at Stanford University in California, US. I would like to thank Professor Russell Poldrack for giving me the opportunity to visit his lab. I would also like to thank Assistant Professor Sanmi Koyejo for all the many discussions and for great collaboration both during and after my visit.

I would like to thank all my fellow PhD students in the Section for Cognitive Systems for a lot of good times both inside and outside the lab. Additionally, I want to thank Rasmus Bonnevie for proofreading this thesis.

Last but not least, I would also like to express my deepest gratitude to Rikke Bech Espersen for her unlimited support and patience and for always believing in me.

Nomenclature

Abbreviations

CCD	Composite Central Design
CDF	Cumulative distribution function
CT	Computer tomography
EEG	Electroencephalography
EP	Expectation propagation
fMRI	Functional magnetic resonance imaging
GP	Gaussian process
IID	Independent and identically distributed
KL	Kullback-Leibler
LASSO	Least absolute shrinkage and selection operator
MAP	Maximum a posteriori
MSE	Mean square error
MEG	Magnetoencephalography
ML	Maximum likelihood
MMV	Multiple measurement vectors
MRI	Magnetic resonance imaging
VB	Variational Bayes

Notation and Symbols

$\mathbb{N}$	Natural numbers, i.e. $\mathbb{N} = \{1, 2, 3, \ldots\}$
$\mathbb{R}$	The real line
$\boldsymbol{x}$	Column vector, i.e. $\boldsymbol{x} \in \mathbb{R}^{D \times 1}$
x_i	The i'th entry of vector $\boldsymbol{x}$
$\boldsymbol{X}$	Matrix, i.e. $\boldsymbol{x} \in \mathbb{R}^{D \times T}$
$X_{i,j}$	Entry in the i'th row and the j'th column of $\boldsymbol{X}$
$\boldsymbol{A}_{n,\cdot}$	n'th row of matrix $\boldsymbol{A}$
$\|\boldsymbol{x}\|$	Norm of $\boldsymbol{x}$
$\|\boldsymbol{x}\|_0$	The 0-'norm' is defined as the number of non-zero entries in $\boldsymbol{x}$.
$\|\cdot\|_{\text{op}}$	Operator norm
$\boldsymbol{I}$	Identity matrix
$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$	Gradient of the function f with respect to $\boldsymbol{\theta}$
$\mathcal{K}(\boldsymbol{A})$	Condition number of matrix $\boldsymbol{A}$
$\mathbb{I}[b]$	Indicator function of the proposition b
$\mathbb{E}_p[\boldsymbol{x}]$	Expectation of $\boldsymbol{x}$ with respect to distribution p
$\mathbb{E}[\boldsymbol{x}\|\boldsymbol{y}]$	Conditional expectation of $\boldsymbol{x}$ with respect to distribution $p(\boldsymbol{x}\|\boldsymbol{y})$
$\mathcal{O}(\cdot)$	Big O notation
$\mathcal{P}$	Space of all probability density functions
$\mathcal{P}_C$	Space of all continuous probability density functions
$\text{Ber}(z\|p_1)$	Bernoulli distribution such that $p(z = 1) = p_1$
$\mathcal{N}(\boldsymbol{y}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ evaluated at $\boldsymbol{y}$.
$[N]$	Set of integers from 1 to N, i.e. $[N] = \{n \| n \in \mathbb{N}, 1 \leq n \leq N\}$
δ	Undersampling-ratio
ρ	Sparsity ratio
$\text{vec}[]$	Vectorization operator
$\boldsymbol{X} \otimes \boldsymbol{Y}$	Kronecker product of $\boldsymbol{X}$ and $\boldsymbol{Y}$
$x \sim f$	x has distribution f
$\text{diag}(x_1, x_2, \ldots, x_D)$	Diagonal matrix with elements $x_1, x_2, \ldots, x_D$

Contents

Summary (English) i

Resumé (Danish) iii

Preface v

List of Publications vii

Acknowledgements ix

Nomenclature xi

1 Introduction 1
- 1.1 Discrete Linear Inverse Problems . . . 3
- 1.2 Time-varying Covariance Estimation . . . 7
- 1.3 Thesis outline . . . 8

2 Bayesian Modeling Preliminaries 9
- 2.1 Bayesian Inference . . . 9
- 2.2 Sparsity Promoting Priors and the Spike-and-slab Distribution . . . 14
- 2.3 Gaussian Processes . . . 18

3 Approximate Inference 21
- 3.1 Variational Bayes . . . 22
- 3.2 Expectation Propagation . . . 24
- 3.3 Approximate Inference for 1D Spike-and-slab Models . . . 29

4 The Structured Spike-and-slab Prior for Linear Models 33
- 4.1 Structured Spike-and-slab Priors . . . 33

4.2 Approximate Inference for Linear Models 37
4.3 Contributions 41

5 Time-varying Covariance Estimation **45**
5.1 A Hierarchical Model for Time-varying Covariance Estimation . . 45
5.2 Contributions 50

6 Discussion and Conclusion **53**

A Bayesian Inference for Structured Spike and Slab Priors **59**

B Spatio-temporal Spike and Slab Priors for Multiple Measurement Vector Problems **71**

C Bayesian Inference for Spatio-temporal Spike-and-slab Priors **79**

D A Hierarchical Model for Time-varying Functional Connectivity **139**

Bibliography **153**

Chapter 1

Introduction

Mathematical modeling plays a crucial role in science and engineering as it allows scientists and engineers to study, understand, and make predictions about the world. A large class of problems within mathematical modeling deals with describing and quantifying the relationship between a set of model parameters $\boldsymbol{x} \in \mathbb{R}^D$ and a set of observations $\boldsymbol{y} \in \mathbb{R}^N$. Many of these problems belong to the so-called *small N, large D* paradigm, where the number of observations N is much smaller than the number of parameters D. These problems can therefore be fundamentally ill-posed in the sense that the solution for $\boldsymbol{x}$ is not unique. In essence, the observations $\boldsymbol{y}$ do not contain enough information to identify the parameters $\boldsymbol{x}$. This situation is exacerbated by the fact that measurements are almost always corrupted by noise. Thus, additional prior information is required to obtain meaningful solutions for these underdetermined problems. The process of using prior information to ensure that the solutions will be constrained to reasonable values is also known as *regularization*.

Regularization can be used for inducing solutions with specific properties, such as minimum energy, smoothness, hierarchical structure, maximum entropy, or sparsity to name a few. From a Bayesian perspective, regularization can be interpreted as imposing certain prior distributions on the parameters of the model. The thesis deals with how to design and apply *structured sparsity* for model regularization in a Bayesian setting.

The concept of sparsity has been successfully applied in many fields, including statistics (Tibshirani, 1994; Vapnik, 1995), machine learning (Lee et al., 2007; Caron and Doucet, 2008), signal processing (Jeffs, 1998; Rao, 1998) and neuroscience (Gorodnitsky et al., 1995; Rish, 2013). The sparsity assumption is equivalent to the assumption that only a subset of the entries in a signal $\boldsymbol{x}$ is non-zero and it can be interpreted as regularization because it limits the complexity of the signal $\boldsymbol{x}$ by constraining the maximum number of non-zero coefficients. Sparsity is sometimes referred to as a manifestation of Occam's razor (Eldar, 2015). That is, when there is more than one solution for a data set, the simplest solution is the best. Not all signals of interest are naturally sparse, but most signals exhibit a sparse or approximately sparse representation when represented in an appropriate (and potentially over-complete) basis (Sallee and Olshausen, 2003; Wright et al., 2010). Furthermore, sparsity can also facilitate interpretation and lead to more efficient representations in terms of computational complexity and memory requirements (Rish and Grabarnik, 2014).

Structured sparsity is a generalization of sparsity (Yuan and Lin, 2006; Huang et al., 2009) and assumes that the set of locations of the non-zero coefficients, i.e. the support of $\boldsymbol{x}$, contains structure, e.g. spatial coherence, block or group structure. Structured sparsity regularization have been shown to yield improved results in many applications, when the appropriate sparsity structure is used. For example, magnetic resonance imaging (MRI) is a technology that enables doctors and neuroscientists to image human tissue in a non-invasive manner. Recently, Chen and Huang (2012) showed that by modeling the tree structure of the sparsity pattern of the wavelet decomposition of an MRI image $\boldsymbol{x}$, the minimum number of required observations N for imaging could be decreased without loss of quality leading to faster scan times for patients. Other examples of sparsity structures include joint sparsity (Cotter et al., 2005; Ziniel and Schniter, 2013b), group and graph structured sparsity (Jacob et al., 2009) and cluster structured sparsity (Yu et al., 2012).

The aim of the work contained in this thesis is to construct flexible models for structured sparsity in a probabilistic setting. The work explores a class of models that uses Gaussian processes (Rasmussen and Williams, 2006) as a latent representation of the structure of sparsity patterns. This allows a priori knowledge of the sparsity structures to be encoded using generic covariance functions. The work also addresses how to perform efficient inference using these models, and how these models can be applied to solve ill-posed problems in the small N, large D regime. Specifically, two specific instances of ill-posed problems are investigated: *linear inverse problems* and *time-varying covariance estimation*, which are described in the following two sections.

1.1 Discrete Linear Inverse Problems

Discrete linear inverse problems is the simplest class of ill-posed inverse problems. In the classical setting, they describe a linear relationship between some observed effect $\boldsymbol{y} \in \mathbb{R}^N$ and its cause $\boldsymbol{x} \in \mathbb{R}^D$,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}, \tag{1.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ is the so-called *forward model* and $\boldsymbol{e} \in \mathbb{R}^N$ is additive noise. Predicting the observations $\boldsymbol{y}$ given the forward model $\boldsymbol{A}$ and the underlying cause $\boldsymbol{x}$ is known as the *forward problem*, while inferring the cause $\boldsymbol{x}$ from the observations $\boldsymbol{y}$ is known as the *inverse problem*.

A wide range of problems can be cast as linear inverse problems, including computer tomography (CT) (Natterer and Wang, 2002), magnetic resonance imaging (MRI) (Seeger et al., 2010), image deblurring (Jeffs, 1998), image denoising (Elad and Aharon, 2006), seismic imaging (Zhang et al., 2013), compressed sensing (Donoho, 2006), electroencephalography (EEG) source localization (Baillet et al., 2001) and feature selection in statistics and machine learning (Tibshirani, 1994; Fan and Lv, 2010).

Despite the simple relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$, linear inverse problems are difficult problems because they are often both ill-posed and ill-conditioned. To illustrate the consequence of an ill-conditioned forward model, assume temporarily that $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is square and non-singular. Applying the inverse operator $\boldsymbol{A}^{-1}$ to the noisy observations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ yields an estimate of $\boldsymbol{x}$

$$\hat{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{A}^{-1}\boldsymbol{e} = \boldsymbol{x} + \boldsymbol{x}_{\text{noise}} \tag{1.2}$$

where $\boldsymbol{x}_{\text{noise}} = \boldsymbol{A}^{-1}\boldsymbol{e}$. Because $\boldsymbol{A}$ is a bounded linear operator, it follows that

$$\|\boldsymbol{y}\| \leq \|\boldsymbol{A}\|_{\text{op}}\|\hat{\boldsymbol{x}}\|, \quad \|\boldsymbol{x}_{\text{noise}}\| \leq \|\boldsymbol{A}^{-1}\|_{\text{op}}\|\boldsymbol{e}\|, \tag{1.3}$$

where $\|\cdot\|_{\text{op}}$ is the operator norm. Re-arranging these inequalities yields

$$\frac{\|\boldsymbol{x}_{\text{noise}}\|}{\|\hat{\boldsymbol{x}}\|} \leq \frac{\|\boldsymbol{A}^{-1}\|_{\text{op}}\|\boldsymbol{e}\|}{\|\hat{\boldsymbol{x}}\|} \leq \|\boldsymbol{A}^{-1}\|_{\text{op}}\|\boldsymbol{A}\|_{\text{op}}\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{y}\|} = \mathcal{K}(\boldsymbol{A})\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{y}\|}, \tag{1.4}$$

where $\mathcal{K}(\boldsymbol{A}) = \|\boldsymbol{A}^{-1}\|_{\text{op}}\|\boldsymbol{A}\|_{\text{op}} \geq 1$ is the *condition number* of $\boldsymbol{A}$ defined by the ratio of maximum and minimum singular values of $\boldsymbol{A}$. Interpreting $\|\boldsymbol{x}_{\text{noise}}\|/\|\hat{\boldsymbol{x}}\|$ as the relative error on the solution, this calculation shows that the bound of the relative error of the (unregularized) solution of a linear system gets amplified by the condition number $\mathcal{K}(\boldsymbol{A})$.

The application of sparsity to linear inverse problems has been a major success giving rise to the widely known *least absolute shrinkage and selection operator*

(LASSO) (Tibshirani, 1994) and the field of *compressed sensing* (Donoho, 2006). Further, Candès et al. (2006) demonstrated that for certain forward models, it is possible to reconstruct the exact solution $\boldsymbol{x}$ of a noiseless linear inverse problem in the undersampled regime $N < D$, if the number of non-zero coefficients $K = \|\boldsymbol{x}\|_0$ is sufficiently small relative to N and D. The relationship between the *undersampling ratio*, $\delta = \frac{N}{D}$, and the degree of sparsity, $\rho = \frac{K}{N}$, gives rise to a phase transition that partitions the (δ, ρ)-space into two phases: *solvable* and *unsolvable* (Donoho and Tanner, 2010). In the solvable phase, the exact solution $\boldsymbol{x}$ can be recovered from noiseless linear measurements, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, with high probability, while this probability drops to zero in the unsolvable phase. Similar phase transitions can also be observed in noisy systems, where the location and shape of the phase transition depends on the noise distribution (Donoho et al., 2011) as well as the solver (Andersen, 2014).

The *multiple measurement vector (MMV) problem* is a natural extension of the linear inverse problem in eq. (1.1), where multiple measurements $\{\boldsymbol{y}_t\}_{t=1}^{T}$ are observed such that $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{e}_t$ for all $t \in [T] = \{t | t \in \mathbb{N}, 1 \leq t \leq T\}$. In matrix notation, the problem becomes

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}, \tag{1.5}$$

where $\boldsymbol{Y} = \big[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\big] \in \mathbb{R}^{N \times T}$, $\boldsymbol{X} = \big[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\big] \in \mathbb{R}^{D \times T}$, and $\boldsymbol{E} = \big[\boldsymbol{e}_1, \ldots, \boldsymbol{e}_T\big] \in \mathbb{R}^{N \times T}$. Solving the MMV problem under the assumption that all signals $\{\boldsymbol{x}_t\}_{t=1}^{T}$ are independent is equivalent to solving each of the T linear inverse problems in eq. (1.1) separately. However, if one can assume *joint sparsity* meaning that all signals $\{\boldsymbol{x}_t\}_{t=1}^{T}$ share the same sparsity pattern, then one can recover the solutions using significantly fewer observations in each measurement vector (Cotter et al., 2005). Thus, the phase transition curves can be improved for signals that exhibit joint sparsity. Moreover, the assumption of joint sparsity can be relaxed by allowing the support of $\boldsymbol{x}_t$ to evolve slowly as a function of t (Ziniel and Schniter, 2013a).

The central hypothesis of this work is that the phase transition curves can be further improved for signals that exhibit structured sparsity. That is, we hypothesize that the minimum number of required samples N can be decreased if the structure of the sparsity patterns is taken into account. This thesis focuses on the setting where the sparsity structure is *spatio-temporal* meaning that the support of $\boldsymbol{x}_t$ is correlated in both space and time. Therefore, we assume that the index t is a temporal index and we assume that each individual variable $x_{i,t}$ in $\boldsymbol{x}_t$ has an associated set of spatial coordinates $\boldsymbol{d}_i \in \mathbb{R}^P$.

From a statistical point of view, this setup is equivalent to a sparse linear regression problem with spatially and temporally correlated support. The probabilistic approach taken in this thesis generalizes easily to other types of observations

as well. For example, by using a binary observation model, the same approach can be used for sparse linear classification problems (Shevade and Keerthi, 2003) with structured support (Meier et al., 2008). Models for sparse linear regression and classification with structured sparsity are widely used in neuroscientific applications. For example, *brain decoding* problems can be cast as sparse classification problems (Michel et al., 2010) and *source localization* problems can be cast as sparse regression problems (Baillet et al., 2001).

Informally, brain decoding deals with predicting external stimuli from brain activity measurements. Consider an experimental setup, where a human subject is being systematically exposed to two different types of external visual stimuli, e.g. pictures of faces and houses (Haxby et al., 2001), while the brain activity of the subject is being recorded using functional magnetic resonance imaging (fMRI) (Friston, 2007). The task of predicting the type of stimuli based on the recorded brain activity is known as *decoding* and it can be cast as a sparse classification problem, where the response variables are the type of stimuli and the input features correspond to the measured activity in the different regions of the brain. The solution to the sparse classification problem yields a map of the brain regions that are relevant for predicting the stimuli. That is, the coefficient $x_{i,t}$ is associated with the i'th brain region at time t and the corresponding spatial coordinates $\boldsymbol{d}_i$ are the physical coordinates of the i'th brain region. Applying *unstructured sparsity* has been shown to increase robustness of the brain maps (Rish, 2013), but the degree of sparsity is important as it gives rise to a trade-off between spatial reproducibility and predictive accuracy (Lautrup et al., 1994; Rasmussen et al., 2012). Applying structured sparsity rather than unstructured sparsity then encourages spatially coherent solutions, which have been shown to further improve interpretation and stability (Michel et al., 2010; Baldassarre et al., 2012; Varoquaux et al., 2016). Source localization problems constitute a large part of the motivation for this work and are described in greater detail in the next section.

1.1.1 The EEG Source Localization Problem

Electroencephalography (EEG) is a non-invasive method to monitor electrical activity generated by the brain using electrodes attached to the scalp of a human subject. The brain and the surrounding tissue act as a volume conductor, which implies that the multivariate EEG signal recorded at the scalp is the result of spatial averaging of the neural activity generated inside the brain (Nunez and Srinivasan, 2006). The purpose of EEG source localization is to localize and identify the regions of the brain, where the neural activity was generated (Baillet et al., 2001), see Figure 1.1.

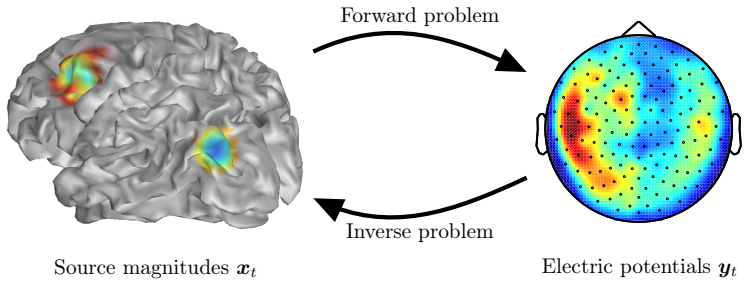


Figure 1.1: Illustration of the forward-inverse relationship for EEG source localization. Figure adapted from (Andersen, 2014).

In this setup, the brain is modeled using a set of D discrete current dipoles distributed across the cortex of the brain. Maxwell's equations then describe the relationship between the dipole sources and the electric potentials measured at the scalp, i.e. how the currents generated by the dipoles propagate through the brain tissue, the skull and the skin. Assuming quasi-stationarity, the relationship between the magnitude of a single source and a scalp potential becomes linear (Baillet et al., 2001). Thus, the model for a single sensor and a single dipole with magnitude x becomes

$$y(\boldsymbol{r}_{\text{sensor}}) = a(\boldsymbol{r}_{\text{sensor}}, \boldsymbol{d}_{\text{dipole}}, \theta)\, x, \tag{1.6}$$

where a is a function that depends on the position of the electrode $\boldsymbol{r}_{\text{sensor}}$, the position of the dipole $\boldsymbol{d}_{\text{dipole}}$, the orientation θ of the dipole as well as the conductivity of the tissue. Extending the model to N sensors and D dipole sources then follows from the principle of linear superposition, i.e. $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, assuming the geometrical configuration of the sensors and the sources is fixed. In a typical EEG setup, the number of sensors is on the order of $N \approx 10^1 - 10^2$, while the number of dipoles is on the order of $D \approx 10^3 - 10^4$ rendering the source localization problem a severely ill-posed problem. Additionally, the forward models are often severely ill-conditioned. To facilitate the study brain dynamics, the source localization model is naturally extended to the MMV setting by letting $\boldsymbol{y}_t$ be the measured EEG response at time t.

As the source localization problem is ill-posed and ill-conditioned, strong regularization is needed. It is often assumed that source distributions $\{\boldsymbol{x}_t\}_{t=1}^T$ are well approximated by sparse vectors (Nunez and Srinivasan, 2006; Delorme et al., 2012) and we will adopt this assumption in this thesis as well. Furthermore, it is also reasonable to assume that the support of the sources is temporally and spatially coherent (Ou et al., 2008; Gerven et al., 2009; Hansen and Hansen, 2017). Chapter 4 describes a model for source localization that incorporates these assumptions in a Bayesian setting.

1.2 Time-varying Covariance Estimation

Covariance matrix estimation is an important problem in multivariate statistics and machine learning. Covariance matrices are interesting quantities by themselves as they represent a measure of the coupling strength among a set of random variables, but covariance matrices also play a central role in several methods, such as linear discriminant analysis and principal component analysis. The covariance matrix of a multivariate random variable $\boldsymbol{x} \in \mathbb{R}^D$ is defined as

$$\mathrm{cov}_p[\boldsymbol{x}] = \mathbb{E}_p\left[(\boldsymbol{x} - \mathbb{E}_p\left[\boldsymbol{x}\right])(\boldsymbol{x} - \mathbb{E}_p\left[\boldsymbol{x}\right])^T\right], \tag{1.7}$$

where $\mathbb{E}$ is the expectation with respect to the density $p(\boldsymbol{x})$. For a D-dimensional stochastic variable, the number of degrees of freedom in the covariance matrix scales as $\mathcal{O}(D^2)$. Informally, this makes covariance matrix estimation in high dimensional spaces a challenging problem as a large number of samples $\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{x}}^2, \ldots, \hat{\boldsymbol{x}}^N \overset{iid}{\sim} p(\boldsymbol{x})$ is required to make the ratio $\frac{D^2}{N}$ small if D is large.

Additionally, in some applications it is of interest to study how the covariance matrix changes over time. More formally, let $p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ be the joint density of a multivariate time series $\{\boldsymbol{x}_t\}_{t=1}^T$ with the special property that the marginal covariance of the time series is time-dependent, i.e. $\mathrm{cov}_{p_t}\left[\boldsymbol{x}_t\right] = \boldsymbol{\Sigma}(t)$, where $p_t(\boldsymbol{x}_t)$ is the time-dependent marginal distribution of $\boldsymbol{x}_t$. The problem *time-varying covariance estimation* then deals with how to estimate the instantaneous covariance matrix $\hat{\boldsymbol{\Sigma}}(t)$ for all time points $t \in [T]$ given a set of N observations $\{\hat{\boldsymbol{x}}_1^n, \hat{\boldsymbol{x}}_2^n, \ldots, \hat{\boldsymbol{x}}_T^n\}_{n=1}^N \overset{iid}{\sim} p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$.

The problem of *dynamic functional connectivity* in neuroimaging (Hutchison et al., 2013; Damaraju et al., 2014; Gonzalez-Castillo et al., 2015; Yang et al., 2016) is an instance of this problem. *Functional connectivity* in the brain is defined as the temporal coupling between spatially remote neurophysiological events (Friston et al., 1993). Functional connectivity has been studied using multiple brain imaging modalities, including positron emission tomography (PET), magnetoencephalography (MEG), electroencephalography (EEG), and functional magnetic resonance imaging (fMRI). The coupling strengths are typically quantified using the correlation or covariance among the measured time series (Hutchison et al., 2013). The field of dynamic functional connectivity investigates how these couplings evolve over time. Thus, estimating the time-varying instantaneous covariance structure of a set of multivariate time series is the underlying statistical problem of dynamic functional connectivity.

The so-called *sliding window approach* is commonly used for time-varying covariance estimation in the neuroscience literature (Allen et al., 2014; Calhoun et al., 2014). The basic idea is to divide the multivariate time series of interest

into a set of overlapping windows and estimate the covariance matrix within each window independently. However, recent work have highlighted several issues with this approach (Hindriks et al., 2016; Shakil et al., 2016), such as the sensitivity to the window size. Also, the estimates of the covariance matrices for each window often suffer from high variance due to the small size N relative to the number of dimensions D. Chapter 5 describes a Bayesian approach to this problem, where sparsity and temporal smoothness are used for regularization.

1.3 Thesis outline

This thesis is structured as follows.

Chapter 2, Bayesian Modeling Preliminaries, introduces the concepts and terminology of Bayesian inference and modeling needed to discuss the contributions of the thesis. This includes an introduction to sparsity promoting priors and Gaussian processes.

Chapter 3, Approximate Inference, gives a brief introduction to approximate inference methods for probabilistic models. In particular, the chapter describes the two methods known as *expectation propagation* and *variational Bayes*, which will be needed in Chapter 4 and 5, respectively.

Chapter 4, The Structured Spike-and-slab Prior for Linear Models, presents the *structured spike and slab prior* for linear models and the associated inference algorithm and discusses how the model can be applied to ill-posed and ill-conditioned discrete linear inverse problems. The chapter also discusses the spatio-temporal extension of the model.

Chapter 5, Time-varying Covariance Estimation, presents a hierarchical model for time-varying covariance estimation and the associated inference algorithm. The chapter also discusses how the model can be applied to investigate dynamic functional connectivity.

Chapter 6, Discussion and Conclusion, summarizes this thesis and proposes interesting research directions for future work.

Chapter 2

Bayesian Modeling Preliminaries

The purpose of this chapter is to introduce the concepts from probability theory and Bayesian modeling that will be needed in the Chapter 4 (The Structured Spike-and-slab Prior for Linear Models) and Chapter 5 (Time-varying Covariance Estimation). Section 2.1 describes the general concepts in Bayesian modeling, and Section 2.2 specifically discusses sparsity from a Bayesian point of view. Finally, Section 2.3 gives a brief introduction to Gaussian processes.

For a more thorough treatment of Bayesian modeling and Gaussian processes, see (Bishop, 2006; Gelman et al., 2013) and (Rasmussen and Williams, 2006), respectively.

2.1 Bayesian Inference

The goal of this section is to introduce the core concepts and terminology in Bayesian inference and to set the notation for the rest of the thesis. A probabilistic model describes the distribution of outcomes of a system or an experiment. Typically, a model has a set of associated parameters $\boldsymbol{x} \in \mathbb{R}^D$ that govern the distribution of the outcomes. Deducing information about $\boldsymbol{x}$ from a set of

observations $\boldsymbol{y}$ is an example of the process of *probabilistic inference*.

First, we briefly describe the classical *maximum likelihood* approach to parameter estimation. The *likelihood function* $L : \mathbb{R}^D \to \mathbb{R}$ is a function of the parameters $\boldsymbol{x} \in \mathbb{R}^D$ and it is one of the central objects in classical statistics (Casella and Berger, 2002). Given a set of observations $\boldsymbol{y} \in \mathbb{R}^N$, the likelihood function is defined as

$$L(\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x}), \tag{2.1}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ is the joint probability density of the observations $\boldsymbol{y}$ given a specific set of parameters $\boldsymbol{x}$. The *maximum likelihood estimate* is defined as

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{x}} L(\boldsymbol{x}), \tag{2.2}$$

corresponding to the specific set of parameters for which the observed measurements $\boldsymbol{y}$ are most likely. Thus, the key operation of the maximum likelihood approach is optimization and the result is a point estimate $\hat{\boldsymbol{x}}_{\mathrm{ML}}$ of some unknown, but deterministic parameters $\boldsymbol{x}$.

Bayesian modeling, on the other hand, treats the unknown parameters $\boldsymbol{x}$ as a set of random variables and hence, the first step of a Bayesian analysis is to construct a joint probability distribution $p(\boldsymbol{x}, \boldsymbol{y})$ of the observed data and the unknown parameters. Using standard rules of probability theory (Bishop, 2006), the joint distribution can be decomposed into a conditional distribution and a marginal distribution as follows

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}), \tag{2.3}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ is referred to as the *sampling distribution* and $p(\boldsymbol{x})$ is the *prior distribution*. The sampling distribution plays a similar role as the likelihood function by describing the uncertainty associated with the sampling process conditioned on the parameters $\boldsymbol{x}$. The prior distribution $p(\boldsymbol{x})$ should capture our assumptions and uncertainty about $\boldsymbol{x}$ before any data is observed. The basic principle of Bayesian inference is then to update and summarize the current beliefs and uncertainty about $\boldsymbol{x}$ by combining information from the measurements $\boldsymbol{y}$ with information from the prior distribution. The updated beliefs are summarized in the *posterior distribution* of the parameters conditioned on the data, i.e. $p(\boldsymbol{x}|\boldsymbol{y})$, and it is derived from the joint distribution in eq. (2.3) by conditioning on $\boldsymbol{y}$

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{x})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}, \tag{2.4}$$

where the marginal distribution $p(\boldsymbol{y})$ is the normalization constant of the posterior distribution. The relationship between the posterior distribution, the

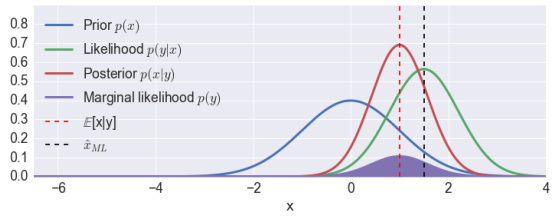


Figure 2.1: Illustration of the relationship between the prior, likelihood, posterior, and marginal likelihood in a Bayesian analysis of a Gaussian model. The marginal likelihood does not depend on $\boldsymbol{x}$, but is the area under the curve given by the product of the likelihood and the prior, i.e. $p(y) = \int p(y|x)p(x)\mathrm{d}x$.

likelihood and the prior in eq. (2.4) is known as the famous *Bayes' theorem* (Bishop, 2006). The posterior distribution contains all available information about $\boldsymbol{x}$, but the posterior expectation provides a summary of the information about $\boldsymbol{x}$ or some function of $\boldsymbol{x}$

$$\mathbb{E}\left[g(\boldsymbol{x})|\boldsymbol{y}\right] = \int g(\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}\boldsymbol{x}. \tag{2.5}$$

The marginal distribution $p(\boldsymbol{y})$ is also referred to as the *marginal likelihood* or *model evidence* and it is obtained by marginalizing the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ with respect to $\boldsymbol{x}$. That is,

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{2.6}$$

The integrals are substituted with summations when the space of parameters is a discrete space. Figure 2.1 illustrates the relationship between the basic components of a Bayesian analysis of a simple Gaussian model.

Suppose we were to make an additional set of observations $\boldsymbol{y}'$ after observing $\boldsymbol{y}$. Which values should we expect for $\boldsymbol{y}'$? The *posterior predictive distribution* answers this question by characterizing the uncertainty of future observations $\boldsymbol{y}'$ conditioned on the current observations $\boldsymbol{y}$

$$p(\boldsymbol{y}'|\boldsymbol{y}) = \int p(\boldsymbol{y}'|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}\boldsymbol{x}. \tag{2.7}$$

From the above marginalization integral, it is seen that both the uncertainty in the sampling distribution $p(\boldsymbol{y}'|\boldsymbol{x})$ and the uncertainty in the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ contribute to the predictive distribution.

The prior or the likelihood is often indexed by a set of parameters $\boldsymbol{\Omega}$ controlling the distributions, e.g. the mean and variance $\boldsymbol{\Omega} = \left\{\mu, \sigma^2\right\}$ for a Gaussian prior.

In a fully Bayesian treatment, this is handled by assigning a prior distribution to the *hyperparameters* $\boldsymbol{\Omega}$ and marginalizing over them. That is, the posterior distribution of $\boldsymbol{x}$ becomes

$$p(\boldsymbol{x}|\boldsymbol{y}) = \int p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Omega}) p(\boldsymbol{\Omega}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\Omega} = \int \frac{p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{\Omega})}{p(\boldsymbol{y})} p(\boldsymbol{\Omega}) \mathrm{d}\boldsymbol{\Omega}. \tag{2.8}$$

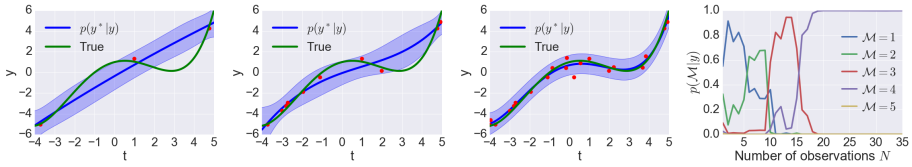
The process of averaging over a space of models with respect to the posterior model density, i.e. $p(\boldsymbol{\Omega}|\boldsymbol{y})$, is known as *Bayesian model averaging*. The conditioning is not explicitly shown in eq. (2.4) and eq. (2.6), but the term $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Omega})$ is equal to the left hand side of eq. (2.4) and $p(\boldsymbol{\Omega}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\Omega}) p(\boldsymbol{\Omega})$, where $p(\boldsymbol{y}|\boldsymbol{\Omega})$ is equal to the left hand side of eq. (2.6). However, we will in general omit the dependence on the hyperparameters to keep the notation uncluttered and only use the full notation when we are discussing Bayesian model averaging.

From eq. (2.5), (2.6), (2.7), and (2.8), it is evident that the core operation in Bayesian inference is integration rather than optimization and that the result is a set of distributions rather than a set of point estimates. Bayesian inference is a very elegant and appealing framework for uncertainty quantification from a mathematical point of view, but it is often the case that the required integrals cannot be evaluated analytically except for a small class of prior distributions known as *conjugate priors*. A prior distribution $p(\boldsymbol{x})$ is said to be conjugate to a likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ if the resulting posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ has the same functional form as the prior distribution (Bishop, 2006; Gelman et al., 2013). Therefore, conjugate priors are algebraically and computationally very convenient as the resulting posterior distribution will have a closed form solution.

On one hand, conjugate priors guarantee fast and exact inference through tractability of the posterior distribution, but there is often little flexibility in the shape of conjugate priors, which makes it difficult to represent specific prior knowledge using these priors. On the other hand, non-conjugate priors can be made arbitrary flexible, but the price is usually intractability and thus, approximate inference. When data is abundant, the specific choice of prior becomes less influential as the likelihood contribution will dominate the posterior distribution. But in the small N, large D regime, the choice of prior distribution becomes very important. Bayesian inference for non-conjugate models can be much more complicated and we will return to this issue in Chapter 3.

We will conclude this section with an example of a conjugate model. Consider the linear inverse problem $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ with a Gaussian prior, $p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}, \Lambda^{-1}\right)$ and Gaussian noise, $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{L}^{-1})$, yielding the joint distribution

$$p(\boldsymbol{y}, \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}, \boldsymbol{L}^{-1}\right) \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right). \tag{2.9}$$



(a) $N = 3$ (b) $N = 10$ (c) $N = 20$ (d) $p(\mathcal{M}|y)$ vs. N

Figure 2.2: Bayesian model averaging for linear regression. The data points (red dots) are noisy observations, $y_n = f(t_n) + e_n$, of the function $f(t_n) = 1 + 0.5t_n - 0.5t_n^2 + 0.02t_n^3 + 0.02t_n^4$ (green curve). Using eq. (2.11), the posterior distribution for five different models of the form $\hat{f}(t) = \sum_{m=0}^{\mathcal{M}} x_m t^m$ for $\mathcal{M} = 1, \ldots, 5$ is obtained. Panels (a)–(c) show the posterior predictive density for increasing sample sizes obtained by averaging over all five models using a uniform prior on the model complexity. Panel (d) shows the posterior distribution of the five models as a function of number of observations.

Assuming the noise covariance matrix $\boldsymbol{L}^{-1}$ is known, the Gaussian prior is conjugate to linear Gaussian models (Bishop, 2006) and hence, the posterior distribution and the model evidence can be in derived closed form. The model evidence can be derived by combining the result in eq. (2.6) with straight forward algebraic manipulations of the joint distribution in eq. (2.9) and is given by

$$p(\boldsymbol{y}|\boldsymbol{\Omega}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\right), \tag{2.10}$$

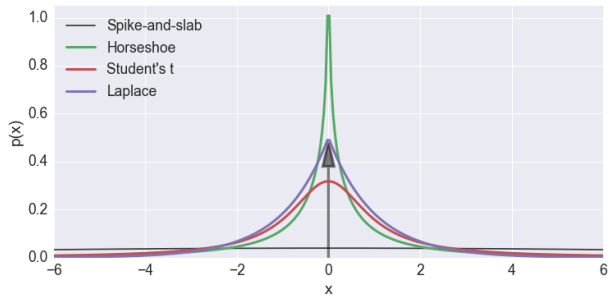
for $\boldsymbol{\Omega} = \{\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{L}\}$. Technically, the model evidence is also conditioned on $\boldsymbol{A}$, but we omit the dependence to keep the notation uncluttered. Similarly, the posterior distribution is given by

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Omega}) = \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{\Sigma}\right), \tag{2.11}$$

for $\boldsymbol{m} = \boldsymbol{\Sigma}\left[\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{y} + \boldsymbol{\Lambda}\boldsymbol{\mu}\right]$ and $\boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\right)^{-1}$. Finally, the predictive distribution for a new set of observations $\boldsymbol{A}' \in \mathbb{R}^{M \times D}, \boldsymbol{y}' \in \mathbb{R}^M$ is given by

$$p(\boldsymbol{y}'|\boldsymbol{y}, \boldsymbol{\Omega}) = \mathcal{N}\left(\boldsymbol{y}'|\boldsymbol{A}'\boldsymbol{m}, \boldsymbol{L}^{-1} + \boldsymbol{A}'\boldsymbol{\Sigma}(\boldsymbol{A}')^T\right). \tag{2.12}$$

Figure 2.2 illustrates Bayesian model averaging for a linear Gaussian model by averaging over five models with different complexity.



(a) $\mathbb{E}(z|y)$ vs y

Figure 2.3: Probability density functions for sparsity promoting priors. The horseshoe prior has an infinitely tall spike at $x = 0$, which is not shown in the figure (Carvalho et al., 2009). The spike-and-slab prior is a mixture of a Dirac delta distribution (black arrow) and Gaussian distribution (black curve) with parameters $p_1 = 0.5$ and $\tau = 10^2$.

2.2 Sparsity Promoting Priors and the Spike-and-slab Distribution

As mentioned in the previous section, the prior distribution $p(\boldsymbol{x})$ summarizes our prior beliefs and assumptions about $\boldsymbol{x}$ before we observe any data. Thus, if we believe $\boldsymbol{x}$ is sparse, this should be reflected in $p(\boldsymbol{x})$. Several *sparsity promoting priors* have been proposed for this purpose and they can mainly be divided in two categories: *weak* and *strong* sparsity priors (Mohamed et al., 2011). Weak sparsity priors are characterized by continuous distribution functions, while strong sparsity priors are *spike-and-slab* distributions.

Sparsity promoting priors for a variable $x \in \mathbb{R}$ characterized by continuous distributions are called weak sparsity priors because they assign zero probability to the event that the variable x is exactly zero. This implies that if one were to draw a finite sample from a weak sparsity prior, one would never observe a sample exactly equal to zero. Examples of weak sparsity priors are the Laplace distribution, the Student's t distribution and the horseshoe prior (Carvalho et al., 2009) (see Figure 2.3).

Strong sparsity priors, on the other hand, are specifically designed such that they assign a strictly positive probability $p_0 \in (0, 1)$ to the event $x = 0$. They

are constructed as discrete mixtures of the form

$$p(x) = p_0\delta(x) + p_1 f(x), \tag{2.13}$$

where $p_1 = 1 - p_0 \in (0, 1)$, f is continuous probability density function and δ is the so-called Dirac delta distribution. Despite the notation, δ is not defined as a function, but instead we define $\delta : \mathcal{P}_C \to \mathbb{R}$ as a linear functional on the space of all continuous probability density functions $\mathcal{P}_C$ with the following two properties

$$\delta_{x_0}(h) = \int_{\mathbb{R}} \delta(x - x_0)h(x)\mathrm{d}x = h(x_0), \tag{2.14}$$

$$\int_{\mathbb{R}} \delta(x - x_0)\mathrm{d}x = 1. \tag{2.15}$$

for all densities $h \in \mathcal{P}_C$ and points $x_0 \in \mathbb{R}$. However, we will continue to abuse the notation and write "$\delta(x)$" interpreting $\delta(x)$ as a density function with infinite density at $x = 0$ and zero elsewhere, but well aware that we cannot evaluate δ for any point $x \in \mathbb{R}$ directly. Nevertheless, we can still combine this prior with a likelihood $p(y|x)$ and do proper posterior inference even though we cannot evaluate the prior $p(x)$ directly.

First, we will compute the moments of the spike-and-slab prior. Using linearity and the normalization property in eq. (2.15), we can verify that $p(x)$ is properly normalized. That is,

$$\int_{\mathbb{R}} p(x)\mathrm{d}x = \int_{\mathbb{R}} (1 - p_0)\delta(x) + p_0 f(x)\mathrm{d}x \tag{2.16}$$

$$= (1 - p_0) \int_{\mathbb{R}} \delta(x)\mathrm{d}x + p_0 \int_{\mathbb{R}} f(x)\mathrm{d}x \tag{2.17}$$

$$= 1. \tag{2.18}$$

We can also compute the n'th moment of p for $n \geq 1$ using the property in eq. (2.14)

$$\mathbb{E}_p\left[x^n\right] = \int_{\mathbb{R}} x^n\, p(x)\mathrm{d}x \tag{2.19}$$

$$= (1 - p_0) \int_{\mathbb{R}} x^n \delta(x)\mathrm{d}x + p_0 \int_{\mathbb{R}} x^n f(x)\mathrm{d}x \tag{2.20}$$

$$= p_0\, \mathbb{E}_f\left[x^n\right]. \tag{2.21}$$

The prior distribution described in eq. (2.13) can be augmented to include an indicator variable $z \in \{0, 1\}$ for the event $x \neq 0$, i.e.

$$p(x|z) = (1 - z)\delta(x) + zf(x), \tag{2.22}$$

$$p(z) = \mathrm{Ber}(z|p_1), \tag{2.23}$$

where $z = \mathbb{I}[x \neq 0]$ is referred to as the support of x, $\text{Ber}(z|p)$ is a Bernoulli distribution with respect to z such that $p(z = 1) = p$ and $p(x) = \sum_{z\in\{0,1\}} p(x|z)p(z)$. This two-stage representation has the additional advantage that one can easily ask what is the posterior probability of x being non-zero, i.e. $p(z|y)$, for some likelihood $p(y|x)$.

From eq. (2.13), it is seen that p is a discrete mixture of a Dirac distribution (spike) and a continuous distribution (slab), hence the name *spike-and-slab prior*. Furthermore, it follows from eq. (2.22) that f is the conditional density of x given $z = 1$ and in principle, f can be chosen as the density function of any continuous distribution, e.g. a Laplace distribution (Ročková and George, 2016), a Gaussian distribution (Mitchell and Beauchamp, 1988) or a Gaussian mixture model (Vila and Schniter, 2013).

Inference for linear models with spike-and-slab priors

Before concluding this section, we will study the posterior distribution of a Gaussian linear model with zero-mean Gaussian spike-and-slab priors on each coefficient x_i for $i = 1, .., D$. Consider first the posterior distribution of a one-dimensional problem with a zero-mean Gaussian spike-and-slab prior on $x \in \mathbb{R}$ and a single Gaussian observation $y \in \mathbb{R}$ defined by the joint distribution

$$p(y, x, z) = \mathcal{N}(y|x, \sigma^2)\left[(1 - z)\delta(x) + z\mathcal{N}(x|0, \tau)\right]\text{Ber}(z|p_1), \tag{2.24}$$

Marginalizing z out and applying Bayes's theorem in eq. (2.4) yields the marginal posterior distribution

$$p(x|y) = \frac{\mathcal{N}(y|x, \sigma^2)\left[(1 - p_1)\delta(x) + p_1\mathcal{N}(x|0, \tau)\right]}{p(y)}. \tag{2.25}$$

The evidence $p(y)$ is obtained by marginalizing with respect to x and z using eq. (2.6)

$$p(y) = (1 - p_1)\mathcal{N}(y|0, \sigma^2) + p_1\mathcal{N}(y|0, \sigma^2 + \tau) \tag{2.26}$$

Substituting the expression for the evidence into eq. (2.25) and rearranging shows that the posterior distribution is also a spike-and-slab distribution,

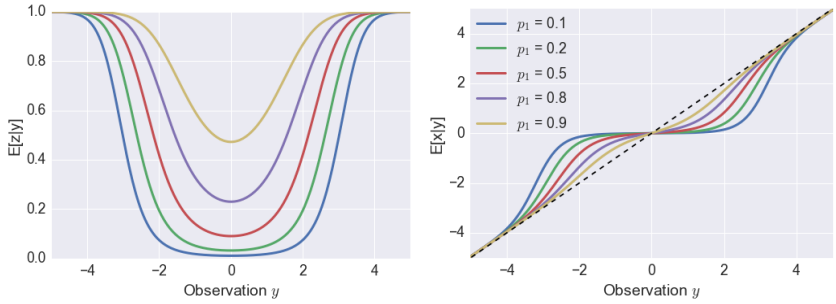
$$p(x|y) = p(z = 0|y)\delta(x) + p(z = 1|y)\mathcal{N}\left(x|\hat{m}, \hat{\sigma}^2\right) \tag{2.27}$$

for

$$\hat{m} = (1 + \sigma^2\tau^{-1})^{-1}y, \tag{2.28}$$

$$\hat{v} = (\sigma^{-2} + \tau^{-1})^{-1}, \tag{2.29}$$

$$p(z = 1|y) = \frac{p_1\mathcal{N}(y|0, \sigma^2 + \tau)}{(1 - p_1)\mathcal{N}(y|0, \sigma^2) + p_1\mathcal{N}(y|\rho_0, \sigma^2 + \tau)}. \tag{2.30}$$



(a) $\mathbb{E}(z|y)$ vs y (b) $\mathbb{E}(x|y)$ vs y

Figure 2.4: Posterior mean of x and z for the spike-and-slab model in eq. (2.24) for $\tau = 10^2$ and for different values of p_1.

Thus, we can readily obtain the posterior mean and variance using eq. (2.21).

Figure 2.3 shows the density of the Gaussian spike-and-slab prior. Figure 2.4(a) and (b) show the posterior mean of z and x, respectively, as a function of the observation y for several values of the prior probability p_1 and Figure 2.5 compares the posterior mean of x for each of the sparsity promoting prior distributions shown in Figure 2.3.

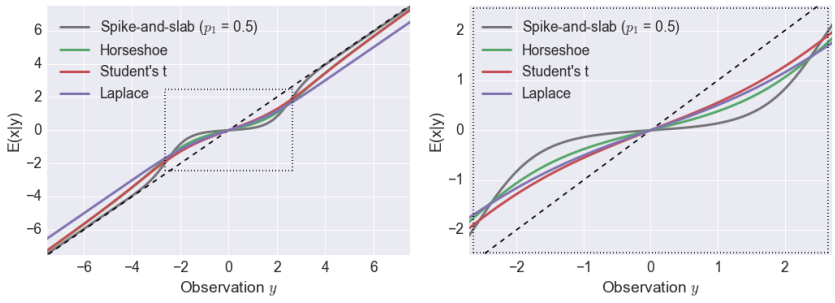
We will now extend the model to the D-dimensional case with N linear measurement $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$. The marginal posterior distribution of $\boldsymbol{x} \in \mathbb{R}^D$ given $\boldsymbol{y} \in \mathbb{R}^N$ becomes

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{\mathcal{N}\left(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}\right) \prod_{i=1}^{D} \left[(1-p_1)\delta(x_i) + p_1 \mathcal{N}\left(x_i|0,\tau\right)\right]}{p(\boldsymbol{y})}, \tag{2.31}$$

where the model evidence is given by

$$p(\boldsymbol{y}) = \sum_{\boldsymbol{z}} \int \mathcal{N}\left(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}, \sigma^2\boldsymbol{I}\right) \prod_{i=1}^{D} \left[(1-z_i)\delta(x_i) + z_i \mathcal{N}\left(x_i|0,\tau\right)\right] \prod_{i=1}^{D} \mathrm{Ber}\left(z_i|p_1\right) \mathrm{d}\boldsymbol{x}. \tag{2.32}$$

The model evidence in eq. (2.32) has a closed-form solution, but it contains 2^D terms due to the product of mixtures and hence, exact posterior inference is infeasible for even moderate sizes of D. Therefore, we have to resort to approximate inference as we will discuss in Chapter 3.



(a) (b)

Figure 2.5: Posterior mean of x for a model with joint density $p(y, x) = \mathcal{N}(y|x, \sigma^2)p(x)$ for each of the sparsity promoting priors shown in Figure 2.3.

2.3 Gaussian Processes

The previous section discussed prior distributions for sparsity and in this section, we will discuss prior distributions for functions. *Gaussian processes* are a non-parametric family of distributions over the uncountably infinite space of functions (Bishop, 2006). A random function f follows a Gaussian process distribution if any linear combination of the function values $f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)$ is Gaussian distributed when the function is evaluated at any finite set of points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N \in \mathcal{X}$ for some input space $\mathcal{X}$ (Rasmussen and Williams, 2006).

A Gaussian process (GP) is completely determined by its mean and covariance function,

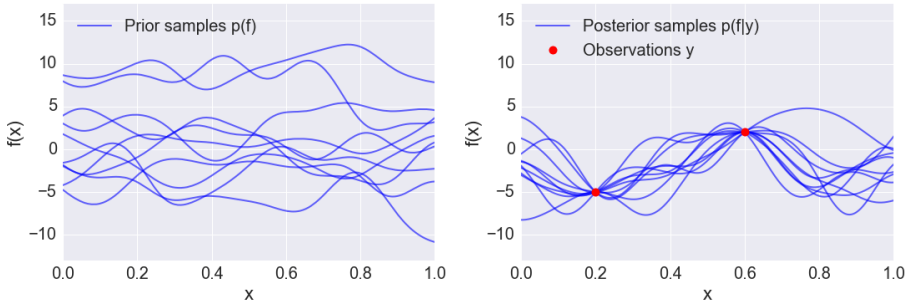
$$f(\boldsymbol{x}) \sim \mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right), \tag{2.33}$$

where the mean function $m : \mathcal{X} \to \mathbb{R}$ and the covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are defined as

$$m(\boldsymbol{x}) = \mathbb{E}\left[f(\boldsymbol{x})\right], \tag{2.34}$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[\left(f(\boldsymbol{x}) - m(\boldsymbol{x})\right)\left(f(\boldsymbol{x}') - m(\boldsymbol{x}')\right)^T\right] \tag{2.35}$$

where the expectations are with respect to the Gaussian process distribution. A function k is a valid covariance function if and only if the *Gram matrix* $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semidefinite for all pairs $\boldsymbol{x}_i, \boldsymbol{x}_j \in \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \subset \mathcal{X}$. If the mean function m can be assumed to be 0 for all $\boldsymbol{x}$, the covariance function k completely specifies the behavior of the Gaussian process. Properties



(a) Samples from the prior

(b) Samples from the posterior

Figure 2.6: (a) Samples from the prior distribution $p(\boldsymbol{f})$ with the kernel in eq. (2.36), (b) Samples from the posterior distribution $p(\boldsymbol{f}|\boldsymbol{y})$ after observing two data points (red dots).

like smoothness, stationarity, and periodicity can easily be encoded into the covariance function. For example, the following covariance function is composed of a *squared exponential* kernel and a bias kernel and is given by

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa_1 \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\kappa_2^2}\right) + \kappa_3. \tag{2.36}$$

The squared exponential kernel is an example of a stationary kernel that generates smooth sample paths with variance $\kappa_1 > 0$ and characteristic length scale $\kappa_2 > 0$. The bias kernel provides an offset with variance $\kappa_3 > 0$. Figure 2.6(a) shows 10 samples generated from a GP using the kernel in eq. (2.36) for points on a uniformly spaced one-dimensional grid.

Gaussian processes are a powerful and widely applied prior for Bayesian non-parametric regression problems. Consider a dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$ with N noisy observations $y_n = f(\boldsymbol{x}_n) + e_n$ for a set of input features $\boldsymbol{x}_n \in \mathcal{X}$ with isotropic Gaussian noise, i.e. $e_n \sim \mathcal{N}(0, \sigma^2)$. The joint distribution of the observations $\boldsymbol{y}$ and the latent function values $\boldsymbol{f}_n = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)]$ becomes

$$p(\boldsymbol{y}, \boldsymbol{f}|\boldsymbol{K}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{f}, \sigma^2\boldsymbol{I}\right)\mathcal{N}\left(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}\right). \tag{2.37}$$

By comparing the joint distribution in eq. (2.37) with the joint distribution of the linear Gaussian model in eq. (2.9), we can readily obtain the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{K})$ in closed-form using eq. (2.10). That is,

$$p(\boldsymbol{y}|\boldsymbol{K}, \sigma^2) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{C}\right), \tag{2.38}$$

for $\boldsymbol{C} = \sigma^2\boldsymbol{I} + \boldsymbol{K}$. Correspondingly, the posterior distribution $p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{K}, \sigma^2)$ can readily be obtained using eq. (2.11). Let us now turn the attention to the predictive distribution $p(f'|\boldsymbol{y})$ for a point $\boldsymbol{x}' \in \mathcal{X}$, which is constructed based on the covariance between the new point $\boldsymbol{x}'$ and each of the training points $\{\boldsymbol{x}_n\}_{n=1}^N$. The extended joint distribution becomes

$$p(\boldsymbol{y}, \boldsymbol{f}, f'|\boldsymbol{K}, \sigma^2) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{f}, \sigma^2\boldsymbol{I}\right)\mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}\\ f'\end{bmatrix} | \, \boldsymbol{0}, \begin{bmatrix}\boldsymbol{K} & \boldsymbol{k}'\\ (\boldsymbol{k}')^T & k(\boldsymbol{x}', \boldsymbol{x}')\end{bmatrix}\right), \tag{2.39}$$

where $\boldsymbol{k}' \in \mathbb{R}^N$ is a covariance vector with elements $\boldsymbol{k}'_n = k(\boldsymbol{x}_n, \boldsymbol{x}')$ for all $n \in [N]$. Since the distribution in eq. (2.39) is jointly Gaussian, the predictive distribution can be obtained using standard results for marginalization and conditioning in Gaussian distributions (Rasmussen and Williams, 2006),

$$p(f'|\boldsymbol{y}, \boldsymbol{K}, \sigma^2) = \mathcal{N}\left(f'\middle|\boldsymbol{k}^T\boldsymbol{C}^{-1}\boldsymbol{y},\ k(\boldsymbol{x}', \boldsymbol{x}') - \boldsymbol{k}^T\boldsymbol{C}^{-1}\boldsymbol{k}\right). \tag{2.40}$$

Although Gaussian processes are a versatile tool for function approximation, eq. (2.38) and (2.40) reveal a potential drawback of the approach. Specifically, the computational complexity scales cubically in the number of observations, i.e. $\mathcal{O}\left(N^3\right)$, and the memory footprint scales quadratically in the number of observations, i.e. $\mathcal{O}\left(N^2\right)$, which can make Gaussian processes prohibitively slow for large datasets. However, researchers have proposed several sparse approximation schemes to reduce the computational load (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Titsias, 2009).

Gaussian processes have also been studied for other sampling distributions, including (but not limited to) probit likelihoods for classification problems (Opper and Winther, 2000) and Student's t likelihoods for robust regression problems (Jylänki et al., 2011). However, the solution for these likelihoods cannot be obtained analytically and thus, approximations are needed. Finally, Gaussian processes are also widely used as building blocks in spatio-temporal models as we will see in Chapter 4 and 5.

Chapter 3

Approximate Inference

The posterior distribution of a random variable of interest conditioned on the observed data is one of the central objects in Bayesian analysis. However, it is often infeasible to evaluate posterior distributions for non-conjugate models because the model evidence in eq. (2.6) is computationally intractable.

For example, if a model contains discrete variables, the marginalization operation in eq. (2.6) requires summing over every possible configuration of the latent space, which can scale exponentially in the number of variables as seen in the spike-and-slab example in Section 2.2. For continuous latent variables, it is often the case that the marginalization integral in eq. (2.6) does not have an analytical solution and is too high dimensional to be calculated using numerical methods within reasonable time.

The field of approximate Bayesian inference strives to develop methods for approximating intractable distributions. These methods can mainly be divided into two categories: *sampling* methods and *deterministic* methods. The sampling-based methods are known as *Monte Carlo* methods (Andrieu et al., 2003), and the general idea is to approximate intractable integrals using finite sums. For example, the posterior mean of some function $g(\boldsymbol{x})$ can be approxi-

mated using the following finite sum

$$\mathbb{E}\left[g(\boldsymbol{x})|\boldsymbol{y}\right] = \int g(\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{y}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{M} \sum_{m=1}^{M} g(\hat{\boldsymbol{x}}_m), \tag{3.1}$$

where $\{\hat{\boldsymbol{x}}_m\}_{m=1}^{M} \sim p(\boldsymbol{x}|\boldsymbol{y})$ are samples from the posterior distribution.

The deterministic approximations, also sometimes referred to as *distributional approximations* (Gelman et al., 2013), are analytical approximations to the posterior distribution. The basic idea is to approximate a complicated distribution p with a simpler and tractable distribution q, e.g. $p(\boldsymbol{x}|\boldsymbol{y}) \approx q(\boldsymbol{x})$. The Laplace approximation (Williams and Barber, 1998), variational Bayes (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2016), and expectation propagation (Minka, 2001; Seeger, 2005) all belong to this category.

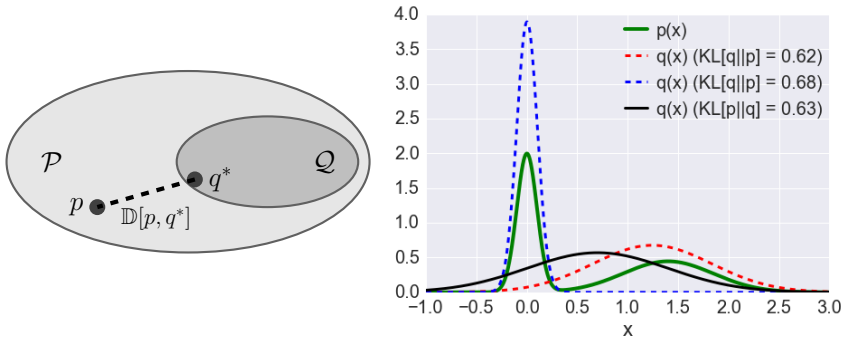
The purpose of the remainder of this chapter is to introduce the basic concepts of variational Bayes and expectation propagation. For more thorough treatments, see (Jordan et al., 1999; Wainwright and Jordan, 2008) and (Minka, 2001, 2004, 2005; Seeger, 2005), respectively.

3.1 Variational Bayes

The goal of variational inference is to approximate a target distribution $p \in \mathcal{P}$ with a tractable distribution $q \in \mathcal{Q}$. The general strategy is to define a space of tractable distributions $\mathcal{Q}$ and then optimize over that space to find a distribution $q \in \mathcal{Q}$, which is as similar to p as possible according to some measure of dissimilarity $\mathbb{D}[p, q]$, see Figure 3.1. In machine learning and Bayesian statistics, the dissimilarity between two distributions is typically measured using the so-called *Kullback-Leibler (KL) divergence* (MacKay, 2003; Bishop, 2006), which is defined by

$$\mathrm{KL}\left[q||p\right] = \int q(\boldsymbol{x}) \ln \left[\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right] \mathrm{d}\boldsymbol{x}. \tag{3.2}$$

The KL divergence is non-negative, i.e. $\mathrm{KL}\left[q||p\right] \geq 0$ for all $q, p \in \mathcal{P}$, and it satisfies $\mathrm{KL}\left[q||p\right] = 0 \iff p = q$, which makes it suitable as a dissimilarity measure. However, the KL divergence is also antisymmetric, i.e. $\mathrm{KL}\left[p||q\right] \neq \mathrm{KL}\left[q||p\right]$, and it fails to satisfy the triangle inequality. Hence, it does not qualify as a formal metric (MacKay, 2003).



(a) (b)

Figure 3.1: (a) Concept of variational inference. The target distribution $p \in \mathcal{P}$ is approximated by the 'closest' tractable distribution $q \in \mathcal{Q}$ where 'closeness' is measured by $\mathbb{D}[p, q^*]$. (b) The target density p (green) is a mixture of two Gaussian distributions, which is approximated by distribution $q(x)$ from the family of Gaussian distributions. The dashed curves (red and blue) show the two solutions, when the KL divergence is minimized in the direction $\mathrm{KL}\,[q||p]$, and the black solid curve shows the unique solution in the opposite direction direction, i.e. $\mathrm{KL}\,[p||q]$.

3.1.1 The Evidence Lower Bound (ELBO)

Approximating a posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ by minimizing the KL divergence $\mathrm{KL}\,[q||p]$ is often referred to as the *variational Bayes (VB)* method (Bishop, 2006). To do so, we substitute the expression for posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ from eq. (2.4) into eq. (3.2) to get

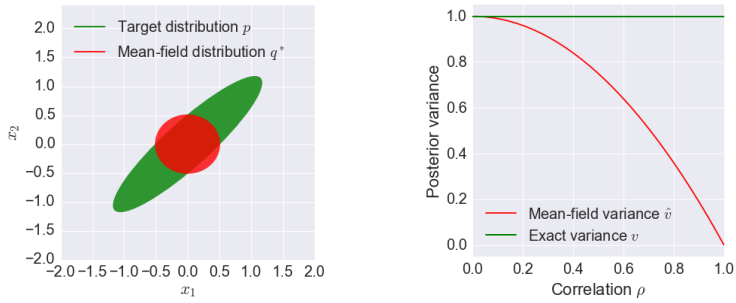
$$0 \leq \mathrm{KL}\,[q||p] = \mathbb{E}_q\left[\ln q(\boldsymbol{x})\right] - \mathbb{E}_q\left[\ln p(\boldsymbol{y}, \boldsymbol{x})\right] + \ln p(\boldsymbol{y}). \tag{3.3}$$

Re-arranging the terms yields the so-called *evidence lower bound (ELBO)*

$$\ln p(\boldsymbol{y}) \geq \mathcal{L}(q) = \mathbb{E}_q\left[\ln p(\boldsymbol{y}, \boldsymbol{x})\right] - \mathbb{E}_q\left[\ln q(\boldsymbol{x})\right], \tag{3.4}$$

where the gap between $\ln p(\boldsymbol{y})$ and $\mathcal{L}(q)$ is the $\mathrm{KL}\,[q||p]$. Thus, maximizing the functional $\mathcal{L}(q) : \mathcal{Q} \to \mathbb{R}$ is equivalent to minimizing $\mathrm{KL}\,[q||p]$ because the model evidence is constant.

The approximation in VB comes from the restriction of the space $\mathcal{Q}$. For example, if we choose $\mathcal{Q} = \mathcal{P}$ to be the space of all distributions, then the "best approximation" within $\mathcal{Q}$ would be the exact posterior distribution itself, i.e.



(a) (b)

Figure 3.2: (a) Approximating a bivariate zero-mean Gaussian distribution p with unit variances and correlation $\rho = 0.9$ with a mean-field approximation $q(x_1, x_2) = q(x_1)q(x_2)$. (b) Marginal variance of the exact distribution and the mean-field distribution as a function of the correlation ρ.

$q^* = p$. Therefore, the choice of $\mathcal{Q}$ is a trade-off between tractability and flexibility. A common choice is to restrict $\mathcal{Q}$ to the space of distributions with a specific parametric form, e.g. Gaussian distributions, or to restrict $\mathcal{Q}$ to the space of factorized distributions, i.e. $q(\boldsymbol{x}) = \prod_i q_i(x_i)$, leading to the *mean-field* approximation (Parisi, 1988; Wainwright and Jordan, 2008), or a combination of the two.

VB is a flexible framework for approximate inference, but one drawback is that the method in general tends to underestimate the uncertainty as illustrated in Figure 3.1 and 3.2. This is a consequence of the rather strong assumptions on $\mathcal{Q}$ combined with the properties of $\mathrm{KL}\left[q||p\right]$ (Minka, 2005). However, several methods for constructing more complex approximating families have recently been proposed (Sohl-Dickstein et al., 2015; Rezende and Mohamed, 2015), but we will not go into details with these methods.

3.2 Expectation Propagation

Expectation propagation (EP) (Minka, 2001; Opper and Winther, 2000) is another deterministic framework for approximating probability distributions. In the EP framework, a target distribution p is approximated by a distribution q from the *exponential family* by solving a series of local variational problems in an iterative fashion.

3.2.1 The Exponential Family

The exponential family $\mathcal{F}$ is the collection of distributions over $\boldsymbol{x} \in \mathbb{R}^D$ with density functions of the form

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - \mathcal{A}\left(\boldsymbol{\theta}\right)\right), \tag{3.5}$$

where $h : \mathbb{R}^D \to \mathbb{R}$ is a base measure, the function $\phi : \mathbb{R}^D \to \mathbb{R}^S$ is known as the *sufficient statistics*, $\boldsymbol{\theta} \in \mathbb{R}^S$ is a set of *natural parameters*, and $\mathcal{A} : \mathbb{R}^S \to \mathbb{R}$ is the *log-partition function* given by

$$\mathcal{A}\left(\boldsymbol{\theta}\right) = \ln \int \exp\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x})\right) \mathrm{d}h(\boldsymbol{x}). \tag{3.6}$$

See the work by Wainwright and Jordan (2008) for a detailed review of exponential families in the context of approximate inference.

3.2.2 The EP approximation

Consider a probability distribution p, that factorizes as follows

$$p(\boldsymbol{y}, \boldsymbol{x}) = \prod_{g=1}^{G} f_g(\boldsymbol{x}_g), \tag{3.7}$$

where $\boldsymbol{x}_g \subseteq \boldsymbol{x}$ is a subvector of $\boldsymbol{x}$. We assume that some of the factors depend on the data $\boldsymbol{y}$ even though the conditioning is not shown explicitly. For example, f_1 could be a prior distribution and f_g for $g \geq 2$ could be likelihood terms. EP approximates the target distribution p with a distribution q_{EP} that factorizes the same way as p

$$q_{\mathrm{EP}}(\boldsymbol{x}) \propto \prod_{g=1}^{G} \tilde{f}_g(\boldsymbol{x}_g), \tag{3.8}$$

where each *site* f_g is approximated with a *site approximation* $\tilde{f}_g$ from the exponential family. Since the exponential family is closed under multiplication (Seeger, 2005), the *global approximation* q_{EP} will also belong to the exponential family. Rather than approximating each site term individually, each site approximation $\tilde{f}_g$ is chosen such that it approximates f_g in the context of the remaining factors (Bishop, 2006). The context for the j'th site is specified by the so-called *cavity distribution*, which is defined as

$$q_{-j}(\boldsymbol{x}) \propto \frac{q_{\mathrm{EP}}(\boldsymbol{x})}{\tilde{f}_j(\boldsymbol{x}_j)} = \prod_{g \neq j} \tilde{f}_g(\boldsymbol{x}_g). \tag{3.9}$$

That is, the cavity distribution is equal to the global approximation, q_{EP}, where the contribution from the j'th site approximation has been removed. The core principle of EP is then to choose $\tilde{f}_j$ such that $\hat{q}_j(\boldsymbol{x}) = \tilde{f}_j(\boldsymbol{x}_j)q_{-j}(\boldsymbol{x})$ is a good approximation of the *tilted distribution* $\hat{p}_j(\boldsymbol{x}) = \frac{1}{Z} f_j(\boldsymbol{x}_j)q_{-j}(\boldsymbol{x})$, where Z is the normalization constant of $\hat{p}_j$. Specifically, the j'th site approximation is updated as follows

$$\tilde{f}_j^* = \arg\min_{\tilde{f}_j} \text{KL}\left[\hat{p}_j(\boldsymbol{x}) || \tilde{f}_j(\boldsymbol{x}_j)q_{-j}(\boldsymbol{x})\right]. \tag{3.10}$$

Note, that this is the KL divergence in the "opposite direction" compared to the variational Bayes approach described in the previous section. The distribution $\hat{q}_j$ can be interpreted as a projection of the tilted distribution $\hat{p}_j$ onto the space of exponential families with respect to the KL divergence (Minka, 2005). Because $\tilde{f}_j(\boldsymbol{x}_j)q_{-j}(\boldsymbol{x})$ belongs to the exponential family, the solution to the KL minimization problem in eq. (3.10) can be obtained by *moment matching* as we will now show.

3.2.3 Moment Matching

Inserting the definition of the exponential family in eq. (3.5) into the definition of the KL divergence from $\hat{p}_j$ to $\hat{q}_j \in \mathcal{F}$ in eq. (3.2) yields

$$\begin{aligned}\text{KL}\left[\hat{p}_j || \hat{q}_j\right] &= \mathbb{E}_{\hat{p}_j}\left[\ln \hat{p}_j(\boldsymbol{x}) - \ln \hat{q}_j(\boldsymbol{x})\right] \\ &= \mathbb{E}_{\hat{p}_j}\left[\ln \hat{p}_j(\boldsymbol{x})\right] - \mathbb{E}_{\hat{p}_j}\left[\ln h(\boldsymbol{x})\right] - \boldsymbol{\theta}^T \mathbb{E}_{\hat{p}_j}\left[\phi(\boldsymbol{x})\right] + \mathcal{A}\left(\boldsymbol{\theta}\right), \end{aligned} \tag{3.11}$$

where $\boldsymbol{\theta}_j$ are the natural parameters of $\hat{q}_j$. Next, we compute the gradient with respect to $\boldsymbol{\theta}_j$ and set it to zero

$$\nabla_{\boldsymbol{\theta}_j} \text{KL}\left[\hat{p}_j || \hat{q}_j\right] = -\mathbb{E}_{\hat{p}_j}\left[\phi(\boldsymbol{x})\right] + \nabla_{\boldsymbol{\theta}_j} \mathcal{A}\left(\boldsymbol{\theta}_j\right) = 0. \tag{3.12}$$

We can now use the fact that the gradient of the log-partition function $\nabla_{\boldsymbol{\theta}_j} \mathcal{A}\left(\boldsymbol{\theta}_j\right) = \mathbb{E}_q\left[\phi(\boldsymbol{x})\right]$ is given by the expected sufficient statistics with respect to $\hat{q}_j$ (Wainwright and Jordan, 2008) to derive the *moment matching condition*

$$\mathbb{E}_{\hat{p}_j}\left[\phi(\boldsymbol{x})\right] = \mathbb{E}_{\hat{q}_j}\left[\phi(\boldsymbol{x})\right], \tag{3.13}$$

which states that we can minimize the KL divergence by matching the expected sufficient statistics. For a Gaussian approximation, this is equivalent to matching the first two moments, i.e.

$$\hat{q}_j^*(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} | \, \mathbb{E}_{\hat{p}_j}\left[\boldsymbol{x}\right], \, \mathbb{E}_{\hat{p}_j}\left[\left(\boldsymbol{x} - \mathbb{E}_{\hat{p}_j}\left[\boldsymbol{x}\right]\right)^2\right]\right). \tag{3.14}$$

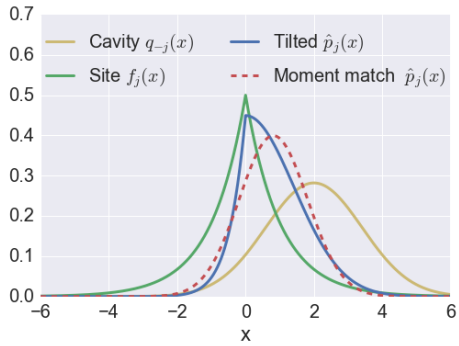


Figure 3.3: The moment matching mechanism in expectation propagation for a Laplacian site f_j (green). The tilted distribution $\hat{p}_j = \frac{1}{Z_j} f_j(x) q_{-j}(x)$ (blue) is approximated by a Gaussian distribution $\hat{q}_j(x)$ (red) by matching the first two moments according to (3.13).

See Figure 3.3 for an example. Once we have obtained a solution $\hat{q}_j^*(\boldsymbol{x})$, we can compute the update for the corresponding site approximation as follows

$$\tilde{f}_j^*(\boldsymbol{x}_j) \propto \frac{q_j^*(\boldsymbol{x})}{q_{-j}(\boldsymbol{x})}. \tag{3.15}$$

Finally, the global approximation is updated using eq. (3.8). This procedure is repeated for each site f_g until convergence or until a maximum number of iterations is reached.

3.2.4 The EP Evidence Approximation

The EP framework also provides an approximation of the model evidence $p(\boldsymbol{y})$. The approximation is obtained by substituting the exact site terms with a scaled version of the corresponding site approximation in the marginalization integral in eq. (2.6). That is,

$$p(\boldsymbol{y}) = \int \prod_{g=1}^{G} f_g(\boldsymbol{x}_g) \mathrm{d}\boldsymbol{x} \approx \int \prod_{g=1}^{G} \tilde{s}_g \tilde{f}_g(\boldsymbol{x}_g) \mathrm{d}\boldsymbol{x} = q_{EP}(\boldsymbol{y}), \tag{3.16}$$

where the scale factors $\tilde{s}_g > 0$ are chosen such that

$$\mathbb{E}_{q_{-j}}\left[f_j\left(\boldsymbol{x}_j\right)\right] = \tilde{s}_j \mathbb{E}_{q_{-j}}\left[\tilde{f}_j\left(\boldsymbol{x}_j\right)\right]. \tag{3.17}$$

After substituting the site approximations into eq. (3.16) , the integral becomes trivial to evaluate as it only contains products of exponential family densities. In contrast to the VB approach, $q_{EP}(\boldsymbol{y})$ is neither a lower or upper bound on the marginal likelihood.

3.2.5 The Expectation Propagation Algorithm

The update of the global approximation in eq. (3.8) is usually computationally expensive. The *parallel EP* algorithm decreases the computational load significantly by only updating the global approximation once after all site approximations have been updated (Gerven et al., 2009). This is in contrast to the *sequential EP* algorithm, where the global approximation is updated every time a site approximation is updated. The parallel EP algorithm is summarized in Algorithm 1.

repeat
  for *each site* f_g **do**
    Compute cavity distribution using eq. (3.9)
    Update site approximation using eq. (3.10) and (3.15)
  end
  Update global approximation using eq. (3.8)
until *converged or maximum number of iterations reached*;
Compute evidence approximation using eq. (3.16)

Algorithm 1: Parallel expectation propagation algorithm

The EP algorithm does indeed minimize local KL divergences, but EP does not minimize the global KL divergence from p to q. But the fixed points of the EP scheme can be shown to correspond to the stationary points of a specific energy function related to the negative logarithm of the marginal likelihood approximation in eq. (3.16) (Minka, 2001; Heskes et al., 2005; Minka, 2007).

We will conclude this section with a small discussion of pros and cons of EP. A major drawback of the EP framework is the lack of theoretical guarantees (Dehaene and Barthelmé, 2015). Specifically, neither the sequential nor the parallel version are guaranteed to converge. However, provably convergent *double loop* algorithms have been proposed (Heskes and Zoeter, 2002; Opper and Winther, 2005), but these algorithms can be much slower than the sequential or parallel algorithms. Furthermore, EP is sensitive to the numerical implementation and it can be numerically unstable for non-log-concave sites (Seeger, 2005; Wainwright, 2008; Jylänki et al., 2011). The computational complexity of EP can be prohibitively slow for large scale problems, e.g. the computational complexity

is $\mathcal{O}\left(D^3\right)$ for Gaussian process-based models (Jylänki et al., 2011). Along the same line, the memory footprint can also be prohibitively large for large scale problems because the number of sites increases with the number of likelihood terms, but the recently proposed *stochastic expectation propagation* (Li et al., 2015) deals with this issue.

On the positive side, EP has been shown to be the method of choice for several problems (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008; Jylänki et al., 2011; Barthelmé and Chopin, 2011; Cunningham et al., 2013; Peltola et al., 2014; Hernández-Lobato et al., 2015). From an uncertainty quantification point of view, EP is an appealing framework as it often produces better posterior uncertainties compared to variational Bayes, which tends to underestimate uncertainties in general (see Figure 3.1 and 3.2). Moreover, several interesting extensions of EP have been proposed, such as *Power EP*, which uses α-divergences as an alternative to KL divergences (Minka, 2005, 2004). Finally, an EP-based framework for distributed Bayesian inference for large datasets has recently been proposed (Gelman et al., 2017).

3.3 Approximate Inference for 1D Spike-and-slab Models

We conclude this chapter by studying the approximate solution obtained by applying the VB and EP methods to the 1D spike-and-slab model discussed in Section 2.2. Specifically, we will compute the approximations to the marginal posterior distributions $p(z|y)$ and $p(x|y)$ and compare these with the exact distributions from eq. (2.25).

We need to re-parametrize the model in eq. (2.24) to be able to apply the VB approximation since the lower bound in (3.4) is not well-defined because of the presence of the Dirac delta distribution in the prior distribution. To avoid this problem, we introduce a new variable $u \in \mathbb{R}$ such that $x = z \cdot u$, which gives rise to the following equivalent joint distribution (Titsias and Lazaro-Gredilla, 2011)

$$p(y, z, u) = \mathcal{N}(y|zu, \sigma^2)\mathcal{N}(u|0, \tau)\text{Ber}(z|p). \tag{3.18}$$

For the VB approximation, we consider two different choices of approximating families $\mathcal{Q}$. In both cases, $\mathcal{Q}$ is restricted to distributions with a fixed parametric form given by a product of a Gaussian density and a Bernoulli density:

$$q_{\text{MF}}(u, z) = \mathcal{N}\left(u|\hat{m}, \hat{v}\right)\text{Ber}\left(z|\hat{p}\right), \tag{3.19}$$

$$q_{\text{MF-TITSIAS}}(u, z) = \mathcal{N}\left(u|z\hat{m}, (1-z)\tau_0 + z\hat{v}\right)\text{Ber}\left(z|\hat{p}\right), \tag{3.20}$$

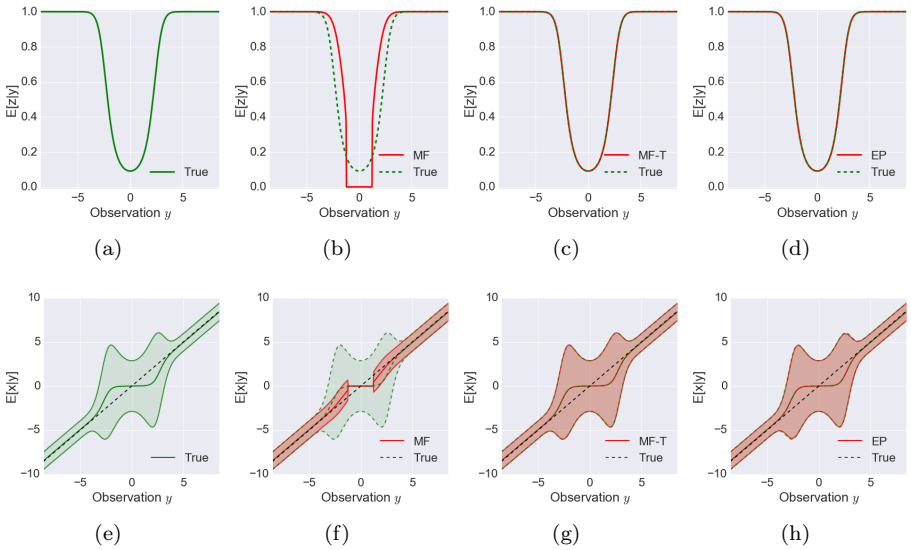


Figure 3.4: Comparison of approximate solutions of $p(z|y)$ and $p(x|y)$ for the spike-and-slab model in eq. (2.24) for $\tau = 10^2$ and $p_1 = 0.5$.

where $\hat{m}, \hat{v}$, and $\hat{p}$ are the variational parameters. In the former approximation, u and z are assumed to be independent in the posterior distribution, while the two variables are coupled in the latter approximation (Titsias and Lazaro-Gredilla, 2011). To find an optimal solution for the q_{MF} approximation given an observation y, we substitute the expression in eq. (3.19) into the expression for the lower bound in eq. (3.4) and optimize with respect to variational parameters. The procedure is the same for the approximation in eq. (3.20) as well.

We will now turn our attention to the EP approximation, which can be applied to both parametrizations of the model. In the parametrization in eq. (2.24), the likelihood, $p(y|x) = \mathcal{N}(y|x, \sigma^2)$, and the prior on the support, $p(z) = \text{Ber}(z|p_1)$, already belong to the exponential family, and hence we only have to approximate the site $p(x|z) = [(1-z)\delta(x) + z\mathcal{N}(x|0, \tau_0)]$ using EP. As $p(x|z)$ depends on both x and z, we choose the site approximations to be of the form

$$\hat{f}(x, z) = \mathcal{N}(x|\hat{m}_{\text{site}}, \hat{v}_{\text{site}}) \,\text{Ber}(z|\hat{p}_{\text{site}}) \,. \tag{3.21}$$

With this choice, the global EP approximation in eq. (3.8) becomes

$$\begin{aligned} q_{\text{EP}}(x, z) &\propto \mathcal{N}(y|x, \sigma^2) \overbrace{\mathcal{N}(x|\hat{m}_{\text{site}}, \hat{v}_{\text{site}}) \,\text{Ber}(z|\hat{p}_{\text{site}})}^{\hat{f}(x,z)} \,\text{Ber}(z|p_1) \\ &\propto \mathcal{N}(x|\hat{m}, \hat{v}) \,\text{Ber}(z|\hat{p}) \,, \end{aligned} \tag{3.22}$$

where the site parameters $\hat{m}_{\text{site}}$, $\hat{v}_{\text{site}}$, and $\hat{p}_{\text{site}}$ are obtained by moment matching using eq. (3.14) and (3.15) and the global parameters $\hat{m}$, $\hat{v}$, and $\hat{p}$ are given by

$$\hat{m} = \hat{v}\left[\sigma^{-2}y + \hat{v}_{\text{site}}^{-1}\hat{m}_{\text{site}}\right], \tag{3.23}$$

$$\hat{v} = \left(\sigma^{-2} + \hat{v}_{\text{site}}^{-1}\right)^{-1}, \tag{3.24}$$

$$\hat{p} = \left[(1-p_1)(1-\hat{p}_{\text{site}}) + p_1\hat{p}_{\text{site}}\right]^{-1} p_1\hat{p}_{\text{site}}. \tag{3.25}$$

The panels in the top row in Figure 3.4 compare the approximate solution for $p(z|y)$ with the exact solution from Section 2.2 and the panels in the bottom row compare the mean and standard deviation of the approximate solution for $p(x|z)$ with the exact solution.

Chapter 4

The Structured Spike-and-slab Prior for Linear Models

The purpose of this chapter is to present the contributions of the papers A, B, and C. All three papers are related to the so-called *structured spike-and-slab prior*, which generalizes the spike-and-slab prior distribution discussed in Section 2.2 to the structured sparsity setting. First, Section 4.1 introduces the structured spike-and-slab prior including temporal extensions. Next, Section 4.2 briefly discusses how to perform inference for linear models using structured spike-and-slab priors in the ill-posed setting with small N, large D. Finally, Section 4.3 summarizes the work in the three papers A, B, and C.

4.1 Structured Spike-and-slab Priors

As mentioned in the introduction in Chapter 1, the central hypothesis of this work is that the phase transition curves for sparse linear inverse problems can be improved for signals that exhibit spatio-temporal sparsity. The aim of this section is to construct a sparsity promoting prior distribution that encourages

spatio-temporal sparsity. First, we will ignore the temporal structure and consider the problem of imposing spatial structure on the support of a vector $\boldsymbol{x}_t \in \mathbb{R}^D$ in a probabilistic setting. Assume each entry $x_{i,t}$ in $\boldsymbol{x}_t$ has a set of associated spatial coordinates $\boldsymbol{d}_i \in \mathbb{R}^P$.

The starting point is the spike-and-slab prior as introduced in Section 2.2

$$p(\boldsymbol{x}_t) = \prod_{i=1}^{D} \left[(1 - p_1)\delta(x_{i,t}) + p_1 \mathcal{N}\left(x_{i,t}|0, \tau\right)\right]. \tag{4.1}$$

All variables, $x_{i,t}$ for all $i \in [D]$, are independent under this distribution as evidenced by the factorization and in the following, we will refer to this distribution as the *unstructured* spike-and-slab distribution. To impose structure onto the support of $\boldsymbol{x}_t$, we replace the fixed hyperparameter p_1 with a smooth function $g : \mathbb{R}^P \to (0, 1)$ that maps the spatial coordinates $\boldsymbol{d}_i$ to probabilities such that

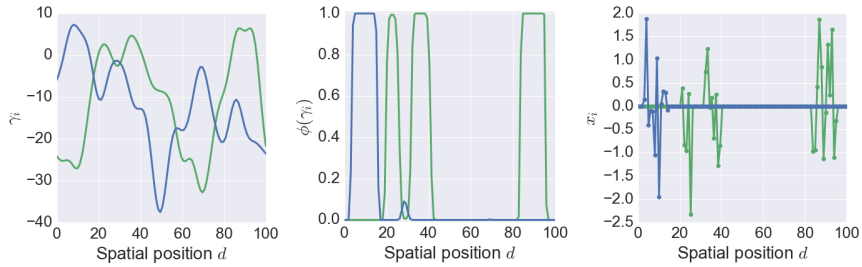
$$p(\boldsymbol{x}_t|\boldsymbol{g}) = \prod_{i=1}^{D} \left[(1 - g_i)\delta(x_{i,t}) + g_i \mathcal{N}\left(x_{i,t}|0, \tau\right)\right], \tag{4.2}$$

where $\boldsymbol{g} = [g(\boldsymbol{d}_1), g(\boldsymbol{d}_2), \ldots, g(\boldsymbol{d}_D)] \in (0, 1)^D$. The function g now controls the support probabilities directly, i.e. $p(x_{i,t} \neq 0|\boldsymbol{g}) = g(\boldsymbol{d}_i)$, and the smoothness assumption of g implies that nearby variables have similar support probabilities. Thus, the structure of the function g is reflected in the support of $\boldsymbol{x}_t$. However, the function g is usually not known in advance, so we will treat it as a random function and assign a prior distribution to it using the following construction

$$g(\boldsymbol{d}) = \phi\left(\gamma(\boldsymbol{d})\right), \tag{4.3}$$
$$\gamma(\boldsymbol{d}) \sim \mathcal{GP}\left(m(\boldsymbol{d}), k\left(\boldsymbol{d}, \boldsymbol{d}'\right)\right), \tag{4.4}$$

where $\phi : \mathbb{R} \to (0, 1)$ is the cumulative distribution function (CDF) of a standardized Gaussian distribution and $\gamma : \mathbb{R}^P \to \mathbb{R}$ is a random function with a Gaussian process prior distribution. This construction ensures that g always maps into the unit interval such that $g(\boldsymbol{d})$ can be interpreted as a probability for all $\boldsymbol{d} \in \mathbb{R}^P$. From a pure modeling perspective, the function ϕ could be any monotonically increasing function from the real line to the unit interval, but the Gaussian CDF is chosen because it simplifies the resulting inference procedure. Specifically, the moment matching integrals (see Section 3.2) in the expectation propagation algorithm can be solved analytically for the Gaussian CDF (Rasmussen and Williams, 2006).

By combining eq. (4.2), (4.3), and eq. (4.4), we can write the *structured spike*



(a) Realizations of $\boldsymbol{\gamma}$ (b) $\phi(\boldsymbol{\gamma})$ (c) Realizations of $\boldsymbol{x}_t$

Figure 4.1: Two realizations of the structured spike-and-slab prior for a problem, where $\boldsymbol{x}_t$ is composed of a set of variables positioned on a one-dimensional grid. (a) Two realizations of the latent Gaussian process prior representing the structure of the sparsity pattern using a squared exponential covariance function with length scale 10. (b) The same two realizations are squeezed into probabilities by the map $\phi : \mathbb{R} \to (0, 1)$. (c) Two realization of $\boldsymbol{x}$ conditioned on the support probabilities shown in panel (b).

and slab prior as follows

$$p(\boldsymbol{x}_t|\boldsymbol{\gamma}) = \prod_{i=1}^{D} \left[(1 - \phi(\gamma_i))\delta(x_{i,t}) + \phi(\gamma_i)\mathcal{N}\left(x_{i,t}|0, \tau\right)\right], \tag{4.5}$$

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}|\boldsymbol{m}, \boldsymbol{K}\right), \tag{4.6}$$

where $\boldsymbol{\gamma} = [\gamma(\boldsymbol{d}_1), \gamma(\boldsymbol{d}_2), \ldots, \gamma(\boldsymbol{d}_D)] \in \mathbb{R}^D$ and $\boldsymbol{m} \in \mathbb{R}^D$, $\boldsymbol{K} \in \mathbb{R}^{D \times D}$ are defined in a similar manner. Using this model, a priori knowledge of the structure of the sparsity pattern of $\boldsymbol{x}_t$ can now be expressed using generic covariance functions through the kernel of the GP. For example, choosing the covariance function to be a squared exponential kernel leads to sparsity patterns with spatial structure (see Figure 4.1), where the length scale parameter of the kernel controls the size of the spatial structures. The structured spike-and-slab prior is not limited to spatial sparsity structure, but it can model any type of sparsity structure that can be expressed using covariance functions, e.g. group structure can be obtained by using a block covariance matrix. Note that the coefficients, i.e. $x_{i,t}$ for all $i \in [D]$, remain conditionally independent given $\boldsymbol{\gamma}$ for all covariance functions.

The marginal support probabilities can also be controlled explicitly by manip-

ulating the mean $\boldsymbol{m}$ and the diagonal of the covariance matrix $\boldsymbol{K}$

$$p(x_{i,j} \neq 0) = \int \phi(\gamma_i)\mathcal{N}\left(\gamma_i | m_i, K_{ii}\right) \mathrm{d}\gamma_i = \phi\left(\frac{m_i}{\sqrt{1+K_{ii}}}\right). \tag{4.7}$$

Thus, if a subset of variables are a priori more likely to be non-zero than others, then this type of information can be encoded into the mean function. Further, when the mean function is constant and the kernel is proportional to an identity matrix, the marginal distribution of the coefficients, $p(\boldsymbol{x}_t)$, reduces to the unstructured spike-and-slab distribution given in eq. (4.1). However, for a general covariance matrix $\boldsymbol{K}$, $p(\boldsymbol{x}_t)$ has no analytical expression because the required marginalization integral is intractable.

We will now extend the structured prior to the spatio-temporal case by introducing $\boldsymbol{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\right] \in \mathbb{R}^{D\times T}$ and $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_T] \in \mathbb{R}^{D\times T}$, where t is assumed to be a temporal index. We will consider four different temporal extensions ordered by increasing complexity: *temporally IID*, *joint sparsity structure*, *first-order structure*, and *Kronecker structure*.

Temporally IID The simplest extension is to assume that the support is independently and identically distributed (IID) with respect to time,

$$p(\boldsymbol{X}|\boldsymbol{\Gamma}) = \prod_{t=1}^{T}\prod_{i=1}^{D}\left[(1-\phi(\gamma_{i,t}))\delta(x_{i,t}) + \phi(\gamma_{i,t})\mathcal{N}\left(x_{i,t}|0,\tau\right)\right], \tag{4.8}$$

$$p(\boldsymbol{\Gamma}) = \prod_{t=1}^{T}\mathcal{N}\left(\boldsymbol{\gamma}_t|\boldsymbol{m},\boldsymbol{K}\right). \tag{4.9}$$

Joint sparsity structure The joint sparsity model assumes that the support probabilities, $\phi(\boldsymbol{\gamma}_t)$, is constant with respect to time,

$$p(\boldsymbol{X}|\boldsymbol{\gamma}) = \prod_{t=1}^{T}\prod_{i=1}^{D}\left[(1-\phi(\gamma_i))\delta(x_{i,t}) + \phi(\gamma_i)\mathcal{N}\left(x_{i,t}|0,\tau\right)\right], \tag{4.10}$$

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}|\boldsymbol{m},\boldsymbol{K}\right). \tag{4.11}$$

First-order structure A more interesting model is to assume that $\boldsymbol{\gamma}_t$ evolves according to a first-order Markov process (Ziniel et al., 2010; Ziniel and Schniter, 2013a)

$$p(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{t-1}) = \mathcal{N}\left(\boldsymbol{\gamma}_t|(1-\alpha)\boldsymbol{m} + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{K}\right), \tag{4.12}$$

where $\alpha \in [0,1]$ controls the temporal correlation and $\beta > 0$ controls the innovation of the process. By choosing the initial distribution to be $\boldsymbol{\gamma}_1 \sim \mathcal{N}\left(\boldsymbol{m},\boldsymbol{K}\right)$,

the marginal distribution of $\boldsymbol{\gamma}_2$ becomes

$$p(\boldsymbol{\gamma}_2) = \int \mathcal{N}(\boldsymbol{\gamma}_2|(1-\alpha)\boldsymbol{m} + \alpha\boldsymbol{\gamma}_1, \beta\boldsymbol{K})\,\mathcal{N}(\boldsymbol{\gamma}_1|\boldsymbol{m}, \boldsymbol{K})\,\mathrm{d}\boldsymbol{\gamma}_1 \tag{4.13}$$

$$= \mathcal{N}\left(\boldsymbol{\gamma}_2|\boldsymbol{m}, (\alpha^2+\beta)\boldsymbol{K}\right), \tag{4.14}$$

where the last equality follows from eq. (2.10) in Section 2.1. Hence, it follows by induction that if α and β satisfy $\alpha^2 + \beta = 1$, then the marginal density of $\boldsymbol{\gamma}_t$ is $p(\boldsymbol{\gamma}_t) = \mathcal{N}(\boldsymbol{\gamma}_t|\boldsymbol{m}, \boldsymbol{K})$ for all $t \in [T]$. The first-order model reduces to the joint sparsity model in the degenerate case $\alpha = 1$ and $\beta = 0$, and to the temporal IID model when $\alpha = 0$ and $\beta = 1$.

Kronecker structure The first-order model in eq. (4.12) has two main advantages. First, it factorizes across time, which makes the resulting inference problem easier. Secondly, it only introduces one additional hyperparameter assuming the constraint $\alpha^2 + \beta = 1$ is satisfied. However, first-order dynamics are often not sufficient for capturing long range correlations. Imposing a joint Gaussian Process on the full $\boldsymbol{\Gamma}$-space circumvents this issue, i.e.

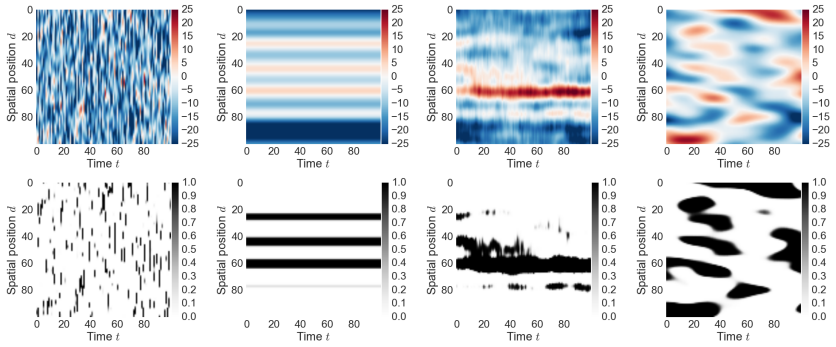
$$\mathrm{vec}[\boldsymbol{\Gamma}] \sim \mathcal{N}(\mathrm{vec}[\boldsymbol{M}], \boldsymbol{K}_{\boldsymbol{\Gamma}}), \tag{4.15}$$

where $\mathrm{vec}[\cdot]$ is the *vectorization operator* and $\boldsymbol{M} \in \mathbb{R}^{D\times T}$ and $\boldsymbol{K}_{\boldsymbol{\Gamma}} \in \mathbb{B}^{DT\times DT}$ are the mean and covariance matrix of $\boldsymbol{\Gamma}$, respectively. Because the input space of the joint Gaussian process takes the form of a Cartesian product grid, the covariance matrix simplifies to a Kronecker product, i.e. $\boldsymbol{K}_{\boldsymbol{\Gamma}} = \boldsymbol{K}_{\text{temporal}} \otimes \boldsymbol{K}$, where $\boldsymbol{K}_{\text{temporal}} \in \mathbb{R}^{T\times T}$ governs the temporal covariance and $\boldsymbol{K} \in \mathbb{R}^{D\times D}$ governs the spatial covariance. This decomposition implies that we never have to work with the full $DT \times DT$ covariance matrix directly, but instead we only have to work with the two smaller matrices $\boldsymbol{K}_{\text{temporal}}$ and $\boldsymbol{K}$, which will greatly reduce the computational complexity of the inference algorithm as well as the memory footprint. Kronecker products have earlier been used for efficient inference in Gaussian process modeling (Stegle et al., 2011; Wilson et al., 2014; Flaxman et al., 2015).

Figure 4.2 shows a realization of both $\boldsymbol{\Gamma}$ and $\phi(\boldsymbol{\Gamma})$ for each of the four different spatio-temporal priors. All four prior distributions can be augmented to include binary support variables $z_{i,t} \in \{0, 1\}$ for all $i \in [D]$ and $t \in [T]$ in the same way as discussed in Section 2.2.

4.2 Approximate Inference for Linear Models

The previous section introduced four different spatio-temporal prior distributions for $\boldsymbol{X}$ and $\boldsymbol{\Gamma}$ and in this section, we will discuss how to apply these priors



(a) I.I.D (b) Joint sparsity (c) First-order (d) Kronecker

Figure 4.2: Realizations of $\boldsymbol{\Gamma} \sim p(\boldsymbol{\Gamma})$ (top row), and $\phi(\boldsymbol{\Gamma})$ (bottom row) for the four different temporal extensions, where both $\boldsymbol{K}$ and $\boldsymbol{K}_{\text{temporal}}$ are chosen to be squared exponential kernels.

to ill-posed problems in the context of linear models. We focus on models where the sampling distributions factor across time,

$$p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}) = \prod_{t=1}^{T} f(\boldsymbol{y}_t|\boldsymbol{x}_t) \prod_{t=1}^{T}\prod_{i=1}^{D} p(x_{i,t}|z_{i,t}) \prod_{t=1}^{T}\prod_{i=1}^{D} p(z_{i,t}|\gamma_{i,t}) p(\boldsymbol{\Gamma}), \quad (4.16)$$

where $f(\boldsymbol{y}_t|\boldsymbol{x}_t)$ is the sampling distribution of $\boldsymbol{y}_t$ and

$$p(x_{i,t}|z_{i,t}) = (1 - z_{i,t})\delta(x_{i,t}) + z_{i,t}\mathcal{N}(x_{i,t}|0, \tau) \quad (4.17)$$

$$p(z_{i,t}|\gamma_{i,t}) = \text{Ber}(z_{i,t}|\phi(\gamma_{i,t}))\,. \quad (4.18)$$

For sparse linear regression problems with observations $\boldsymbol{y}_t \in \mathbb{R}^N$ for all $t \in [T]$, we use Gaussian sampling distributions of the form $f(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{A}\boldsymbol{x}_t, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ is the noise covariance. For sparse linear classification problems with observations $\boldsymbol{y}_t \in \{-1, 1\}^N$ for all $t \in [T]$, we use the probit likelihood $f(\boldsymbol{y}_t|\boldsymbol{x}_t) = \prod_{n=1}^{N} \phi(y_{nt}\boldsymbol{A}_{n,\cdot}\boldsymbol{x}_t)$, where $\boldsymbol{A}_{n,\cdot} \in \mathbb{R}^D$ denotes the n'th row of $\boldsymbol{A}$. Exact posterior inference is intractable for both observation models due to the product of mixture distribution in the prior as discussed in Section 2.2. Thus, we have to resort to approximate inference.

It has been shown empirically that expectation propagation-based algorithms (see Section 3.2) perform well for sparse linear models; both in terms of reconstructing the true weights $\boldsymbol{X}$ as measured by the mean squared error (MSE) and in terms of predictive power (Wainwright, 2008; Hernandez-Lobato et al., 2010; Hernández-Lobato et al., 2015). Specifically, Hernández-Lobato et al. (2013) proposed an EP algorithm for approximating the posterior distribution of linear

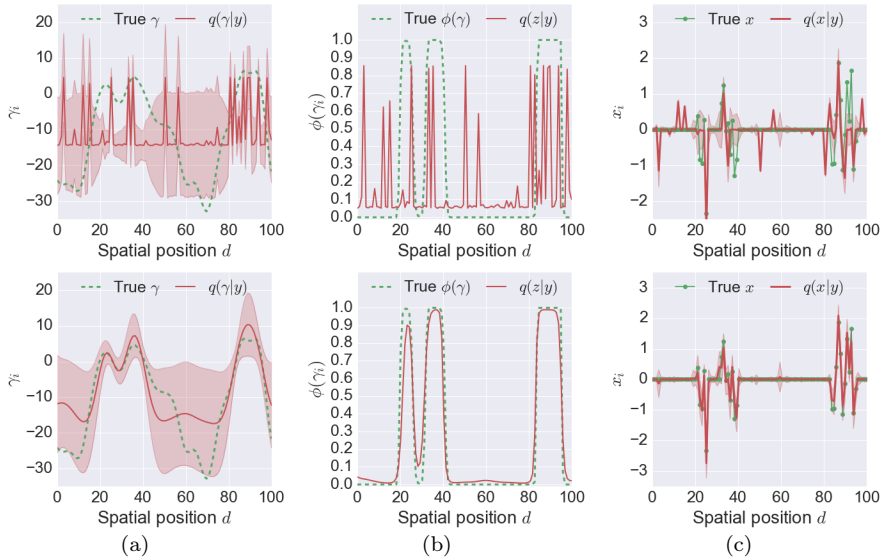
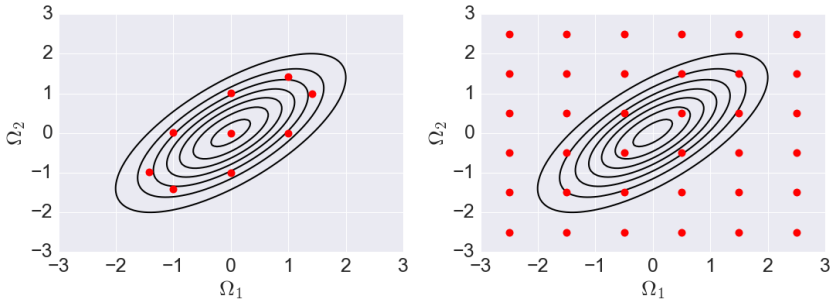


Figure 4.3: Illustration of the benefit obtained by modeling the structure of the sparsity pattern. The panels show the approximate posterior distributions for problem with $D = 100$ and $N = \frac{1}{2}D$ linear Gaussian measurements obtained using the expectation propagation algorithm without structure in the prior (top row) and with structure in the prior (bottom row) (a) Posterior distribution of $\boldsymbol{\gamma}$. (b) Posterior distribution of $\boldsymbol{z}$. (c) Posterior distribution of $\boldsymbol{x}$.

models with (unstructured) spike-and-slab prior distributions and showed that it outperformed competing methods. Therefore, we derive an expectation propagation algorithm for approximating the posterior distribution $p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y})$ by extending the work by Hernández-Lobato et al. (2013).

For the both observation models, all moment matching integrals can be evaluated analytically. Figure 4.3 shows an example of the EP approximation for a toy problem with a single Gaussian measurement vector ($T = 1$) and compares the results with the posterior distribution obtained using an unstructured prior distribution.

As described in Section 3.2, EP provides approximate posterior distributions for $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{\Gamma}$ conditioned on both the observations $\boldsymbol{Y}$ as well as the hyperparameters of the model. The set of hyperparameters $\boldsymbol{\Omega}$ for these models include hyperparameters of the likelihood, e.g. the noise covariance matrix, the variance of the "slab"-component, i.e. τ, and the hyperparameters of the kernel.



(a) Central composite design (b) Regular grid design

Figure 4.4: Illustration of the central composite design (CCD) for approximate numerical integration with respect to a posterior distribution $p(\boldsymbol{\Omega}|\boldsymbol{Y})$ over two hyperparameters $\boldsymbol{\Omega} = (\Omega_1, \Omega_2)$. The CCD grid (panel (a)) is constructed using the mode and the curvature at the mode of the distribution p and it uses significantly fewer points for numerical integration compared to a regular grid (panel (b)). If B is the number of hyperparameters, then the resulting number of CCD points is $M = 9, 15, 25, 43, 77$ for $B = 2, 3, 4, 5, 6$, respectively.

The ideal approach would be to take the uncertainty of the hyperparameters into account by assigning priors all hyperparameters and marginalize over them using eq. (2.8). However, this is intractable in the EP framework as the moment matching integrals are intractable. Instead, we consider two different evidence approximation schemes (MacKay, 1996). The first is an *empirical Bayes*-type approximation that simply uses the maximum a posteriori (MAP) value of $\boldsymbol{\Omega}$ as a point estimate

$$\hat{\boldsymbol{\Omega}}_{\text{MAP}} = \arg\max_{\boldsymbol{\Omega}} \left[\ln q_{\text{EP}}\left(\boldsymbol{Y}|\boldsymbol{\Omega}\right) + \ln p(\boldsymbol{\Omega})\right], \tag{4.19}$$

where $q_{\text{EP}}\left(\boldsymbol{y}|\boldsymbol{\Omega}\right)$ is the approximation of the model evidence in eq. (3.16) and $p(\boldsymbol{\Omega})$ is a prior distribution on the hyperparameters. In the second approximation, we approximate the marginalization integral eq. (2.8) using numerical integration as follows

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}) \approx \int q_{\text{EP}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}) q_{\text{EP}}\left(\boldsymbol{\Omega}\middle|\boldsymbol{Y}\right) \mathrm{d}\boldsymbol{\Omega} \tag{4.20}$$

$$\approx \sum_{m=1}^{M} q_{\text{EP}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_m) q_{\text{EP}}\left(\boldsymbol{\Omega}_m\middle|\boldsymbol{Y}\right) w_m, \tag{4.21}$$

where $q_{\text{EP}}\left(\boldsymbol{\Omega}\middle|\boldsymbol{Y}\right) \propto q_{\text{EP}}\left(\boldsymbol{Y}|\boldsymbol{\Omega}\right) p(\boldsymbol{\Omega})$ and $\{\boldsymbol{\Omega}_m\}_{m=1}^M$, $\{w_m\}_{m=1}^M$ is a set of integration points and weights, respectively. Thus, the resulting approximate

marginal posterior distribution becomes a Gaussian mixture model with mixing weights $\pi_m = q_{\text{EP}}\left(\boldsymbol{\Omega}_m \middle| \boldsymbol{Y}\right) w_m$ and components $q_{\text{EP}}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma} | \boldsymbol{Y}, \boldsymbol{\Omega}_m)$. We use a so-called Central Composite Design (CCD) (Rue and Martino) to keep the number of points and weights to a minimum, see Figure 4.4. The latter approach has been shown to yield better results as it takes the uncertainty of the hyperparameters into account (Vanhatalo et al., 2010) at the cost of increased computational complexity.

4.3 Contributions

4.3.1 Paper A: Bayesian Inference for Structured Spike and Slab Priors

Paper A proposes the structured spike-and-slab prior for a single measurement vector, as formulated in eq. (4.5) and (4.6), for models with linear Gaussian observations and isotropic noise. The paper also provides an expectation propagation algorithm for approximate inference scaling as $\mathcal{O}(ND^2 + D^3)$, where the term $\mathcal{O}(ND^2)$ comes from the update of the posterior distribution of $\boldsymbol{x}$ and the cubic term comes from the update of the posterior distribution of the Gaussian process. Since $\mathcal{O}(D^3)$ can be prohibitively slow for large scale problems, Paper A also proposes a low rank approximation scheme that reduces the computational complexity to $\mathcal{O}(ND^2 + RD^2)$, where $R \in [D]$ is a parameter controlling the trade-off between accuracy of the posterior distribution and the computational complexity.

Furthermore, Paper A also describes a series of numerical experiments for quantifying the performance of the proposed model and the associated inference algorithm. The normalized mean square error between the estimated solution $\hat{\boldsymbol{x}} = \mathbb{E}_{q_{\text{EP}}}[\boldsymbol{x}|\boldsymbol{y}]$ and the true solution $\boldsymbol{x}$ was used as a performance metric to quantify the algorithm's ability to recover the true solution $\boldsymbol{x}$. Similarly, the F-measure (Van Rijsbergen, 1979) between the estimated support $\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} q_{\text{EP}}(\boldsymbol{z})$ and true support $\boldsymbol{z}$ was used for quantifying the ability to recover the true support $\boldsymbol{z}$.

Importantly, the paper verifies the hypothesis that the location of the phase transition improves when the spatial structure of the sparsity pattern is taken into account. Moreover, the paper shows that the method is robust to the specific choice of covariance function for the kernels. Through experiments with an EEG forward model $\boldsymbol{A}$ (see Section 1.1.1) with condition number $\mathcal{K}(\boldsymbol{A}) \approx 8{\cdot}10^{15}$, the paper also demonstrates that the proposed algorithm is capable of inferring

the true sources despite the enormous condition number. To summarize, Paper A was a proof-of-concept paper for the structured spike-and-slab prior and the associated EP algorithm.

4.3.2 Paper B: Spatio-temporal Spike and Slab Priors for Multiple Measurement Vector Problems

Paper B extends the structured spike-and-slab prior and the inference algorithm to the spatio-temporal case with $T \geq 1$ by introducing the first-order model from eq. (4.12). The resulting EP inference scheme for the first-order model scales as $\mathcal{O}\left(TD^3\right)$, i.e. linearly in the number of measurement vectors T. Based on numerical experiments, paper B demonstrated that going from the single measurement case ($T = 1$) with spatial sparsity to the spatio-temporal case ($T \geq 1$) further improved the location of the phase transition.

4.3.3 Paper C: Bayesian Inference for Spatio-temporal Spike-and-slab Priors

Paper C extends and generalizes the work in Paper A and Paper B by allowing more general observation models, e.g. the probit likelihood for sparse classification. Paper C also introduces the spatio-temporal extension using the Kronecker formulation in eq. (4.15). The computational complexity of the EP algorithm for the Kronecker formulation is $\mathcal{O}\left(T^3D^3\right)$ with a memory footprint of $\mathcal{O}\left(T^2D^2\right)$ due to the full Gaussian Process prior on $\mathbf{\Gamma}$ in eq. (4.15). To reduce the computational complexity, the paper considers three different approximation schemes: *the low-rank approximation* as in Paper A, the *group approximation* and the *common precision approximation*.

As the name suggests, the group approximation clusters the variables in $\mathbf{\Gamma}$ into meaningful groups based on their spatio-temporal coordinates to reduce the dimension of the Gaussian Process. This approximation can yield significant speed ups for some problems without sacrificing accuracy and works well for problems, where the underlying spatial structure is a P-dimension regular grid. The drawback is that the approximation requires application-specific customization for problems associated with non-regular grids. The resulting computational complexity is $\mathcal{O}\left(T_g^3D_g^3\right)$, where $T_g \in [T]$, $D_G \in [D]$ are the number of groups in the temporal dimension and the spatial dimension, respectively.

The common precision approximation forces all site precision (inverse variance) parameters of the site approximations of the Bernoulli-sites in eq. (4.16) to have

the same value. By tying the site precision parameters together, it is possible to utilize properties of Kronecker products to reduce the computational complexity to $\mathcal{O}\left(TD^2 + T^2D\right)$ and the memory footprint to $\mathcal{O}\left(D^2 + T^2\right)$. The common precision approximation sacrifices the accuracy of the posterior uncertainty of $\boldsymbol{\Gamma}$, but because $\boldsymbol{\Gamma}$ is an auxiliary variable introduced to induce structure, the effect on the posterior distribution of $\boldsymbol{X}$ and $\boldsymbol{Z}$ is negligible.

Paper C also discusses several methods to handle the unknown values of the hyperparamaters of the model. The simplest method is to optimize the approximate model evidence provided by EP in eq. (3.16) using gradient-based methods yielding a ML estimate. A slightly more robust approach is to apply prior distributions to the hyperparameters and use the MAP value as a point estimate. Finally, the last option is to take the uncertainty of the hyperparameters into account approximating the marginalization integral in eq. (2.2) using an efficient numerical integration scheme called composite central design (CCD) (Rue and Martino; Vanhatalo et al., 2010). The latter slightly increases the computational load, but Paper C demonstrates that it provides better posterior uncertainties and it increases the accuracy of the model.

Finally, Paper C also describe a series of numerical experiments using both synthetic data and real data. The experiments with synthetic data serves to validate the proposed inference algorithm as well as study the phase-transition curves for inverse problems with Kronecker structure. The real data experiments include compressed sensing, phoneme classification, and EEG source localization. Both the compressed sensing problem and the phoneme classification problem show that the algorithm is capable of extracting meaningful information of high dimensional parameters in the small N, large D regime. These experiments also show that the proposed algorithm compare well with competing methods from the literature.

The EEG source localization problem was a setup with $N = 128$ electrodes, $D = 5124$ sources, and $T = 161$ measurement vectors yielding a total of $D \cdot T = 824964$ unknown sources. The specific dataset originated from a face perception study (Henson et al., 2003) and data for a single subject was analyzed using the spatio-temporal prior distribution with Kronecker structure. The results showed that the dataset was not informative about the hyperparameters of the kernel, i.e. the hyperparameters of the kernel could not be inferred from the data. However, by providing the algorithm with a priori knowledge of the length scale parameters of the spatial- and temporal covariance matrix, the algorithm detected four well-localized brain regions that are consistent with the findings from fMRI studies on the same experimental paradigm.

Chapter 5

Time-varying Covariance Estimation

The purpose of this chapter is to present the contributions of Paper D, which proposes a Gaussian process-based approach to time-varying covariance estimation. Section 5.1 introduces the proposed model and Section 5.2 summarizes the contributions of Paper D.

5.1 A Hierarchical Model for Time-varying Covariance Estimation

In this section, we will describe a hierarchical model for analysis of non-stationary multivariate time series with time-varying covariance structure. The model is applicable to multivariate time series in general, but it is specifically developed to analyze time-varying functional connectivity using fMRI time series data from multiple subjects as described in Section 1.2. However, the work included in this thesis is limited to making a model for time-varying covariance matrices and hence, the subsequent functional connectivity analysis based the estimated covariance matrices will not be discussed here.

As mentioned in Section 1.2, the dimension D of each time series is typically

larger than both the length of the time series T as well as the number of subjects N, which makes time-varying covariance estimation a challenging problem. The general idea is to model the time series for each subject using a multivariate Gaussian distribution, where the instantaneous covariance matrix of each time series changes slowly as a function of time. Furthermore, we assume that the instantaneous covariance matrix for each subject can be expressed using a latent representation that is shared across all subjects in a hierarchical manner.

Consider first the observation model. Assume we observe a set of multivariate time series, where $\boldsymbol{x}_t^n \in \mathbb{R}^D$ denotes the values of the time series at time $t \in [T]$ for subject $n \in [N]$. Let $\mathcal{D}^n = \{\boldsymbol{x}_t^n\}_{t=1}^T$ denote the entire time series for the n'th subject and let $\mathcal{D} = \{\mathcal{D}^n\}_{n=1}^N$ denote the collection of the observed time series for all subjects. We assume that the sampling distribution of $\boldsymbol{x}_t^n$ is a time-dependent multivariate Gaussian distribution

$$\boldsymbol{x}_t^n \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_t^n\right), \tag{5.1}$$

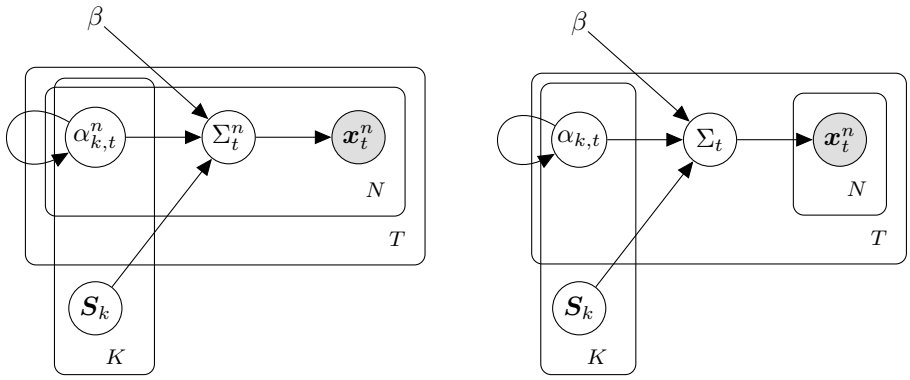
where $\boldsymbol{\Sigma}_t^n$ is the instantaneous covariance matrix time at t for the n'th subject. Thus, the observations $\{\boldsymbol{x}_t^n\}_{t=1}^T$ are assumed to be conditionally independent given $\{\boldsymbol{\Sigma}_t^n\}_{t=1}^T$ for all $n \in [N]$. We further assume that $\boldsymbol{\Sigma}_t^n$ decomposes into a non-negative linear combination of components $\boldsymbol{S}_k \in \mathbb{R}^{D \times D}$ as follows

$$\boldsymbol{\Sigma}_t^n = \beta^{-1}\boldsymbol{I} + \sum_{k=1}^K \alpha_{k,t}^n \boldsymbol{S}_k, \tag{5.2}$$

where $\mathcal{S} = \{\boldsymbol{S}_k\}_{k=1}^K$ is a dictionary of positive semi-definite *covariance matrix components*, $\mathcal{A} = \left\{\alpha_{k,t}^n \geq 0 \,|n \in [N], k \in [K], t \in [T]\right\}$ is a set of non-negative mixing weights, and K is the number of components. The parameter $\beta > 0$ controls the amount of additive white noise.

The noise precision β and the covariance matrix components $\boldsymbol{S}_k$ for all $k \in [K]$ are assumed to be independent of time. Hence, the temporal evolution of $\boldsymbol{\Sigma}_t^n$ for the n'th subject is controlled solely by the coefficients $\mathcal{A}^n = \left\{\alpha_{k,t}^n \,|k \in [K], t \in [T]\right\}$ and thus, the second order dynamics of $\mathcal{D}^n$ are completely determined by $\mathcal{A}^n$. Intuitively, $\mathcal{S}$ acts as a common basis for covariance matrices for all time points and all subjects, where $(\alpha_{1,t}^n, \alpha_{2,t}^n, \ldots, \alpha_{K,t}^n)$ are the coordinates for the instantaneous covariance matrix for the n'th subject at time t.

We will refer to the model defined in eq. (5.1) and (5.2) as the model with *subject-specific mixing weights*. The model is general in the sense that it assumes that the covariance matrix trajectories $\{\boldsymbol{\Sigma}_t^n\}_{t=1}^T$ are different for each subject as the mixing weights $\alpha_{k,t}^n$ depends on the subject index n. However, we will also consider a special case of eq. (5.2), where all subjects share the same mixing



(a) Subject specific mixing weights (b) Shared mixing weights

Figure 5.1: Graphical representation of the proposed models. The left-most figure shows the model with subject-specific mixing weights as in eq. (5.2) and the right-most figure shows the model with shared mixing weights across subjects as in eq. (5.3). The self-connections of $\alpha^n_{k,t}$ (left-panel) and $\alpha_{k,t}$ (right-panel) are indicating that the mixing weights are correlated in time (Hensman et al., 2013). The variables $a^n_{k,t}, v_{k,i}$, and $s_{k,i}$ have been left out of the figure for clarity.

weights, i.e. $\alpha_{k,t} = \alpha^n_{k,t}$ for all $n \in [N]$. Under this assumptions, the model simplifies to

$$\boldsymbol{\Sigma}_t = \beta^{-1}\boldsymbol{I} + \sum_{k=1}^{K} \alpha_{k,t}\boldsymbol{S}_k. \tag{5.3}$$

This implies that set the of subject time series become identically distributed on subject level. That is,

$$p(\mathcal{D}|\,\{\boldsymbol{\Sigma}_t\}_{t=1}^T) = \prod_{n=1}^{N} p(\mathcal{D}^n|\,\{\boldsymbol{\Sigma}_t\}_{t=1}^T). \tag{5.4}$$

We will refer to this model as the model with *shared mixing weights*. Both models are depicted as graphical models in Figure 5.1.

The idea is to infer the dictionary $\mathcal{S}$ and mixing weights $\mathcal{A}$ simultaneously from the set of observed time series $\mathcal{D}$. The hierarchical construction, where the covariance matrix components are shared across both time and subjects, allows us to pool data across multiple subjects when inferring the covariance matrix components. In contrast, the sliding window approaches estimate the local

covariance matrices independently for each time window and for each subject (Hutchison et al., 2013; Calhoun et al., 2014; Allen et al., 2014; Hindriks et al., 2016; Shakil et al., 2016).

After defining the sampling distributions in eq. (5.1), (5.2), and (5.3), we need to assign prior distributions to the mixing weights $\mathcal{A}$ and to the dictionary $\mathcal{S}$ to complete the Bayesian model. For simplicity, we will assume that the noise precision β is a deterministic hyperparameter.

Consider first the prior distribution for the mixing weights. First of all, it is imperative to ensure that $\boldsymbol{\Sigma}_t^n$ remains a valid covariance matrix for all $t \in [T]$ and for all $n \in [N]$. Secondly, we want to impose temporal smoothness on $\boldsymbol{\Sigma}_t^n$ for regularization. This assumption implies that two samples $\boldsymbol{x}_t^n$ and $\boldsymbol{x}_{t'}^n$ are more likely to have similar second order moments, if t and t' are close in time. Finally, we want to encourage the model to explain the instantaneous covariance matrix using as few covariance matrix components as possible at any given time point. These three desired properties for $\boldsymbol{\Sigma}_t^n$ can easily be translated into the following three properties for the mixing weights $\mathcal{A}^n$: non-negativity, temporal smoothness and sparsity, respectively. These properties are all satisfied by the following construction

$$\alpha_{k,t}^n = \max\left(0, a_{k,t}^n\right), \tag{5.5}$$

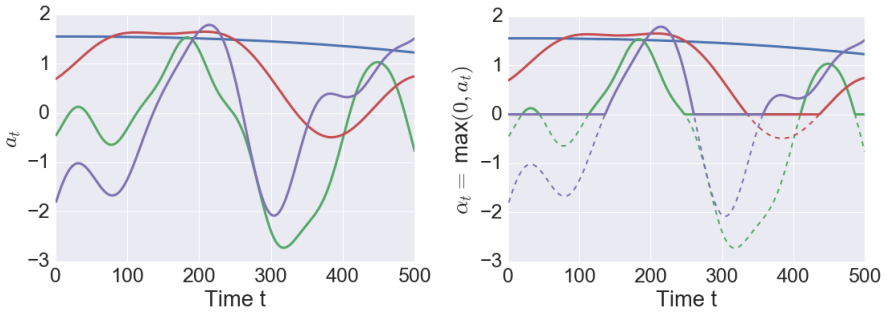
$$\boldsymbol{a}_k^n \sim \mathcal{N}\left(\boldsymbol{m}_k^n, \boldsymbol{K}_k^n\right), \tag{5.6}$$

where $\boldsymbol{a}_k^n = [a_k^n(1), a_k^n(2), \ldots, a_k^n(T)] \in \mathbb{R}^T$ is a random function with a Gaussian process prior evaluated at $t = 1, 2, \ldots, T$ and $\boldsymbol{m}_k^n \in \mathbb{R}^T$, $\boldsymbol{K}_k^n \in \mathbb{R}^{T \times T}$ are the prior mean and covariance matrix of $\boldsymbol{a}_k^n$, respectively. That is, the vector $\boldsymbol{\alpha}_k^n = \left[\alpha_{k,1}^n, \alpha_{k,2}^n, \ldots, \alpha_{k,T}^n\right] \in \mathbb{R}^T$ is obtained by element-wise thresholding of the vector $\boldsymbol{a}_k^n$. This construction can be interpreted as applying a rectified linear transformation (Nair and Hinton, 2010) to Gaussian processes. As with the structured spike-and-slab prior, the prior covariance matrix $\boldsymbol{K}_k^n$ governs the correlation structure of $\boldsymbol{a}_k^n$. Hence, the temporal smoothness properties of $\boldsymbol{\alpha}_k^n$ can be controlled through the covariance matrix $\boldsymbol{K}_k^n$. By allowing the prior covariance matrix for each component to have separate length scales, the model can encode both slowly and rapidly fluctuating components, see Figure 5.2.

The marginal support probability of $\alpha_{k,t}^n$, i.e. the marginal probability of the event $\alpha_{k,t}^n > 0$, is given by

$$p(\alpha_{k,t}^n > 0) = p(a_{k,t}^n > 0) = \int_0^\infty \mathcal{N}\left(a_{k,t}^n | \mu_{k,t}^n, C_{k,tt}^n\right) = \phi\left(\frac{\mu_{k,t}^n}{\sqrt{C_{k,tt}^n}}\right), \tag{5.7}$$

where $\phi : \mathbb{R} \to (0, 1)$ is the CDF of a standardized normal distribution. Thus,



(a) (b)

Figure 5.2: (a) Four realizations of the Gaussian process $\boldsymbol{a}_k^n \sim \mathcal{GP}\left(\boldsymbol{m}_k^n, \boldsymbol{K}_k^n\right)$ with squared exponential kernels with different length scales. (b) Linear rectified Gaussian processes, $\alpha_{k,t}^n = \max\left(0, a_{k,t}^n\right)$.

the expected number of non-zero entries in $\boldsymbol{\alpha}_k^n$ is controlled by the interplay between the prior mean and variance of the Gaussian process analogously to the structured spike-and-slab prior in Chapter 4. Similarly, the marginal density $p(\boldsymbol{\alpha}_k^n)$ has no analytical expression either. However, note that the structured spike-and-slab prior is a prior distribution for structured sparse signals, where the non-zero coefficients are Gaussian distributed and conditionally independent given the support. In contrast, the distribution defined in eq. (5.5) and (5.6) describes structured sparse signals, where the non-zero coefficients are non-negative and temporally smooth.

We will now turn our attention to the task of assigning prior distributions to the components $\boldsymbol{S}_k$. We will assume that each $\boldsymbol{S}_k$ is sparse, symmetric and of rank one. That is, $\boldsymbol{S}_k = \boldsymbol{v}_k \boldsymbol{v}_k^T$, where $\boldsymbol{v}_k \in \mathbb{R}^D$ is a sparse vector. This construction ensures that $\boldsymbol{S}_k$ is positive semidefinite as required. Using the (unstructured) spike-and-slab distribution from eq. (2.22) and (2.23) as a sparsity promoting prior for $\boldsymbol{v}_k$, the prior distribution becomes

$$p(\boldsymbol{v}_k, \boldsymbol{s}_k) = \prod_{i=1}^{D} \left[(1 - s_{i,k})\delta(v_{i,k}) + s_{i,k}\mathcal{N}\left(0, \tau_k\right)\right] \mathrm{Ber}\left(s_{i,k}|p_k\right) \tag{5.8}$$

where $s_{k,i} \in \{0, 1\}$ is a binary support variable for $v_{k,i}$, and $p_k \in (0, 1)$ is a component-specific hyperparameter controlling the degree of sparsity of the k'th component.

After completing the model specification, we will now discuss how to perform inference using the model. Recall, the primary goal is to estimate the instan-

taneous covariance matrix as a function of time for a set of time series $\mathcal{D}$. We will use the posterior expectation (see Section 2.1) of $\boldsymbol{\Sigma}_t^n$ conditioned on the data $\mathcal{D}$ to estimate the instantaneous covariance matrix for each subject, i.e. $\hat{\boldsymbol{\Sigma}}_t^n = \mathbb{E}\left[\boldsymbol{\Sigma}_t^n | \mathcal{D}\right]$. A secondary goal is to obtain the posterior distribution of the dictionary of components as well as the posterior of the mixing weights as they might convey meaningful information by themselves. However, the posterior distributions of interest are intractable and thus, we will resort to approximate inference using the variational Bayes method described in Section 3.1.

Because the covariance matrix components are of rank one, i.e. $\boldsymbol{S}_k = \boldsymbol{v}_k \boldsymbol{v}_k^T$, we can rewrite eq. (5.2) as

$$\boldsymbol{\Sigma}_t^n = \beta^{-1}\boldsymbol{I} + \sum_{k=1}^{K} \alpha_{k,t}^n \boldsymbol{v}_k \boldsymbol{v}_k^T = \beta^{-1}\boldsymbol{I} + \boldsymbol{V}\boldsymbol{A}_t^n \boldsymbol{V}^T, \tag{5.9}$$

where $\boldsymbol{V} = \begin{bmatrix}\boldsymbol{v}_1 \, \boldsymbol{v}_2 \, \ldots \, \boldsymbol{v}_K\end{bmatrix} \in \mathbb{R}^{D \times K}$ and $\boldsymbol{A}_t^n = \operatorname{diag}\left(\alpha_{1,t}^n \, \alpha_{2,t}^n \, \ldots \, \alpha_{K,t}^n\right) \in \mathbb{R}^{K \times K}$. By comparing this expression with eq. (2.10), this form is recognized as the marginalized covariance matrix of a linear factor model (Bishop, 2006) with Gaussian latent variables, i.e. $\boldsymbol{x}_t^n \sim \mathcal{N}\left(\boldsymbol{V}\boldsymbol{z}_t^n, \beta^{-1}\boldsymbol{I}\right)$ with $\boldsymbol{z}_t^n \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}_t^n\right)$. Thus, the model can be re-cast as a factor model, where the variances of the factors are time-dependent and controlled by linear rectified Gaussian processes. The same argument can be made for the covariance model with shared weights in eq. (5.3). Using the factor model representation, we derive an approximate inference algorithm for the model using a mean-field approximation. Specifically, we use factorized Gaussian distributions for approximating the posterior distributions of $\boldsymbol{z}_t^n$ and $\boldsymbol{a}_k^n$, and we use the "Titsias" distribution (see Section 3.3) for approximating the posterior distribution of the spike-and-slab variables $\boldsymbol{V}$. The hyperparameters of the model are learned from the data by optimizing the evidence lower bound with respect to the hyperparameters in a expectation-maximization (Bishop, 2006) manner.

5.2 Contributions

Paper D: A hierarchical model for time-varying functional connectivity

Paper D presents two hierarchical models for time-varying covariance estimation given a set of observed time series. The models are proposed as an alternative to the sliding window approach for analyzing dynamic functional connectivity. The first model assumes that second order dynamics of the observed data are

subject-specific in eq. (5.2), while the second model assumes that all subjects follow the same dynamics as in eq. (5.3). Both models assume that the instantaneous covariance matrix for each time series can be expressed using a latent representation that is shared across all subjects in a hierarchical manner. The paper also provides a variational Bayes algorithm for approximate posterior inference. The inference algorithm yields an estimate of the instantaneous covariance matrix for all time points for all subjects as well as posterior distributions of mixing weights $\mathcal{A}$ and the shared dictionary of covariance matrix components $\mathcal{S}$. The computational complexity of the inference algorithm is $\mathcal{O}\left(NTDK + KT^3\right)$.

Furthermore, the paper describes a series of numerical experiments using both synthetic data and real data. The experiments with synthetic data serve to evaluate the inference algorithm as well as compare the proposed method to reference methods (sliding window methods and hidden Markov models (Rabiner, 1989)). The log-euclidean Riemannian metric (LERM) (Vemulapalli and Jacobs, 2015; Huang et al., 2015), which defines a metric on the manifold of symmetric positive definite matrices, was employed to quantify the quality of the estimated covariance matrices.

Finally, the model with shared mixing weights was applied to an fMRI dataset from a motor task experiment (Van Essen et al., 2013). To validate the estimated sequence of instantaneous covariance matrices, the paper demonstrated that the task conditions could be predicted from the estimated covariance matrices.

Chapter 6

Discussion and Conclusion

The goal of this thesis was to develop probabilistic models for structured sparsity with the purpose of regularizing ill-posed problems in the small N, large D regime. The goal was two-fold. First, to construction prior distributions for structured sparsity and second, to design efficient inference algorithms for these prior distributions. Using Gaussian processes and spike-and-slab distributions as building blocks, this thesis has focused on models with structured sparsity for two specific ill-posed problems: linear inverse problems and time-varying covariance estimation.

Structured Spike-and-slab Priors for Linear Models

Paper A introduced the structured spike-and-slab prior as a generalization of the spike-and-slab prior. It is a sparsity promoting prior for structured sparse signals $\boldsymbol{X} \in \mathbb{R}^{D \times T}$, where the non-zero coefficients are Gaussian distributed and conditionally independent given the support. The structured spike-and-slab prior uses a latent Gaussian process to induce structure to the support of $\boldsymbol{X}$. This means that prior knowledge of the sparsity structure can be encoded using generic covariance functions through the kernel of the Gaussian process. The Gaussian process representation is advantageous to Ising models (Hernandez-Lobato et al., 2011) and more general Markov Random Field models (Cevher

et al., 2009) for structured sparsity, as global covariance functions are easier to interpret than clique potentials and partial correlations. Furthermore, several extensions were proposed (temporally IID, joint sparsity, first-order structure, and Kronecker structure) to extend the model to multiple measurement vector problems with spatially and temporally correlated support.

The proposed prior distribution was combined with Gaussian and probit likelihoods to form Bayesian models for sparse linear regression and sparse linear classification, respectively. An expectation propagation (EP) algorithm for approximate posterior inference was derived for both models. As the standard EP algorithm can be prohibitively slow for large scale problems, three additional approximations (the low rank approximation, the group approximation, and the common precision approximation) were proposed to reduce the computational complexity. The resulting algorithms were studied and evaluated intensively using numerical experiments with synthetic data in papers A, B, and C. These experiments demonstrated that the phase transition curves for linear inverse problems can be significantly improved when the structure of the sparsity patterns is taken into account. That is, the minimum number of noisy linear measurements $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}$ required to reconstruct $\boldsymbol{X}$ can be reduced significantly if $\boldsymbol{X}$ exhibits structured sparsity and if this sparsity structure is taken into account by the model. Paper A and C applied the proposed algorithms to two compressed sensing problems and a phoneme classification problem. The experiments showed that the proposed method was able to perform as well or better than competing methods from the literature.

Finally, Paper C also applied spatio-temporal spike-and-slab model to an EEG source localization problem using a dataset from a face perception study (Henson et al., 2009). This experiment established several important points. First of all, it was shown that the algorithm can be scaled to large scale problems. Secondly, it was also shown that the method was able to detect and localize (both spatially and temporally) four well-defined brain regions known to be associated with face perception (Henson et al., 2009). Finally, it was also concluded that the hyperparameters of the kernel of the underlying Gaussian process, i.e. the magnitude and length scales of the Gaussian process, could not be inferred from the data. This is not completely surprising as model section and hyperparameter learning are known to be difficult for ill-posed problems in neuroimaging (Rasmussen et al., 2012; Varoquaux et al., 2017). However, there are a number of possible extensions of this work that might alleviate this issue: *reduce model misspecification*, *incorporate more prior information*, or *extend the model to multiple subjects*. We will briefly discuss these potential research directions one by one.

The term "model misspecification" refers to inconsistencies between model and reality, e.g. wrong model assumptions. The proposed model assumes that the

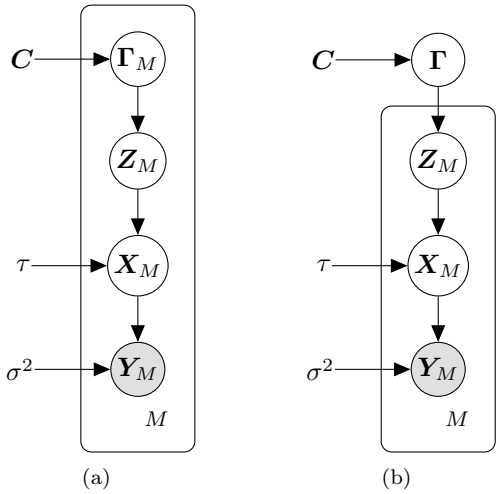


Figure 6.1: Generalization of the spatio-temporal spike-and-slab model to the source localization setup with M subjects. (a) only hyperparameters are shared across subjects. (b) hyperparameters and $\mathbf{\Gamma}$ are across subjects.

noise follows an isotropic Gaussian distributed. That is, the noise is assumed to be spatially and temporally independent, which is a rather crude assumption for EEG. For example, the work by Jun et al. (2006) and Engemann et al. (2015) shows that modeling the noise covariance is important for source localization. Incorporating a more realistic noise model will introduce additional hyperparameters that must be inferred, but using a realistic noise model is likely to render the posterior inference process intractable. However, it is likely that there exist a set of model assumptions in the spectrum between isotropic noise models and realistic noise models that are tractable and will improve robustness of the source localization. Pursuing such models is an interesting direction for future research.

The second potential research direction is centered around the idea of constructing better prior distributions for source localization i.e. to incorporate more prior information. This work has focused on the spatio-temporal structure of the support of the EEG sources, but there are other types of information that can be incorporated. For example, it might be beneficial to model the bilateral symmetry of the two hemispheres of the brain (Onton and Makeig, 2006; Friston et al., 2008; Hansen and Hansen, 2017). This can easily be achieved using the spatio-temporal spike-and-slab prior by adding a component to the covariance function that encourages spatial symmetry of the support between

the two hemispheres. Furthermore, the proposed model assumes that the non-zero coefficients in $\boldsymbol{X}$ are conditionally independent given the support. But it would also be interesting to include some degree of both spatial and temporal correlation structure to the non-zero coefficients of $\boldsymbol{X}$ for regularization (Baillet et al., 2001; Stahlhut et al., 2012; Jatoi et al., 2014) . Again, this is a trade-off between modeling more structure and keeping the number of hyperparameters to a minimum. However, a simple first-order temporal process on the non-zero coefficients will only introduce one additional hyperparameter.

Source localization problems are severely ill-posed because the number of sensors/measurements is much smaller than the number of parameters/sources and for practical reasons, it is almost impossible to get more than $\mathcal{O}(100)$ measurements per measurement vector (Nunez and Srinivasan, 2006). Another interesting research direction is therefore to extend the structured spike-and-slab model to a hierarchical model for multiple subjects as this is the most straightforward way to utilize more data. Consider a setup, where a number of subjects are exposed to identical stimuli in some simple experimental setup. The simplest hierarchical extension would be to assume that the kernel parameter of the individual subjects are shared such that both $\boldsymbol{\Gamma}$, $\boldsymbol{Z}$, and $\boldsymbol{X}$ are subject-specific, see Figure 6.1(a). A drawback of this model is that it has subject-specific Gaussian processes, which implies that the computationally expensive step of computing the posterior distribution of a Gaussian process must be carried out for every subject in every iteration. However, it is hypothesized that the support sets for each subject will be overlapping and hence, the hierarchical model shown in Figure 6.1(b) might be a more appropriate model. This model assumes that the support probabilities, i.e. $\phi(\boldsymbol{\Gamma})$, are shared across subjects, but that the support $\boldsymbol{Z}_m$ and the coefficients $\boldsymbol{X}_m$ are subject-specific. Using this model, all subjects will contribute to the estimation of the latent Gaussian process, which means that there is more data available for hyperparameter inference.

A Hierarchical model for Time-varying Covariance Estimation

Paper D introduces a hierarchical model for simultaneous analysis of multiple time series with time-varying covariance structure. The model is applicable to general time series data, but it is specifically developed with the problem of dynamic functional connectivity in mind (see Chapter 1). The general idea of the model is to assume that the instantaneous covariance matrix for each time series at any given time is expressed using a latent representation that is shared across all time series in a hierarchical manner. Informally, the model assumes that the instantaneous covariance matrix of each time series can be described in terms of a common basis and a set of time-varying non-negative coordinates.

To impose sparsity and smoothness onto the time-varying coordinates, Paper D proposes a sparsity promoting prior for structured sparse signals, where the non-zero coefficients are non-negative and smooth in time. The ”common basis” consists of a set of sparse rank one matrices, which are modeled as outer products of sparse vectors. A mean-field algorithm was derived for approximate posterior inference.

The model and the associated inference algorithm was studied and evaluated using a set of numerical experiments with synthetic data in Paper D. These experiments served to validate the inference algorithm and to compare the performance of the method with competing methods for datasets with known ground truth. Paper D also describes a numerical experiment, where the proposed model was applied to a real fMRI dataset, where subjects were exposed to a motor task paradigm (Van Essen et al., 2013). A sequence of time-varying covariance estimates were obtained using the model. To validate the results, the paper demonstrated that these covariance matrices was predictive of the task conditions of the experiment.

Paper D only scratches the surface of this class of models and there is a lot of interesting work to be done in terms of modeling, inference, and experiments. The field of dynamic functional connectivity is a relatively new and rapidly evolving field and thus, best practices and gold standards have not yet been established. Therefore, there are still many unanswered questions. For example, it is not clear whether the continuous dynamics approach (Smith et al., 2012) in this work is preferred over models with discrete switching dynamics, e.g. hidden Markov models (Nielsen et al., 2016; Sourty et al., 2016; Vidaurre et al., 2016), or vice versa. Therefore, it is necessary to test and evaluate the model on more datasets.

The proposed algorithm uses a variational approximation with a factorized mean-field distribution for inference, i.e. the posterior distribution ignores any correlation between the variables. As discussed in Chapter 3, these approximation are crude and they can often lead to severely underestimated uncertainties. It is therefore of interest to improve the quality of the posterior approximation, either by introducing more structure to the family of approximate distributions or by using a different inference method. But nevertheless, Paper D does indeed provide proof-of-concept that the method works.

To summarize the research contributions from both parts of the thesis, papers A, B, C, and D proposed and evaluated two approaches for imposing structured sparsity in the probabilistic setting and demonstrated the advantages of structured sparsity for linear inverse problems and time-varying covariance estimation.

Appendix A

Bayesian Inference for Structured Spike and Slab Priors

A Andersen, M. R., Winther, O., and Hansen, L. K. (2014), 'Bayesian inference for structured spike and slab priors'. Advances in Neural Information Processing Systems (NIPS) 2014, 9 pages

Bayesian Inference for Structured Spike and Slab Priors

Michael Riis Andersen, Ole Winther & Lars Kai Hansen
DTU Compute, Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark
{miri, olwi, lkh}@dtu.dk

Abstract

Sparse signal recovery addresses the problem of solving underdetermined linear inverse problems subject to a sparsity constraint. We propose a novel prior formulation, the structured spike and slab prior, which allows to incorporate a priori knowledge of the sparsity pattern by imposing a spatial Gaussian process on the spike and slab probabilities. Thus, prior information on the structure of the sparsity pattern can be encoded using generic covariance functions. Furthermore, we provide a Bayesian inference scheme for the proposed model based on the expectation propagation framework. Using numerical experiments on synthetic data, we demonstrate the benefits of the model.

1 Introduction

Consider a linear inverse problem of the form:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ is the measurement matrix, $\boldsymbol{y} \in \mathbb{R}^N$ is the measurement vector, $\boldsymbol{x} \in \mathbb{R}^D$ is the desired solution and $\boldsymbol{e} \in \mathbb{R}^N$ is a vector of corruptive noise. The field of sparse signal recovery deals with the task of reconstructing the sparse solution $\boldsymbol{x}$ from $(\boldsymbol{A}, \boldsymbol{y})$ in the ill-posed regime where $N < D$. In many applications it is beneficial to encourage a structured sparsity pattern rather than independent sparsity. In this paper we consider a model for exploiting a priori information on the sparsity pattern, which has applications in many different fields, e.g., structured sparse PCA [1], background subtraction [2] and neuroimaging [3].

In the framework of probabilistic modelling sparsity can be enforced using so-called sparsity promoting priors, which conventionally has the following form

$$p(\boldsymbol{x}|\lambda) = \prod_{i=1}^{D} p(x_i|\lambda), \tag{2}$$

where $p(x_i|\lambda)$ is the marginal prior on x_i and λ is a fixed hyperparameter controlling the degree of sparsity. Examples of such sparsity promoting priors include the Laplace prior (LASSO [4]), and the Bernoulli-Gaussian prior (the spike and slab model [5]). The main advantage of this formulation is that the inference schemes become relatively simple due to the fact that the prior factorizes over the variables x_i. However, this fact also implies that the models cannot encode any prior knowledge of the structure of the sparsity pattern.

One approach to model a richer sparsity structure is the so-called *group sparsity* approach, where the set of variables $\boldsymbol{x}$ has been *p*artitioned into groups beforehand. This

approach has been extensively developed for the ℓ_1 minimization community, i.e. *group LASSO*, *sparse group LASSO* [6] and *graph LASSO* [7]. Let $\mathcal{G}$ be a partition of the set of variables into G groups. A Bayesian equivalent of group sparsity is the group spike and slab model [8], which takes the form

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{g=1}^{G} \left[(1-z_g)\,\delta\left(\boldsymbol{x}_g\right) + z_g \mathcal{N}\left(\boldsymbol{x}_g|0, \tau \boldsymbol{I}_g\right)\right], \qquad p(\boldsymbol{z}|\boldsymbol{\lambda}) = \prod_{g=1}^{G} \text{Bernoulli}\left(z_g|\lambda_g\right), \quad (3)$$

where $\boldsymbol{z} \in [0,1]^G$ are binary support variables indicating whether the variables in different groups are active or not. Other relevant work includes [9] and [10]. Another more flexible approach is to use a Markov random field (MRF) as prior for the binary variables [2].

Related to the MRF-formulation, we propose a novel model called the *Structured Spike and Slab* model. This model allows us to encode a priori information of the sparsity pattern into the model using generic covariance functions rather than through clique potentials as for the MRF-formulation [2]. Furthermore, we provide a Bayesian inference scheme based on expectation propagation for the proposed model.

2 The structured spike and slab prior

We propose a hierarchical prior of the following form:

$$p(\boldsymbol{x}|\boldsymbol{\gamma}) = \prod_{i=1}^{D} p(x_i|g(\gamma_i)), \qquad p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma}_0\right), \quad (4)$$

where $g : \mathbb{R} \to \mathbb{R}$ is a suitable injective transformation. That is, we impose a Gaussian process [11] as a prior on the parameters γ_i. Using this parametrization, prior knowledge of the structure of the sparsity pattern can be encoded using $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. The mean value $\boldsymbol{\mu}_0$ controls the prior belief of the support and the covariance matrix determines the prior correlation of the support. In the remainder of this paper we restrict $p(x_i|g(\gamma_i))$ to be a spike and slab model, i.e.

$$p(x_i|z_i) = (1-z_i)\delta(x_i) + z_i \mathcal{N}\left(x_i|0, \tau_0\right), \qquad z_i \sim \text{Ber}\left(g(\gamma_i)\right). \quad (5)$$

This formulation clearly fits into eq. (4) when z_i is marginalized out. Furthermore, we will assume that g is the standard Normal CDF, i.e. $g(x) = \phi(x)$. Using this formulation, the marginal prior probability of the i'th weight being active is given by:

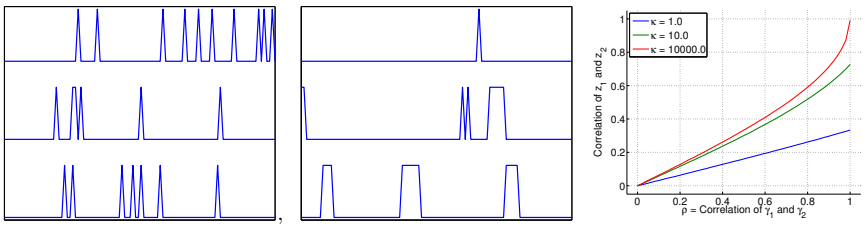
$$p(z_i = 1) = \int p(z_i = 1|\gamma_i) p(\gamma_i) \mathrm{d}\gamma_i = \int \phi(\gamma_i) \mathcal{N}\left(\gamma_i|\mu_i, \Sigma_{ii}\right) \mathrm{d}\gamma_i = \phi\left(\frac{\mu_i}{\sqrt{1+\Sigma_{ii}}}\right). \quad (6)$$

This implies that the probability of $z_i = 1$ is 0.5 when $\mu_i = 0$ as expected. In contrast to the ℓ_1-based methods and the MRF-priors, the Gaussian process formulation makes it easy to generate samples from the model. Figures 1(a), 1(b) each show three realizations of the support from the prior using a squared exponential kernel of the form: $\Sigma_{ij} = 50 \exp(-(i-j)^2/2s^2)$ and μ_i is fixed such that the expected level of sparsity is 10%. It is seen that when the scale, s, is small, the support consists of scattered spikes. As the scale increases, the support of the signals becomes more contiguous and clustered, where the sizes of the clusters increase with the scale.

To gain insight into the relationship between $\boldsymbol{\gamma}$ and $\boldsymbol{z}$, we consider the two dimensional system with $\mu_i = 0$ and the following covariance structure

$$\boldsymbol{\Sigma}_0 = \kappa \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \kappa > 0. \quad (7)$$

The correlation between z_1 and z_2 is then computed as a function of ρ and κ by sampling. The resulting curves in Figure 1(c) show that the desired correlation is an increasing function of ρ as expected. However, the figure also reveals that for $\rho = 1$, i.e. 100% correlation between the γ parameters, does not imply 100% correlation of the support variables $\boldsymbol{z}$. This



(a) Scale $s = 0.1$ (b) Scale $s = 5$ (c) Correlation of support

Figure 1: (a,b) Realizations of the support $\boldsymbol{z}$ from the prior distribution using a squared exponential covariance function for $\boldsymbol{\gamma}$, i.e. $\Sigma_{ij} = 50\exp(-(i-j)^2/2s^2)$ and $\boldsymbol{\mu}$ is fixed to match an expected sparsity rate K/D of 10%. (c) Correlation of z_1 and z_2 as a function of ρ for 5 different values of A obtained by sampling. This prior mean function is fixed at $\mu_i = 0$ for all i.

is due to the fact that there are two levels of uncertainty in the prior distribution of the support. That is, first we sample $\boldsymbol{\gamma}$, and then we sample the support $\boldsymbol{z}$ conditioned on $\boldsymbol{\gamma}$.

The proposed prior formulation extends easily to the multiple measurement vector (MMV) formulation [12, 13, 14], in which multiple linear inverse problems are solved simultaneously. The most straightforward way is to assume all problem instances share the same support variable, commonly known as joint sparsity [14]

$$p\left(\boldsymbol{X}\middle|\boldsymbol{z}\right) = \prod_{t=1}^{T}\prod_{i=1}^{D}\left[(1-z_i)\delta(x_i^t) + z_i\mathcal{N}\left(x_i^t\middle|0,\tau\right)\right], \tag{8}$$

$$p(z_i|\gamma_i) = \text{Ber}\left(z_i\middle|\phi(\gamma_i)\right), \tag{9}$$

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right), \tag{10}$$

where $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}^1 & \dots & \boldsymbol{x}^T\end{bmatrix} \in \mathbb{R}^{D\times T}$. The model can also be extended to problems, where the sparsity pattern changes in time

$$p\left(\boldsymbol{X}\middle|\boldsymbol{z}\right) = \prod_{t=1}^{T}\prod_{i=1}^{D}\left[(1-z_i^t)\delta(x_i^t) + z_i^t\mathcal{N}\left(x_i^t\middle|0,\tau\right)\right], \tag{11}$$

$$p(z_i^t|\gamma_i^t) = \text{Ber}\left(z_i^t\middle|\phi(\gamma_i^t)\right), \tag{12}$$

$$p(\boldsymbol{\gamma}_1,...,\boldsymbol{\gamma}_T) = \mathcal{N}\left(\boldsymbol{\gamma}_1\middle|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right)\prod_{t=2}^{T}\mathcal{N}\left(\boldsymbol{\gamma}_t\middle|(1-\alpha)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1},\beta\boldsymbol{\Sigma}_0\right), \tag{13}$$

where the parameters $0 \leq \alpha \leq 1$ and $\beta \geq 0$ controls the temporal dynamics of the support.

3 Bayesian inference using expectation propagation

In this section we combine the structured spike and slab prior as given in eq. (5) with an isotropic Gaussian noise model and derive an inference algorithm based on expectation propagation. The likelihood function is $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y}\middle|\boldsymbol{A}\boldsymbol{x},\sigma_0^2\boldsymbol{I}\right)$ and the joint posterior distribution of interest thus becomes

$$p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{1}{Z}p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}) \tag{14}$$

$$= \frac{1}{Z}\underbrace{\mathcal{N}\left(\boldsymbol{y}\middle|\boldsymbol{A}\boldsymbol{x},\sigma_0^2\boldsymbol{I}\right)}_{f_1}\underbrace{\prod_{i=1}^{D}\left[(1-z_i)\delta(x_i) + z_i\mathcal{N}\left(x_i\middle|0,\tau_0\right)\right]}_{f_2}\underbrace{\prod_{i=1}^{D}\text{Ber}\left(z_i\middle|\phi\left(\gamma_i\right)\right)}_{f_3}\underbrace{\mathcal{N}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right)}_{f_4},$$

where Z is the normalization constant independent of $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{\gamma}$. Unfortunately, the true posterior is intractable and therefore we have to settle for an approximation. In particular, we apply the framework of expectation propagation (EP) [15, 16], which is an iterative deterministic framework for approximating probability distributions using distributions from the exponential family. The algorithm proposed here can be seen as an extension of the work in [8].

As shown in eq. (14), the true posterior is a composition of 4 factors, i.e. f_a for $a = 1, .., 4$. The terms f_2 and f_3 are further decomposed into D conditionally independent factors

$$f_2(\boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{D} f_{2,i}(x_i, z_i) = \prod_{i=1}^{D} \left[(1 - z_i)\delta(x_i) + z_i \mathcal{N}\left(x_i \middle| 0, \tau_0\right)\right], \tag{15}$$

$$f_3(\boldsymbol{z}, \boldsymbol{\gamma}) = \prod_{i=1}^{D} f_{3,i}(z_i, \gamma_i) = \prod_{i=1}^{D} \mathrm{Ber}\left(z_i \middle| \phi\left(\gamma_i\right)\right) \tag{16}$$

The idea is then to approximate each term in the true posterior density, i.e. f_a, by simpler terms, i.e. $\tilde{f}_a$ for $a = 1, .., 4$. The resulting approximation $Q\left(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}\right)$ then becomes

$$Q\left(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}\right) = \frac{1}{Z_{EP}} \prod_{a=1}^{4} \tilde{f}_a\left(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}\right). \tag{17}$$

The terms $\tilde{f}_1$ and $\tilde{f}_4$ can be computed exact. In fact, $\tilde{f}_4$ is simply equal to the prior over $\boldsymbol{\gamma}$ and $\tilde{f}_1$ is a multivariate Gaussian distribution with mean $\tilde{\boldsymbol{m}}_1$ and covariance matrix $\tilde{\boldsymbol{V}}_1$ determined by $\tilde{\boldsymbol{V}}_1^{-1}\tilde{\boldsymbol{m}}_1 = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{y}$ and $\tilde{\boldsymbol{V}}_1^{-1} = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{A}$. Therefore, we only have to approximate the factors $\tilde{f}_2$ and $\tilde{f}_3$ using EP. Note that the exact term f_1 is a distribution of $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$, whereas the approximate term $\tilde{f}_1$ is a function of $\boldsymbol{x}$ that depends on $\boldsymbol{y}$ through $\tilde{\boldsymbol{m}}_1$ and $\tilde{\boldsymbol{V}}_1$ etc. In order to take full advantage of the structure of the true posterior distribution, we will further assume that the terms $\tilde{f}_2$ and $\tilde{f}_3$ also are decomposed into D independent factors.

The EP scheme provides great flexibility in the choice of the approximating factors. This choice is a trade-off between analytical tractability and sufficient flexibility for capturing the important characteristics of the true density. Due to the product over the binary support variables $\{z_i\}$ for $i = 1, .., D$, the true density is highly multimodal. Finally, f_2 couples the variables $\boldsymbol{x}$ and $\boldsymbol{z}$, while f_3 couples the variables $\boldsymbol{z}$ and $\boldsymbol{\gamma}$. Based on these observations, we choose $\tilde{f}_2$ and $\tilde{f}_3$ to have the following forms

$$\tilde{f}_2\left(\boldsymbol{x}, \boldsymbol{z}\right) \propto \prod_{i=1}^{D} \mathcal{N}\left(x_i \middle| \tilde{m}_{2,i}, \tilde{v}_{2,i}\right) \prod_{i=1}^{D} \mathrm{Ber}\left(z_i \middle| \phi\left(\tilde{\gamma}_{2,i}\right)\right) = \mathcal{N}\left(\boldsymbol{x} \middle| \tilde{\boldsymbol{m}}_2, \tilde{\boldsymbol{V}}_2\right) \prod_{i=1}^{D} \mathrm{Ber}\left(z_i \middle| \phi\left(\tilde{\gamma}_{2,i}\right)\right),$$

$$\tilde{f}_3\left(\boldsymbol{z}, \boldsymbol{\gamma}\right) \propto \prod_{i=1}^{D} \mathrm{Ber}\left(z_i \middle| \phi\left(\tilde{\gamma}_{3,i}\right)\right) \prod_{i=1}^{D} \mathcal{N}\left(\gamma_i \middle| \tilde{\mu}_{3,i}, \tilde{\sigma}_{3,i}\right) = \mathcal{N}\left(\boldsymbol{\gamma} \middle| \tilde{\boldsymbol{\mu}}_3, \tilde{\boldsymbol{\Sigma}}_3\right) \prod_{i=1}^{D} \mathrm{Ber}\left(z_i \middle| \phi\left(\tilde{\gamma}_{2,i}\right)\right),$$

where $\tilde{\boldsymbol{m}}_2 = [\tilde{m}_{2,1}, .., \tilde{m}_{2,D}]^T$, $\tilde{\boldsymbol{V}}_2 = \mathrm{diag}\left(\tilde{v}_{2,1}, ..., \tilde{v}_{2,D}\right)$ and analogously for $\tilde{\boldsymbol{\mu}}_3$ and $\tilde{\boldsymbol{\Sigma}}_3$. These choices lead to a joint variational approximation $Q(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma})$ of the form

$$Q\left(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}\right) = \mathcal{N}\left(\boldsymbol{x} \middle| \tilde{\boldsymbol{m}}, \tilde{\boldsymbol{V}}\right) \prod_{i=1}^{D} \mathrm{Ber}\left(z_i \middle| g\left(\tilde{\gamma}_i\right)\right) \mathcal{N}\left(\boldsymbol{\gamma} \middle| \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right), \tag{18}$$

where the joint parameters are given by

$$\tilde{\boldsymbol{V}} = \left(\tilde{\boldsymbol{V}}_1^{-1} + \tilde{\boldsymbol{V}}_2^{-1}\right)^{-1}, \qquad \tilde{\boldsymbol{m}} = \tilde{\boldsymbol{V}}\left(\tilde{\boldsymbol{V}}_1^{-1}\tilde{\boldsymbol{m}}_1 + \tilde{\boldsymbol{V}}_2^{-1}\tilde{\boldsymbol{m}}_2\right) \tag{19}$$

$$\tilde{\boldsymbol{\Sigma}} = \left(\tilde{\boldsymbol{\Sigma}}_3^{-1} + \tilde{\boldsymbol{\Sigma}}_4^{-1}\right)^{-1}, \qquad \tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}}\left(\tilde{\boldsymbol{\Sigma}}_3^{-1}\tilde{\boldsymbol{\mu}}_3 + \tilde{\boldsymbol{\Sigma}}_4^{-1}\tilde{\boldsymbol{\mu}}_4\right) \tag{20}$$

$$\tilde{\gamma}_j = \phi^{-1}\left[\left(\frac{\left(1 - \phi(\tilde{\gamma}_{2,j})\right)\left(1 - \phi(\tilde{\gamma}_{3,j})\right)}{\phi(\tilde{\gamma}_{2,j})\phi(\tilde{\gamma}_{3,j})} + 1\right)^{-1}\right], \quad \forall j \in \{1, .., D\}. \tag{21}$$

where $\phi^{-1}(x)$ is the probit function. The function in eq. (21) amounts to computing the product of two Bernoulli densities parametrized using $\phi\left(\cdot\right)$.

- Initialize approximation terms $\tilde{f}_a$ for $a = 1, 2, 3, 4$ and Q
- Repeat until stopping criteria
  - For each $\tilde{f}_{2,i}$:
    - Compute cavity distribution: $Q^{\setminus 2,i} \propto \frac{Q}{\tilde{f}_{2,i}}$
    - Minimize: $\mathrm{KL}\big(f_{2,i}Q^{\setminus 2,i} \big|\big| Q^{2,\mathrm{new}}\big)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{2,i} \propto \frac{Q^{2,\mathrm{new}}}{Q^{\setminus 2,i}}$ to update parameters $\tilde{m}_{2,i}, \tilde{v}_{2,i}$ and $\tilde{\gamma}_{2,i}$.
  - Update joint approximation parameters: $\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\gamma}}$
  - For each $\tilde{f}_{3,i}$:
    - Compute cavity distribution: $Q^{\setminus 3,i} \propto \frac{Q}{\tilde{f}_{3,i}}$
    - Minimize: $\mathrm{KL}\big(f_{3,i}Q^{\setminus 3,i} \big|\big| Q^{3,\mathrm{new}}\big)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{3,i} \propto \frac{Q^{3,\mathrm{new}}}{Q^{\setminus 3,i}}$ to update parameters $\tilde{\mu}_{3,i}, \tilde{\sigma}_{3,i}$ and $\tilde{\gamma}_{3,i}$
  - Update joint approximation parameters: $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\gamma}}$

Figure 2: Proposed algorithm for approximating the joint posterior distribution over $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{\gamma}$.

3.1 The EP algorithm

Consider the update of the term $\tilde{f}_{a,i}$ for a given a and a given i, where $\tilde{f}_a = \prod_i \tilde{f}_{a,i}$. This update is performed by first removing the contribution of $\tilde{f}_{a,i}$ from the joint approximation by forming the so-called cavity distribution

$$Q^{\setminus a,i} \propto \frac{Q}{\tilde{f}_{a,i}} \tag{22}$$

followed by the minimization of the Kullbach-Leibler [17] divergence between $f_{a,i}Q^{\setminus a,i}$ and $Q^{a,\mathrm{new}}$ w.r.t. $Q^{a,\mathrm{new}}$. For distributions within the exponential family, minimizing this form of KL divergence amounts to matching moments between $f_{a,i}Q^{\setminus 2,i}$ and $Q^{a,\mathrm{new}}$ [15]. Finally, the new update of $\tilde{f}_{a,i}$ is given by

$$\tilde{f}_{a,i} \propto \frac{Q^{a,\mathrm{new}}}{Q^{\setminus a,i}}. \tag{23}$$

After all the individual approximation terms $\tilde{f}_{a,i}$ for $a = 1, 2$ and $i = 1, .., D$ have been updated, the joint approximation is updated using eq. (19)-(21). To minimize the computational load, we use parallel updates of $\tilde{f}_{2,i}$ [8] followed by parallel updates of $\tilde{f}_{3,i}$ rather than the conventional sequential update scheme. Furthermore, due to the fact that $\tilde{f}_2$ and $\tilde{f}_3$ factorizes, we only need the marginals of the cavity distributions $Q^{\setminus a,i}$ and the marginals of the updated joint distributions $Q^{a,\mathrm{new}}$ for $a = 2, 3$.

Computing the cavity distributions and matching the moments are tedious, but straightforward. The moments of $f_{a,i}Q^{\setminus 2,i}$ require evaluation of the zeroth, first and second order moment of the distributions of the form $\phi(\gamma_i)\mathcal{N}\left(\gamma_i|\mu_i, \Sigma_{ii}\right)$. Derivation of analytical expressions for these moments can be found in [11]. See the supplementary material for more details. The proposed algorithm is summarized in figure 2. Note, that the EP framework also provides an approximation of the marginal likelihood [11], which can be useful for learning the hyperparameters of the model. Furthermore, the proposed inference scheme can easily be extended to the MMV formulation eq. (8)-(10) by introducing a $\tilde{f}^t_{2,i}$ for each time step $t = 1, .., T$.

3.2 Computational details

Most linear inverse problems of practical interest are high dimensional, i.e. D is large. It is therefore of interest to simplify the computational complexity of the algorithm as much as possible. The dominating operations in this algorithm are the inversions of the two $D \times D$ covariance matrices in eq. (19) and eq. (20), and therefore the algorithm scales as $\mathcal{O}\left(D^3\right)$. But $\tilde{\boldsymbol{V}}_1$ has low rank and $\tilde{\boldsymbol{V}}_2$ is diagonal, and therefore we can apply the Woodbury matrix identity [18] to eq. (19) to get

$$\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}_2 - \tilde{\boldsymbol{V}}_2 \boldsymbol{A}^T \left(\sigma_o^2 \boldsymbol{I} + \boldsymbol{A}\tilde{\boldsymbol{V}}_2 \boldsymbol{A}^T\right)^{-1} \boldsymbol{A}\tilde{\boldsymbol{V}}_2. \tag{24}$$

For $N < D$, this scales as $\mathcal{O}\left(ND^2\right)$, where N is the number of observations. Unfortunately, we cannot apply the same identity to the inversion in eq. (20) since $\tilde{\boldsymbol{\Sigma}}_4$ has full rank and is non-diagonal in general. The eigenvalue spectrum of many prior covariance structures of interest, i.e. simple neighbourhoods etc., decay relatively fast. Therefore, we can approximate $\boldsymbol{\Sigma}_0$ with a low rank approximation $\boldsymbol{\Sigma}_0 \approx \boldsymbol{P}\Lambda\boldsymbol{P}^T$, where $\Lambda \in \mathbb{R}^{R\times R}$ is a diagonal matrix of the R largest eigenvalues and $\boldsymbol{P} \in \mathbb{R}^{D\times R}$ is the corresponding eigenvectors. Using the R-rank approximation, we can now invoke the Woodbury matrix identity again to get:

$$\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\Sigma}}_3 + \tilde{\boldsymbol{\Sigma}}_3 \boldsymbol{P} \left(\Lambda + \boldsymbol{P}^T \tilde{\boldsymbol{\Sigma}}_3 \boldsymbol{P}\right)^{-1} \boldsymbol{P}^T \tilde{\boldsymbol{\Sigma}}_3. \tag{25}$$

Similarly, for $R < D$, this scales as $\mathcal{O}\left(RD^2\right)$. Another better approach that preserves the total variance would be to use probabilistic PCA [19] to approximate $\boldsymbol{\Sigma}_0$. A third alternative is to consider other structures for $\boldsymbol{\Sigma}_0$, which facilitate fast matrix inversions such as block structures and Toeplitz structures. Numerical issues can arise in EP implementations and in order to avoid this, we use the same precautions as described in [8].

4 Numerical experiments

This section describes a series of numerical experiments that have been designed and conducted in order to investigate the properties of the proposed algorithm.

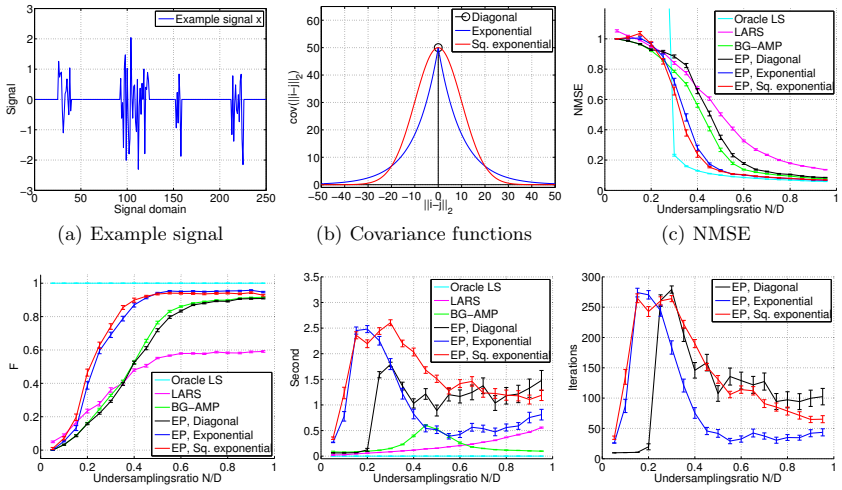
4.1 Experiment 1

The first experiment compares the proposed method to the LARS algorithm [20] and to the BG-AMP method [21], which is an approximate message passing-based method for the spike and slab model. We also compare the method to an "oracle least squares estimator" that knows the true support of the solutions. We generate 100 problem instances from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}$, where the solutions vectors have been sampled from the proposed prior using the kernel $\Sigma_{i,j} = 50\exp(-||i-j||_2^2/(2\cdot 10^2))$, but constrained to have a fixed sparsity level of the $K/D = 0.25$. That is, each solution $\boldsymbol{x}_0$ has the same number of non-zero entries, but different sparsity patterns. We vary the degree of undersampling from $N/D = 0.05$ to $N/D = 0.95$. The elements of $\boldsymbol{A} \in \mathbb{R}^{N\times 250}$ are i.i.d Gaussian and the columns of $\boldsymbol{A}$ have been scaled to unit ℓ_2-norm. The SNR is fixed at 20dB. We apply the four methods to each of the 100 problems, and for each solution we compute the Normalized Mean Square Error (NMSE) between the true signal $\boldsymbol{x}_0$ and the estimated signal $\hat{\boldsymbol{x}}$ as well as the F-measure:

$$\text{NMSE} = \frac{||\boldsymbol{x}_0 - \hat{\boldsymbol{x}}||_2}{||\boldsymbol{x}_0||_2} \qquad\qquad F = 2\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}, \tag{26}$$

where precision and recall are computed using a MAP estimate of the support. For the structured spike and slab method, we consider three different covariance structures: $\Sigma_{ij} = \kappa\cdot\delta(i-j)$, $\Sigma_{ij} = \kappa\exp(-||i-j||_2/s)$ and $\Sigma_{ij} = \kappa\exp(-||i-j||_2^2/(2s^2))$ with parameters $\kappa = 50$ and $s = 10$. In each case, we use a $R = 50$ rank approximation of Σ. The average results are shown in figures 3(a)-(f). Figure (a) shows an example of one of the sampled vectors $\boldsymbol{x}_0$ and figure (b) shows the three covariance functions.

From figure 3(c)-(d), it is seen that the two EP methods with neighbour correlation are able to improve the phase transition point. That is, in order to obtain a reconstruction



(a) Example signal (b) Covariance functions (c) NMSE

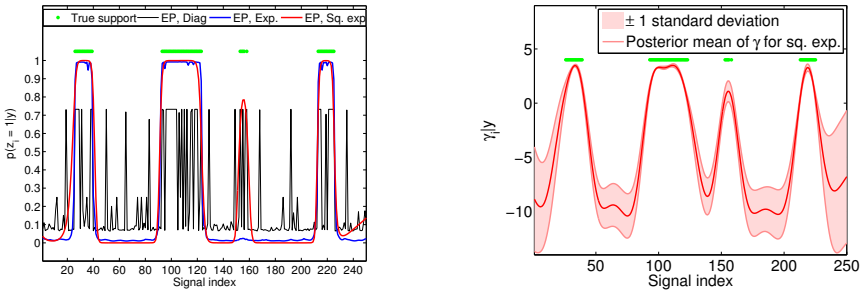
(d) F-measure (e) Run times (f) Iterations

Figure 3: Illustration of the benefit of modelling the additional structure of the sparsity pattern. 100 problem instances are generated using the linear measurement model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$, where elements of $\boldsymbol{A} \in \mathbb{R}^{N\times 250}$ are i.i.d Gaussian and the columns are scaled to unit ℓ_2-norm. The solutions $\boldsymbol{x}_0$ are sampled from the prior in eq. (5) with hyperparameters $\Sigma_{ij} = 50\exp\left[-||i-j||_2 / \left(2\cdot 10^2\right)\right]$ and a fixed level of sparsity of $K/D = 0.25$. For EP methods, the $\boldsymbol{\Sigma}_0$ matrix is approximated using a rank 50 matrix. SNR is fixed at 20dB.

of the signal such that $F \approx 0.8$, EP with diagonal covariance and BG-AMP need an undersamplingratio of $N/D \approx 0.55$, while the EP methods with neighbour correlation only need $N/D \approx 0.35$ to achieve $F \approx 0.8$. For this specific problem, this means that utilizing the neighbourhood structure allows us to reconstruct the signal with 50 fewer observations. Note that, the reconstruction using the exponential covariance function does also improve the result even if the true underlying covariance structure corresponds to a squared exponential function. Furthermore, we see similar performance of BG-AMP and EP with a diagonal covariance matrix. This is expected for problems where A_{ij} is drawn iid as assumed in BG-AMP. However, the price of the improved phase transition is clear from figure 3(e). The proposed algorithm has significantly higher computational complexity than BG-AMP and LARS. Figure 4(a) shows the posterior mean of $\boldsymbol{z}$ for the signal shown in figure 3(a). Here it is seen that the two models with neighbour correlation provide a better approximation to the posterior activation probabilities. Figure 4(b) shows the posterior mean of $\boldsymbol{\gamma}$ for the model with the squared exponential kernel along with $\pm$ one standard deviation.

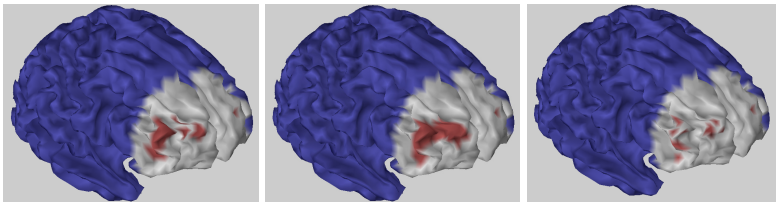
4.2 Experiment 2

In this experiment we consider an application of the MMV formulation as given in eq. (8)-(10), namely EEG source localization with synthetic sources [22]. Here we are interested in localizing the active sources within a specific region of interest on the cortical surface (grey area on figure 5(a)). To do this, we now generate a problem instance of $\boldsymbol{Y} = \boldsymbol{A}_{\text{EEG}}\boldsymbol{X}_0 + \boldsymbol{E}$ using the procedure as described in experiment 1, where $\boldsymbol{A}_{\text{EEG}} \in \mathbb{R}^{128\times 800}$ is now a submatrix of a real EEG forward matrix corresponding to the grey area on the figure. The condition number of $\boldsymbol{A}_{\text{EEG}}$ is $\approx 8\cdot 10^{15}$. The true sources $\boldsymbol{X}_0 \in \mathbb{R}^{800\times 20}$ are sampled from the structured spike and slab prior in eq. (8) using a squared exponential kernel with parameters $A = 50$, $s = 10$ and $T = 20$. The number of active sources is 46, i.e. $\boldsymbol{x}$ has 46 non-zero rows. SNR is fixed to $20dB$. The true sources are shown in figure 5(a). We now use the EP algorithm to recover the sources using the true prior, i.e. squared exponential kernel and



(a) (b)

Figure 4: (a) Marginal posterior means over $\boldsymbol{z}$ obtained using the structured spike and slab model for the signal in figure 3(a). The experiment set-up is the as described in figure 3, except the undersamplingsratio is fixed to $N/D = 0.5$. (b) The posterior mean of $\boldsymbol{\gamma}$ superimposed with $\pm$ one standard deviation. The green dots indicate the true support.



(a) True sources (b) EP, Sq. exponential (c) EP, Diagonal

Figure 5: Source localization using synthetic sources. The $\boldsymbol{A} \in \mathbb{R}^{128\times 800}$ is a submatrix (grey area) of a real EEG forward matrix. (a) True sources. (b) Reconstruction using the true prior , $F_{sq} = 0.78$. (c) Reconstruction using a diagonal covariance matrix, $F_{\text{diag}} = 0.34$.

the results are shown in figure 5(b). We see that the algorithm detects most of the sources correctly, even the small blob on the right hand side. However, it also introduces a small number of false positives in the neighbourhood of the true active sources. The resulting F-measure is $F_{sq} = 0.78$. Figure 5(c) shows the result of reconstructing the sources using a diagonal covariance matrix, where $F_{\text{diag}} = 0.34$. Here the BG-AMP algorithm is expected to perform poorly due to the heavy violation of the assumption of A_{ij} being Gaussian iid.

4.3 Experiment 3

We have also recreated the Shepp-Logan Phantom experiment from [2] with $D = 10^4$ unknowns, $K = 1723$ non-zero weights, $N = 2K$ observations and SNR $= 10dB$ (see supplementary material for more details). The EP method yields $F_{sq} = 0.994$ and $\text{NMSE}_{sq} = 0.336$ for this experiment, whereas BG-AMP yields $F = 0.624$ and NMSE $= 0.717$. For reference, the oracle estimator yields NMSE $= 0.326$.

5 Conclusion and outlook

We introduced the structured spike and slab model, which allows incorporation of a priori knowledge of the sparsity pattern. We developed an expectation propagation-based algorithm for Bayesian inference under the proposed model. Future work includes developing a scheme for learning the structure of the sparsity pattern and extending the algorithm to the multiple measurement vector formulation with slowly changing support.

References

[1] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *AISTATS*, pages 366–373, 2010.

[2] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery using markov random fields. In *NIPS*, Vancouver, B.C., Canada, 8–11 December 2008.

[3] M. Pontil, L. Baldassarre, and J. Mouro-Miranda. Structured sparsity models for brain decoding from fMRI data. *Proceedings - 2012 2nd International Workshop on Pattern Recognition in NeuroImaging, PRNI 2012*, pages 5–8, 2012.

[4] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society series b-methodological*, 58(1):267–288, 1996.

[5] T. J. Mitchell and J. Beauchamp. Bayesian variable selection in linear-regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[6] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal Of Computational And Graphical Statistics*, 22(2):231–245, 2013.

[7] G. Obozinski, J. P. Vert, and L. Jacob. Group lasso with overlap and graph lasso. *ACM International Conference Proceeding Series*, 382:–, 2009.

[8] D. Hernandez-Lobato, J. Hernandez-Lobato, and P. Dupont. Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal Of Machine Learning Research*, 14:1891–1945, 2013.

[9] L. Yu, H. Sun, J. P. Barbot, and G. Zheng. Bayesian compressive sensing for cluster structured sparse signals. *Signal Processing*, 92(1):259 – 269, 2012.

[10] M. Van Gerven, B. Cseke, R. Oostenveld, and T. Heskes. Bayesian source localization with the multivariate laplace prior. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1901–1909. Curran Associates, Inc., 2009.

[11] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[12] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, pages 2477–2488, 2005.

[13] D. P. Wipf and B. D. Rao. An empirical bayesian strategy for solving the, simultaneous sparse approximation problem. *IEEE Transactions On Signal Processing*, 55(7):3704–3716, 2007.

[14] J. Ziniel and P. Schniter. Dynamic compressive sensing of time-varying signals via approximate message passing. *IEEE Transactions On Signal Processing*, 61(21):5270–5284, 2013.

[15] T. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 362–369, San Francisco, CA, 2001. Morgan Kaufmann.

[16] M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

[17] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[18] K. B. Petersen and M. S. Pedersen. The matrix cookbook. 2012.

[19] M. E Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999.

[20] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[21] P. Schniter and J. Vila. Expectation-maximization gaussian-mixture approximate message passing. *2012 46th Annual Conference on Information Sciences and Systems, CISS 2012*, pages –, 2012.

[22] S. Baillet, J. C. Mosher, and R. M. Leahy. Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, 18(6):14–30, 2001.

Appendix B

Spatio-temporal Spike and Slab Priors for Multiple Measurement Vector Problems

---

B Andersen, M. R., Winther, O. and Hansen, L. K. (2015), 'Spatio-temporal spike and slab priors for MMV problems'. Signal Processing with Adaptive Sparse Structured Representations (SPARS) 2015, 6 pages

Spatio-temporal spike and slab priors for MMV problems

Michael Riis Andersen, Ole Winther & Lars Kai Hansen
DTU Compute, Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark
Email: {miri, olwi, lkh}@dtu.dk

***Abstract*—We are interested in solving the multiple measurement vector (MMV) problem for instances, where the underlying sparsity pattern exhibit spatio-temporal structure motivated by the electroencephalogram (EEG) source localization problem. We propose a probabilistic model that takes this structure into account by generalizing the structured spike and slab prior and the associated Expectation Propagation inference scheme. Based on numerical experiments, we demonstrate the viability of the model and the approximate inference scheme.**

I. Introduction

The multiple measurement vector problem (MMV) [1] is given by:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times D}$ is the forward matrix, $\mathbf{Y} \in \mathbb{R}^{N\times T}$ is the measurement matrix, $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_T\end{bmatrix} \in \mathbb{R}^{D\times T}$ is the desired solution and $\mathbf{E} \in \mathbb{R}^{N\times T}$ is a matrix of corruptive noise. We are interested in finding sparse solutions to eq. (1) in the ill-posed regime, where $N < D$. Furthermore, the sparsity pattern of $\mathbf{X}$ is assumed to have certain structural properties. In particular, we are considering problems where the sparsity pattern exhibit spatio-temporal structure as in EEG source localization [2], [3] or in background subtraction in computer vision [4]. Let $\mathbf{z}_t$ be an indicator for the support of $\mathbf{x}_t$, i.e. $\mathbf{z}_t = \mathbb{I}[\mathbf{x}_t \neq 0]$, then $\mathbf{z}_t$ is assumed to be spatially correlated. Furthermore, we assume that the support vectors $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T$ slowly evolve through time as well - rendering the joint sparsity assumption invalid [5].

The main contribution of this work is to propose a model for spatio-temporal sparsity patterns by extending the structured spike and slab prior [6] to account for temporal evolution of the sparsity pattern as well. Furthermore, we demonstrate the benefits of the model through numerical experiments.

A. Related work

The field of structured sparsity has received a great deal of attention in the recent years. In this section we highlight some of the related work, but this list is by no means exhaustive. The LASSO-community have introduced the Group and Graph LASSO methods, which generalize the standard ℓ_1-norm minimization approach to promote different kinds of structured sparsity [7]. In the probabilistic setting, the standard workhorse for sparsity is the so-called spike and slab prior [8]. This has also been generalized to model group sparsity [9] and cluster sparsity [10]. In the context of compressed sensing [11], Cevher et al. [4] used a Markov random field to enforce spatially correlated sparsity patterns, whereas Ziniel et al. used binary Markov chains to model temporally correlated sparsity patterns [12].

II. The structured spike and slab prior

In this section we briefly introduce the conventional spike and slab prior [8] and the structured spike and slab prior [6] before we move on to the spatio-temporal spike and slab prior on the next section. The conventional spike and slab prior decomposes each $x_{i,t}$ as a product of a binary variable $z_{i,t}$ and a real number $c_{i,t}$, i.e. $x_{i,t} = z_{i,t}c_{i,t}$, where $z_{i,t} \sim \text{Ber}(p_0)$ and $c_{i,t} \sim \mathcal{N}(0, \tau_0)$ for $i \in \{1, 2, .., D\}$ and $t \in \{1, 2, .., T\}$. The structured spike and slab prior generalized this formulation by imposing structure on the binary variable for each time t as follows

$$p(\mathbf{z}_t|\phi(\boldsymbol{\gamma}_t)) = \prod_{i=1}^{D} \text{Ber}\left(z_{i,t}|\phi(\gamma_{i,t})\right), \tag{2}$$

$$p(\boldsymbol{\gamma}_t) = \mathcal{N}\left(\boldsymbol{\gamma}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right), \tag{3}$$

where the Bernoulli probabilities are parametrized using the standard normal CDF $\phi : \mathbb{R} \to (0, 1)$. The hyperparameters $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ encode the prior belief of the support for time t. Specifically, the prior mean value $\boldsymbol{\mu}_t$ controls the prior belief of the number of non-zero variables and the covariance matrix $\boldsymbol{\Sigma}_t$ determines the prior correlation of the support at time t. Thus, we can impose structure on the binary support variables $\mathbf{z}_t$ by means of imposing generic covariance functions on $\boldsymbol{\gamma}$. For example, say we choose $\Sigma_{i,j}$ to be the squared exponential covariance function, then the resulting prior distribution will promote sparsity patterns where neighbouring support variables have the same state. Under the other hand, when $\boldsymbol{\Sigma}$ is diagonal, we recover the independent spike and slab prior.

The marginal prior probability of the $x_{i,t}$ being non-zero is given by

$$\begin{aligned} p(z_{i,t} = 1) &= \int p(z_{i,t} = 1|\gamma_{i,t})p(\gamma_{i,t})\mathrm{d}\gamma_{i,t} \\ &= \int \phi(\gamma_{i,t})\mathcal{N}\left(\gamma_{i,t}|\mu_{i,t}, \Sigma_{ii,t}\right)\mathrm{d}\gamma_{i,t} \\ &= \phi\left(\frac{\mu_{i,t}}{\sqrt{1+\Sigma_{ii,t}}}\right). \end{aligned} \tag{4}$$

Thus, if the prior on $\boldsymbol{\gamma}_t$ has zero mean, then the prior belief of $p(z_{i,t})$ is unbiased, i.e. $p(z_{i,t}) = 0.5$. On the other hand, if $\mu_{i,t}$ is negative, the prior belief of $z_{i,t}$ is biased towards zero and vice versa.

III. The spatio-temporal spike and slab prior

In this section we describe the temporal extension of the structured spike and slab prior. Instead of considering $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ as fixed hyperparameters, we propose to impose a prior on $\boldsymbol{\Gamma} = \begin{bmatrix}\boldsymbol{\gamma}_1 & \boldsymbol{\gamma}_2 & \dots & \boldsymbol{\gamma}_T\end{bmatrix}$ to model problems where the support of the solution $\mathbf{X}$ changes over time. In particular, we impose a first order process Markov process on $\boldsymbol{\Gamma}$ to model the slowly changing sparsity pattern

$$p\left(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\gamma}_t|\left(1-\alpha\right)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right), \tag{5}$$

where the hyperparameters α and β control the temporal correlation and the "innovation" of the process, respectively. Furthermore, we assume that the prior distribution on $\boldsymbol{\gamma}_1$ is given by

$$p(\boldsymbol{\gamma}_1) = \mathcal{N}\left(\boldsymbol{\gamma}_1|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right). \tag{6}$$

Under these assumptions the marginal distribution of $\boldsymbol{\gamma}_2$ becomes

$$\begin{aligned} p(\boldsymbol{\gamma}_2) &= \int p\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\gamma}_1\right) p(\boldsymbol{\gamma}_1) \mathrm{d}\boldsymbol{\gamma}_1 \\ &= \mathcal{N}\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\mu}_0, \left(\alpha^2 + \beta\right) \boldsymbol{\Sigma}_0\right). \end{aligned} \tag{7}$$

Therefore, it follows by induction that if α and β satisfy $\alpha^2 + \beta = 1$, then the marginal distribution of $\boldsymbol{\gamma}_t$ is $p(\boldsymbol{\gamma}_t) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ for all t. Furthermore, we also see that for $\alpha = 1$ and $\beta = 0$, the prior reduces to the structured spike and slab prior in the joint sparsity setting. In the other extreme, at $\alpha = 0$ and $\beta = 1$, the prior reduces to the structured spike and slab prior in the time-independent setting. Hence, the spatio-temporal spike and slab prior can be seen as a generalization of the two extreme cases.

This choice of model is also motivated by the fact that the first order structure in the temporal dimension gives rise to an inference scheme that scales linearly in the number of time steps T as we will see in the next section.

IV. Bayesian Inference using the spatio-temporal spike and slab prior

The goal of this section is to describe an inference procedure for solving the problem in eq. (1) using the proposed prior in a fully Bayesian setting. We combine the spatio-temporal spike and slab prior with a time-independent isotropic Gaussian noise model of the form

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{y}_t \middle| \mathbf{A}\mathbf{x}_t, \sigma_0^2 \mathbf{I}\right). \tag{8}$$

This gives rise to the following joint distribution

$$\begin{aligned} p(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}) = & \underbrace{\prod_{t=1}^{T} \mathcal{N}\left(\mathbf{y}_t \middle| \mathbf{A}\mathbf{x}_t, \sigma_0^2 \mathbf{I}\right)}_{f_1(\mathbf{X})} \\ & \underbrace{\prod_{t=1}^{T} \prod_{i=1}^{D} \left[(1 - z_{i,t}) \delta(x_{i,t}) + z_{i,t} \mathcal{N}\left(x_{i,t} \middle| 0, \tau_0\right)\right]}_{f_2(\mathbf{X}, \mathbf{Z})} \\ & \underbrace{\prod_{t=1}^{T} \prod_{i=1}^{D} \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\gamma_{i,t}\right)\right)}_{f_3(\mathbf{Z}, \boldsymbol{\Gamma})} \\ & \underbrace{\mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \prod_{t=2}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| (1-\alpha)\boldsymbol{\mu}_0 + \alpha \boldsymbol{\gamma}_{t-1}, \beta \boldsymbol{\Sigma}_0\right)}_{f_4(\boldsymbol{\Gamma})} \end{aligned} \tag{9}$$

The desired posterior distribution $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}|\mathbf{Y})$ is obtained from Bayes' Rule [13]. Unfortunately, this posterior distribution is intractable due to the product of mixtures and hence, we have to settle for approximate inference. Specifically, we use Expectation Propagation [14]–[16] for approximate inference by extending the proposed inference scheme in [6].

A. Approximate Inference using Expectation Propagation

Expectation propagation (EP) is an iterative deterministic method for approximating probability distributions using simpler distributions from the exponential family. As indicated in eq. (9), the exact posterior can be decomposed as follows

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}|\mathbf{Y}) = & \frac{1}{Z} \prod_{t=1}^{T} f_{1,t}\left(\mathbf{x}_t\right) \prod_{t=1}^{T} \prod_{i=1}^{D} f_{2,i,t}\left(x_{i,t}, z_{i,t}\right) \\ & \prod_{t=1}^{T} \prod_{i=1}^{D} f_{3,i,t}\left(z_{i,t}, \gamma_{i,t}\right) \prod_{t=1}^{T} f_{4,t}\left(\boldsymbol{\gamma}_t\right), \end{aligned} \tag{10}$$

where $Z = p(\mathbf{Y})$ is the normalization constant. Moreover, note that each factor in the decomposition only depends on a subset of the variables in the model, i.e. $f_{2,i,t}$ depends only on the variables $x_{i,t}$ and $z_{i,t}$ and so on and so forth. The EP framework takes advantage of this decomposition by approximating each factor in eq. (10) with a distribution from the exponential family. First we describe the functional form of the approximation and then we briefly explain how to estimate the parameters of the approximation using the EP algorithm.

Let $\tilde{f}_{1,t}$ denote the approximation of $f_{1,t}$ etc. First, we note that each of the factors in the first term, i.e. $f_{1,t}$ for all t, are already a member of the exponential family and hence does not have to be approximated. Therefore, for each t we have

$$\tilde{f}_{1,t}\left(\mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_t \middle| \tilde{\mathbf{m}}_{1,t}, \tilde{\boldsymbol{V}}_{1,t}\right), \tag{11}$$

where the parameters are determined by $\tilde{\boldsymbol{V}}_{1,t}^{-1} \tilde{\boldsymbol{m}}_{1,t} = \frac{1}{\sigma_0^2} \mathbf{A}^T \mathbf{y}_t$ and $\tilde{\boldsymbol{V}}_{1,t}^{-1} = \frac{1}{\sigma_0^2} \mathbf{A}^T \mathbf{A}$. Note that the exact term $f_{1,t}$ is a distribution on $\mathbf{y}_t$ conditioned on $\mathbf{x}_t$, whereas the approximate term $\tilde{f}_{1,t}$ is a function of $\mathbf{x}_t$ that depends on $\mathbf{y}_t$ through $\tilde{\mathbf{m}}_{1,t}$ and $\tilde{\boldsymbol{V}}_{1,t}$ etc. Next, we turn to the factors in the second term, i.e. $f_{2,i,t}$. Since each of these factors depends on $x_{i,t}$ and $z_{i,t}$, we choose $\tilde{f}_{2,i,t}$ to be

$$\tilde{f}_{2,i,t} = \mathcal{N}\left(x_{i,t} \middle| \tilde{m}_{2,i,t}, \tilde{V}_{2,i,t}\right) \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\tilde{\gamma}_{2,i,t}\right)\right), \tag{12}$$

where $\tilde{m}_{2,i,t}$, $\tilde{V}_{2,i,t}$ and $\tilde{\gamma}_{2,i,t}$ have to determined using the EP algorithm. Based on similar arguments $\tilde{f}_{3,i,t}$ and $\tilde{f}_{4,t}$ are chosen as follows

$$\tilde{f}_{3,i,t} = \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\tilde{\gamma}_{3,i,t}\right)\right) \mathcal{N}\left(\gamma_{3,i,t} \middle| \tilde{\mu}_{3,i,t}, \tilde{\Sigma}_{3,i,t}\right), \tag{13}$$

$$\tilde{f}_{4,t} = \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \tilde{\boldsymbol{\mu}}_{4,t}, \tilde{\boldsymbol{\Sigma}}_{4,t}\right). \tag{14}$$

Note that $f_{4,1}$ does not have to approximated either, it is simply $\tilde{f}_{4,1} = \mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$. Furthermore, note that the approximations to the factors $f_{4,t}$ for all t do not factorize w.r.t. $\gamma_{t,1}, \gamma_{t,2}, ...$ in order to capture potentially strong correlations in the support.

After specifying all the individual approximation terms, we derive the joint approximation of the desired posterior $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}|\mathbf{Y})$. Since the exponential family is closed under products, the approximate joint distribution has the following form

$$\begin{aligned} Q\left(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}\right) = & \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{x}_t \middle| \tilde{\mathbf{m}}_t, \tilde{\boldsymbol{V}}_t\right) \prod_{t=1}^{T} \prod_{i=1}^{D} \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\tilde{\gamma}_{i,t}\right)\right) \\ & \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t\right). \end{aligned} \tag{15}$$

Let $\mathbf{m}_{2,t} = [\tilde{m}_{2,1,t}, \tilde{m}_{2,2,t}, \ldots, \tilde{m}_{2,D,t}]^T$ and $\boldsymbol{V}_{2,t} = \mathrm{diag}\left(\tilde{V}_{2,1,t}, \tilde{V}_{2,2,t}, \ldots, \tilde{V}_{2,D,t}\right)$, and analogously for $\tilde{\boldsymbol{\mu}}_3$, $\tilde{\boldsymbol{\Sigma}}_3$

and $\boldsymbol{\gamma}_3$, then the parameters of the joint approximation are given by

$$\tilde{\boldsymbol{V}}_t = \left(\tilde{\boldsymbol{V}}_{1,t}^{-1} + \tilde{\boldsymbol{V}}_{2,t}^{-1}\right)^{-1}, \tag{16}$$

$$\tilde{\mathbf{m}}_t = \tilde{\boldsymbol{V}}_t\left(\tilde{\boldsymbol{V}}_{1,t}^{-1}\tilde{\mathbf{m}}_{1,t} + \tilde{\boldsymbol{V}}_{2,t}^{-1}\tilde{\mathbf{m}}_{2,t}\right), \tag{17}$$

$$\tilde{\boldsymbol{\Sigma}}_t = \left(\tilde{\boldsymbol{\Sigma}}_{3,t} + \tilde{\boldsymbol{\Sigma}}_{4,t}\right)^{-1}, \tag{18}$$

$$\tilde{\boldsymbol{\mu}}_t = \tilde{\boldsymbol{\Sigma}}_t\left(\tilde{\boldsymbol{\Sigma}}_{3,t}^{-1}\tilde{\boldsymbol{\mu}}_{3,t} + \tilde{\boldsymbol{\Sigma}}_{4,t}^{-1}\tilde{\boldsymbol{\mu}}_{4,t}\right), \tag{19}$$

$$\tilde{\gamma}_{i,t} = \phi^{-1}\left[\left(\frac{(1-\phi(\tilde{\gamma}_{2,i,t}))\,(1-\phi(\tilde{\gamma}_{3,i,t}))}{\phi(\tilde{\gamma}_{2,i,t})\phi(\tilde{\gamma}_{3,i,t})} + 1\right)^{-1}\right]. \tag{20}$$

The posterior covariance matrices $\tilde{\boldsymbol{V}}_t$ and $\tilde{\boldsymbol{\Sigma}}_t$ are (potentially) fully dense matrices, which makes the approximation able to cope with non-orthogonal forward matrices $\mathbf{A}$.

B. The Expectation Propagation Algorithm

In this section we describe how to compute the parameters of the individual approximations using the EP algorithm. The EP algorithm works by updating each of the individual approximation terms one by one until convergence. Consider the update of the term $\tilde{f}_{a,i,t}$ for a given a, i and t. The update is obtained by performing the following three steps of the EP algorithm. The first step is to remove the contribution of $\tilde{f}_{a,i,t}$ from the joint approximation in eq. (15) by forming the so-called cavity distribution

$$Q^{\setminus a,i,t} \propto \frac{Q}{\tilde{f}_{a,i,t}}. \tag{21}$$

In the next step we minimize the Kullbach-Leibler [13] divergence between $f_{a,i,t}Q^{\setminus a,i,t}$ and $Q^{a,t,\text{new}}$ w.r.t. $Q^{a,t,\text{new}}$. That is, we minimize $\text{KL}\left(\frac{1}{Z_{a,i,t}}f_{a,i,t}Q^{\setminus a,i,t}||Q^{a,t,\text{new}}\right)$, where $Z_{a,i,t}$ is the normalization constant of $f_{a,i,t}Q^{\setminus a,i,t}$. For distributions within the exponential family, minimizing this form of KL divergence amounts to matching moments between $f_{a,i,t}Q^{\setminus a,i,t}$ and $Q^{a,t,\text{new}}$ [14]. Finally, the third and last step is to compute the new update of $\tilde{f}_{a,i,t}$ as follows

$$\tilde{f}_{a,i,t} \propto \frac{Q^{a,t,\text{new}}}{Q^{\setminus a,i,t}}. \tag{22}$$

After the individual approximation terms $\tilde{f}_{a,i,t}$ for all i and t for a given a have been updated, the relavant parts of the joint approximation are updated using eq. (16)-(20). To minimize the computational load, we use parallel updates of $\tilde{f}_{2,i,t}$ [9] followed by parallel updates of $\tilde{f}_{3,i,t}$ rather than the conventional sequential update scheme. Furthermore, due to the fact that $\tilde{f}_2$ and $\tilde{f}_3$ factorizes w.r.t. both i and t, we only need the marginals of the cavity distributions $Q^{\setminus a,i,t}$, which simplifies the computations. Computing the cavity distributions and matching the moments are straightforward. However, when matching the moments, we are required to evaluate of the zero'th, first and second order moment of the distributions of the form $\phi(\gamma_i)\mathcal{N}\left(\gamma_i|\mu_i,\Sigma_{ii}\right)$. Derivation of analytical expressions for these moments can be found in the appendix to chapter 3 in [17].

The proposed EP algorithm is summarized in figure 1. The computational complexity of the algorithm is dominated by the matrix inversions in eq. (16) and (19). However, when $N < D$, the covariance matrices $\tilde{\boldsymbol{V}}_{1,t}$ have low rank and hence, eq. (16) can be carried out in $\mathcal{O}\left(ND^2\right)$ using the Matrix Inversion Lemma [18]. Therefore, the resulting inference scheme scales as $\mathcal{O}\left(TD^3\right)$, i.e. it scales linearly in the number of measurement vectors T.

- Initialize approximation terms $\tilde{f}_a$ for $a = 1, 2, 3, 4$ and Q
- Repeat until stopping criteria
  - For each $\tilde{f}_{2,i,t}$:
    - Compute cavity distribution: $Q^{\setminus 2,i,t} \propto \frac{Q}{\tilde{f}_{2,i,t}}$
    - Minimize: $\text{KL}\left(f_{2,i,t}Q^{\setminus 2,i,t}||Q^{2,t,\text{new}}\right)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{2,i,t} \propto \frac{Q^{2,t,\text{new}}}{Q^{\setminus 2,i,t}}$ to update parameters $\tilde{m}_{2,i,t}, \tilde{v}_{2,i,t}$ and $\tilde{\gamma}_{2,i,t}$.
  - Update joint approximation parameters: $\tilde{\mathbf{m}}, \tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\gamma}}$
  - For each $\tilde{f}_{3,i,t}$:
    - Compute cavity distribution: $Q^{\setminus 3,i,t} \propto \frac{Q}{\tilde{f}_{3,i,t}}$
    - Minimize: $\text{KL}\left(f_{3,i,t}Q^{\setminus 3,i,t}||Q^{3,t,\text{new}}\right)$ w.r.t. $Q^{3,t,\text{new}}$
    - Compute: $\tilde{f}_{3,i,t} \propto \frac{Q^{3,t,\text{new}}}{Q^{\setminus 3,i,t}}$ to update parameters $\tilde{\mu}_{3,i,t}, \tilde{\sigma}_{3,i,t}$ and $\tilde{\gamma}_{3,i,t}$
  - Update joint approximation parameters: $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\gamma}}$
  - For each $\tilde{f}_{4,t}$
    - Compute cavity distribution: $Q^{\setminus 4,t} \propto \frac{Q}{\tilde{f}_{4,t}}$
    - Minimize: $\text{KL}\left(f_{4,t}Q^{\setminus 4,i}||Q^{4,t,\text{new}}\right)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{4,t} \propto \frac{Q^{4,t,\text{new}}}{Q^{\setminus 4,t}}$ to update parameters $\tilde{m}_{4,t}, \tilde{v}_{4,t}$ and $\tilde{\gamma}_{4,t}$.
  - Update joint approximation parameters: $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$

Fig. 1. Proposed algorithm for approximating the joint posterior distribution over $\mathbf{X}, \mathbf{Z}$ and $\boldsymbol{\Gamma}$ conditioned on $\mathbf{Y}$.

C. Tuning of hyperparameters

The algorithm requires tuning of multiple hyperparameters for optimal performance. The Expectation Propagation framework provides a neat alternative to typical cross-validation schemes. Besides the approximation to the posterior distribution $P(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}|\mathbf{Y})$, EP also provides an approximation to the marginal likelihood $P(\mathbf{Y})$, which is very useful for model selection and tuning of hyperparameters [13]. The exact marginal likelihood is obtained by marginalizing out $\mathbf{X}, \mathbf{Z}$ and $\boldsymbol{\Gamma}$ from the joint distribution in eq. (9). The EP approximation to the marginal likelihood is obtained by substituting all the (scaled) individual approximation terms into the resulting formula. Finally, it is also possible to get closed form expression for the gradients of the marginal likelihood approximation w.r.t. to the hyperparameters [16], [17], which allows efficient tuning of the hyperparameters. However, a detailed treatment of the marginal likelihood approximation and its gradient w.r.t. hyperparameters are out of scope for this extended abstract.

V. Numerical Experiments

In order to investigate the properties of the proposed algorithm, we have designed and conducted two numerical experiments. The first experiment addresses the reconstruction performance of the algorithm, whereas the second experiment investigate the algorithm's robustness towards coherent forward models.

A. Experiment 1

To evaluate the proposed method, we have compared the method to several related solvers: BG-AMP[1] [19], DCS-AMP[2] [20], Spatial

[1]We used the implementation in GAMP-toolbox by Sundeep Rangan et al: http://gampmatlab.wikia.com/wiki/

[2]We used the implementation in the DCS-AMP-toolbox by Justin Ziniel: http://www2.ece.ohio-state.edu/ zinielj/dcs/

EP (implements the structured spike and slab prior) [6] and Spatial MMV EP. The BG-AMP method combines the conventional spike and slab prior with approximate message passing-based [21] inference. We include this method to have a baseline result without any structural assumptions on the sparsity pattern. The DCS-AMP can be seen as an extension of BG-AMP, which assumes that the sparsity pattern evolves slowly in time according to a binary Markov chain. The Spatial EP method assumes spatial correlation in the sparsity pattern, but no temporal correlation. Finally, the Spatial MMV method is similar to Spatial EP but with static sparsity across time, i.e. it assume joint sparsity across time.

To set up the first test we first sampled one realization of $\mathbf{Z}$ using eq. (2)-(5) with $D = 100$, $T = 100$, $\alpha = 0.99$ and $\beta = 1 - \alpha^2$, see figure 4(a). The average number of non-zero weights per column is fixed to 20. We note that the resulting sample exhibits the spatio-temporal structure as desired. Afterwards, we sample the nonzero coefficients in $\mathbf{X}$ from a standard normal distribution and from these we generate compressive measurements using eq. (1), where $A_{ij} \sim \mathcal{N}(0, 1/N)$, the SNR $= 10$dB and the undersampling ratio N/D is varied from from 0.05 to 0.95. To quantify the performance of the methods we use Normalized Mean Square Error (NMSE) between the true $\mathbf{X}$ and the estimated $\hat{\mathbf{X}}$ given by

$$NMSE = \frac{\sum_{i,t}\left(X_{i,t} - \hat{X}_{i,t}\right)^2}{\sum_{i,t} X_{i,t}^2}. \tag{23}$$

Furthermore, we evaluate each method's ability to recover the true support $\mathbf{Z}$ using the F-measure [22] based on a MAP estimate of the support $\hat{\mathbf{Z}}$,

$$F = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{24}$$

The results are averaged over 100 realizations of the noise $\mathbf{E}$ and non-zero coefficients in $\mathbf{X}$ and are shown in figures 2-3. It is seen that the proposed spatio-temporal method outperforms the other methods both in terms of NMSE and F-measure, but in general it is seen that richer prior assumptions on the support improves the results significantly. We also note that for very undersampled problems, the Spatial MMV EP method with static sparsity actually performs best. But as the undersampling ratio increases, all the other methods, including BG-AMP, outperforms it due to the very high bias of the model.

Figures 4(b)-4(f) shows the reconstructed support sets for the undersampling ratio $N/D = 0.4$. It is seen that DCS-AMP and Spatial EP, which models temporal and spatial structure, respectively, clearly outperforms BG-AMP. Furthermore, it is also seen that joint sparsity assumption (fig. 4(e)) is too restrictive for these kinds of signals. Again, we note that the spatio-temporal model gives superior results in terms of both F-measure and NMSE.

B. Experiment 2

The forward model $\mathbf{A}$ in the EEG source localization problem contains highly correlated columns, i.e. $\mathbf{A}$ is coherent. Therefore, it is of interest to investigate the proposed algorithm's robustness to coherent forward models. The set-up in this experiment is basically the same as for the first experiment, except that undersampling ratio is now fixed to $N/D = 0.4$ and the elements in the forward model A_{ij} are no longer Gaussian i.i.d. Instead we sample the rows of $\mathbf{A}$ from a correlated multivariate normal distribution, such that the

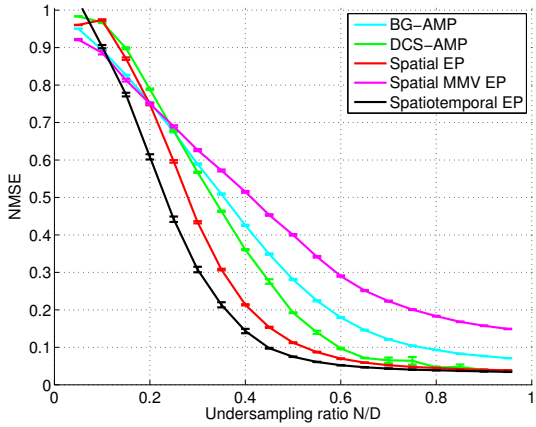


Fig. 2. Normalized mean square error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a), where $D = 100, T = 100$ and SNR $= 10dB$. The entries in $\mathbf{A}$ are Gaussian i.i.d, i.e. $A_{i,j} \sim \mathcal{N}(0, 1/N)$. The results are averaged over 100 realizations.

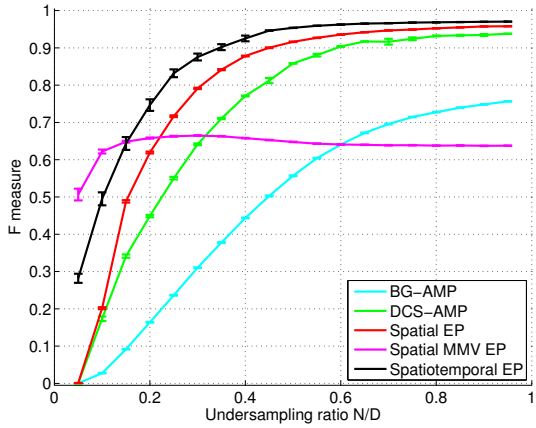


Fig. 3. F-measure error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a), where $D = 100, T = 100$ and SNR $= 10dB$. The entries in $\mathbf{A}$ are Gaussian i.i.d, i.e. $A_{i,j} \sim \mathcal{N}(0, 1/N)$. The results are averaged over 100 realizations.

columns of $\mathbf{A}$ will be correlated. In particular, the correlation of the i'th and j'th column of $\mathbf{A}$ is given by $r^{|i-j|}$. We compute the NMSE and F-measure as a function of the correlation r. Note that the BG-AMP and DCS-AMP methods are designed for Gaussian i.i.d forward. These two methods are therefore not expected to perform well in this experiment, but we include them for completeness. The results are averaged over 50 realizations and are shown in figures 5 and 6. The EP-based methods show some robustness to correlation in the columns of $\mathbf{A}$, but the performance does degrade gradually when we increase the correlation. In particular, when changing the correlation r from 0.05 to 0.95, the F-measure for the spatio-temporal method only drops from approximate 0.92 to 0.89, but the NMSE increases from approximately 0.15 to 0.45.

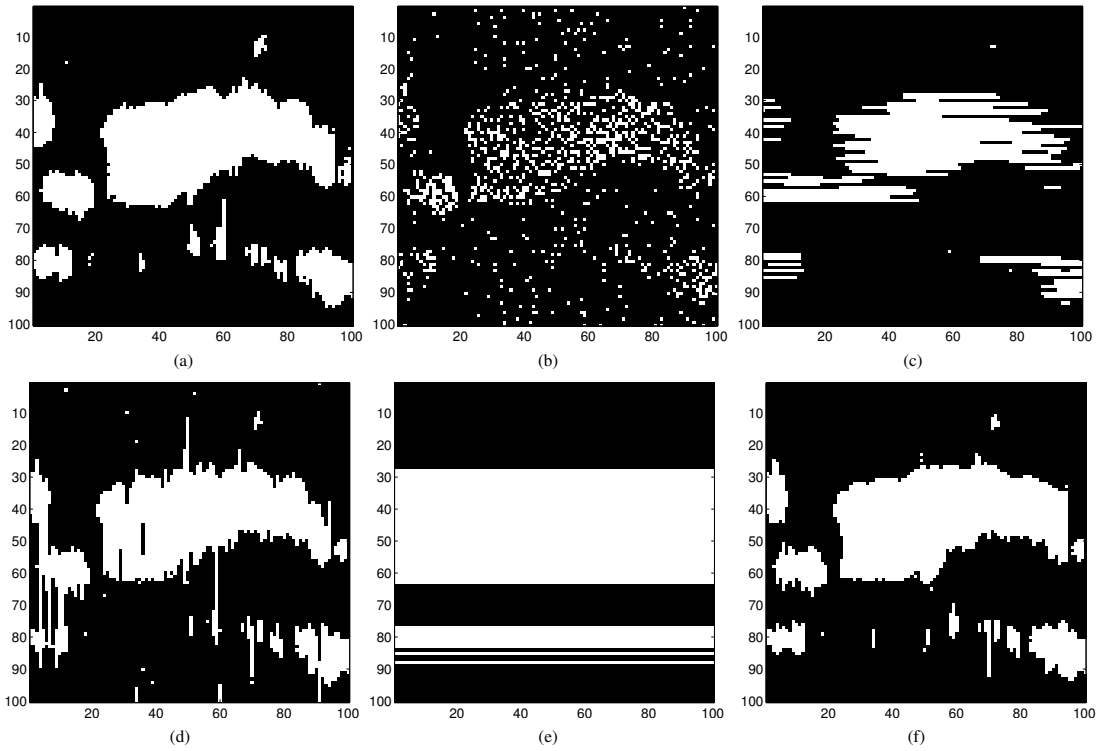


Fig. 4. True and reconstructed support for the 5 considered methods. The undersampling ratio is $N/D = 0.4$ and $D = 100, T = 100$ and $SNR = 10$dB. a) True support, b) BG-AMP (NMSE = 0.805, F = 0.450), c) DCS-AMP (NMSE = 0.777, F = 0.763), d) Spatial EP (NMSE = 0.658, F = 0.902), e) Spatial MMV EP (NMSE = 0.833, F = 0.663), f) Spatio-temporal EP (NMSE = 0.618, F = 0.935).

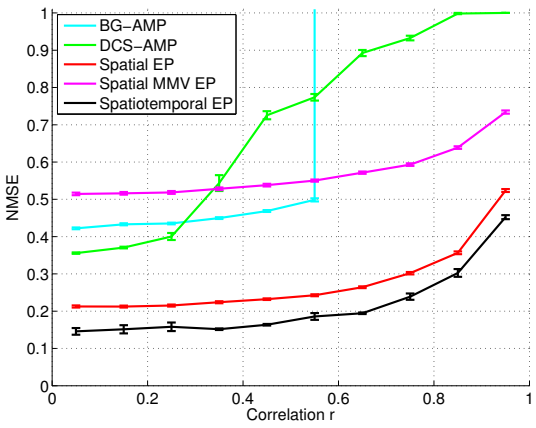


Fig. 5. NMSE error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a). The correlation of the i'th and j'th column of $\mathbf{A}$ is given by $r^{|i-j|}$. The results are averaged over 50 realizations.

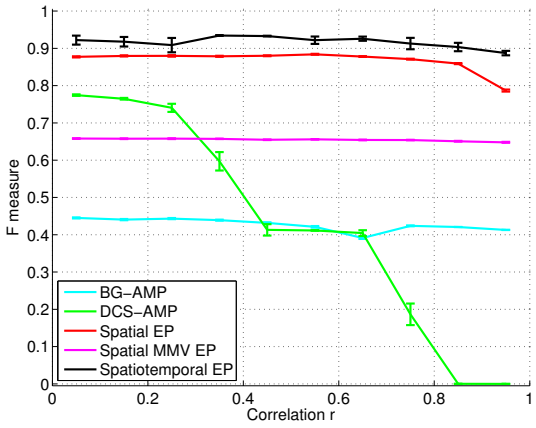


Fig. 6. F-measure error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a). The correlation of the i'th and j'th column of $\mathbf{A}$ is given by $r^{|i-j|}$. The results are averaged over 50 realizations.

VI. Conclusion & outlook

We extended the structured spike and slab prior and the associated Expectation Propagation inference scheme to cope with smooth temporal evolution of the sparsity pattern. Based on numerical experiments with synthetic data we demonstrated the benefits of the extended model. In particular, we showed that the method outperformed the reference methods. Future work includes developing an automated approach learning the hyperparameters of the prior and applying the proposed method to a real EEG source localization problem.

Acknowledgment

The authors would like to thank Sundeep Rangan et al. and Justin Ziniel for making their toolboxes available online.

References

[1] S. F. Cotter, B. D. Rao, K. Engan, K. Kreutz-delgado, and S. Member, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, pp. 2477–2488, 2005.

[2] S. Baillet, J. C. Mosher, and R. M. Leahy, "Electromagnetic brain mapping," *IEEE Signal Processing Magazine*, vol. 18, no. 6, pp. 14–30, 2001.

[3] J. M. Antelis and J. Minguez, "EEG source localization based on dynamic bayesian estimation techniques," 2012.

[4] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using markov random fields," *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, pp. 257–264, 2009.

[5] E. van den Berg and M. P. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Transactions On Information Theory*, vol. 56, no. 5, pp. 2516–2527, 2010.

[6] M. R. Andersen, O. Winther, and L. K. Hansen, "Bayesian inference for structured spike and slab priors," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1745–1753.

[7] L. Jacob, J.-P. Vert, G. Obozinski, and G. Obozinski, "Group lasso with overlap and graph lasso," *Proceedings of the 26th International Conference on Machine Learning, ICML 2009*, pp. 433–440, 2009.

[8] T. J. Mitchell and J. Beauchamp, "Bayesian variable selection in linear-regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[9] D. Hernandez-Lobato, J. Hernandez-Lobato, and P. Dupont, "Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation," *Journal Of Machine Learning Research*, vol. 14, pp. 1891–1945, 2013.

[10] L. Yu, H. Sun, J. P. Barbot, and G. Zheng, "Bayesian compressive sensing for clustered sparse signals," *Icassp, Ieee International Conference on Acoustics, Speech and Signal Processing - Proceedings, Icassp Ieee Int Conf Acoust Speech Signal Process Proc*, pp. 3948–3951, 2011.

[11] D. Donoho, "Compressed sensing," *IEEE TRANSACTIONS ON INFORMATION THEORY*, vol. 52, no. 4, pp. 1289–1306, 2006.

[12] J. Ziniel, L. C. Potter, and P. Schniter, "Tracking and smoothing of time-varying sparse signals via approximate belief propagation," *Conference Record - Asilomar Conference on Signals, Systems and Computers, Conf. Rec. Asilomar Conf. Signals Syst. Comput*, pp. 808–812, 2010.

[13] C. M. Bishop, "Pattern recognition and machine learning," 2006.

[14] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, ser. UAI '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[15] M. Opper and O. Winther, "Gaussian processes for classification: Mean-field algorithms," *Neural Computation*, vol. 12, no. 11, pp. 2655–2684, 2000.

[16] M. Seeger, "Expectation propagation for exponential families," 2009.

[17] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[18] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2012.

[19] J. P. Vila and P. Schniter, "Expectation-maximization gaussian-mixture approximate message passing," *IEEE Transactions On Signal Processing*, vol. 61, no. 19, pp. 4658–4672, 2013.

[20] J. Ziniel and P. Schniter, "Dynamic compressive sensing of time-varying signals via approximate message passing," *IEEE transactions on signal processing*, vol. 61, no. 21, pp. 5270–5284, 2013.

[21] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," *Ieee International Symposium on Information Theory - Proceedings, Ieee Int Symp Inf Theor Proc*, pp. 2168–2172, 2011.

[22] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

Appendix C

Bayesian Inference for Spatio-temporal Spike-and-slab Priors

C Andersen, M. R., Winther, O. and Hansen, L. K. (2015), 'Bayesian inference for spatio-temporal spike and slab priors'. *Re-submitted to the Journal of Machine Learning Research (JMLR) (8/3-2017), 57 pages*

Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Bayesian inference for spatio-temporal spike-and-slab priors

Michael Riis Andersen MIRI@DTU.DK
Department of Applied Mathematics and Computer Science
Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark

Aki Vehtari AKI.VEHTARI@AALTO.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science, Aalto University
P.O. Box 15400, FI-00076, Finland

Ole Winther OLWI@DTU.DK
Department of Applied Mathematics and Computer Science
Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark

Lars Kai Hansen LKAI@DTU.DK
Department of Applied Mathematics and Computer Science
Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark

Editor: Lawrence Carin

Abstract

In this work, we address the problem of solving a series of underdetermined linear inverse problems subject to a sparsity constraint. We generalize the spike-and-slab prior distribution to encode a priori correlation of the support of the solution in both space and time by imposing a transformed Gaussian process on the spike-and-slab probabilities. An expectation propagation (EP) algorithm for posterior inference under the proposed model is derived. For large scale problems, the standard EP algorithm can be prohibitively slow. We therefore introduce three different approximation schemes to reduce the computational complexity. Finally, we demonstrate the proposed model using numerical experiments based on both synthetic and real data sets.

Keywords: Linear inverse problems, bayesian inference, expectation propagation, sparsity-promoting priors, spike-and-slab priors

1. Introduction

Many problems of practical interest in machine learning involve a high dimensional feature space and a relatively small number of observations. Inference is in general difficult for such underdetermined problems due to high variance and therefore regularization is often the key to extracting meaningful information from such problems (Tibshirani, 1994). The classical approach is Tikhonov regularization (also known as ℓ_2 regularization), but during the last

©2000 Michael Riis Andersen, Aki Vehtari, Ole Winther and Lars Kai Hansen.

few decades sparsity has been an increasingly popular choice of regularization for many problems, giving rise to methods such as the LASSO (Tibshirani, 1994), Sparse Bayesian Learning (Tipping, 2001) and sparsity promoting priors (Mitchell and Beauchamp, 1988).

In this work, we address the problem of finding sparse solutions to linear inverse problems of the form

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^D$ is the desired solution, $\boldsymbol{y} \in \mathbb{R}^N$ is an observed measurement vector, $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ is a known forward model and $\boldsymbol{e} \in \mathbb{R}^N$ is additive measurement noise. We are mainly interested in the underdetermined regime, where the number of observations is smaller than the number of unknowns, that is $N < D$. In the sparse recovery literature, it has been shown that the sparsity constraint is crucial for recovering $\boldsymbol{x}$ from a small set of linear measurements (Candès et al., 2006). Furthermore, the ratio between the number non-zero coefficients $K = \|\boldsymbol{x}\|_0$ and the dimension D dictates the required number of measurements N for robust reconstruction of $\boldsymbol{x}$ and this relationship has given rise to so-called *phase transition curves* (Donoho and Tanner, 2010). A large body of research has been dedicated to improve these phase transition curves and these endeavors have lead to the concepts of *multiple measurement vectors* (Cotter et al., 2005) and *structured sparsity* (Huang et al., 2009).

The multiple measurement vector problem (MMV) is a natural extension of eq. (1), where multiple measurements $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T$ are observed and assumed to be generated from a series of signals $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$, which share a common sparsity pattern. In matrix notation, we can write the problem as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}, \tag{2}$$

where the desired solution is now a matrix $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_T \end{bmatrix} \in \mathbb{R}^{D \times T}$ and similarly for the measurement matrix $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$ and the noise term $\boldsymbol{E} \in \mathbb{R}^{N \times T}$. The assumption of *joint sparsity* allows one to recover $\boldsymbol{X}$ with significantly fewer observations compared to solving each of the T inverse problems in eq. (1) separately (Cotter et al., 2005). The MMV approach has also been generalized to problems, where the sparsity pattern is evolving slowly in time (Ziniel and Schniter, 2013a). Structured sparsity, on the other hand, is a generalization of simple sparsity and seeks to exploit the fact that the sparsity patterns of many natural signals contain a richer structure than simple sparsity, for example, *group sparsity* (Jacob et al., 2009b) or *cluster structured sparsity* (Yu et al., 2012).

In this paper, we combine these two approaches and focus on problems, where the sparsity pattern of $\boldsymbol{X}$ exhibits a spatio-temporal structure. In particular, we assume that the row and column indices of $\boldsymbol{X}$ can be associated with a set of spatial and temporal coordinates, respectively. This can equivalently be interpreted as a sparse linear regression problem, where the support of the regressors is correlated in both space and time. Applications of such a model include dynamic compressed sensing (Ziniel and Schniter, 2013a), background subtraction in computer vision (Cevher et al., 2009) and EEG source localization problem

(Baillet et al., 2001).

We take a Bayesian approach to modeling this structure since it provides a natural way of incorporating such prior knowledge in a model. In particular, we propose a hierarchical probabilistic model for $\boldsymbol{X}$ based on the so-called spike-and-slab prior (Mitchell and Beauchamp, 1988). We introduce a smooth latent variable controlling the spatio-temporal structure of the support of $\boldsymbol{X}$ by extending the work by Andersen et al. (2014). We aim for full Bayesian inference under the proposed probabilistic model, but inference w.r.t. the exact posterior distribution of interest is intractable. Instead we resort to approximate inference using Expectation Propagation (Minka, 2001; Opper and Winther, 2000), which has been shown to provide accurate inference for spike-and-slab priors (Hernández-Lobato et al., 2013; Hernandez-Lobato et al., 2010; Jylänki et al., 2014; Peltola et al., 2014). Our model formulation is generic and generalizes easily to other types of observations. In particular, we also combine the proposed prior with a probit observation model to model binary observations in a sparse linear classification setting.

The contribution of this paper is three-fold. First we extend the structured spike-and-slab prior and the associated EP inference scheme to incorporate both spatial and temporal smoothness of the support. However, the computational complexity of the resulting EP algorithm is prohibitively slow for problems of even moderate sizes of signal dimension D and length T. To alleviate the computational bottleneck of the EP algorithm we propose three different approximation schemes. Finally, we discuss several approaches for learning the hyperparameters and evaluate them based on synthetic and real data sets.

1.1 Related work

In this section, we briefly review some of the most common approaches to simple sparsity and their generalization to structured sparsity. The classical approach to sparsity is the LASSO (Tibshirani, 1994), which operates by optimizing a least squares cost function augmented with an ℓ_1 penalty on the regression weights. Several extensions have been proposed in the literature to generalize the LASSO to the structured sparsity setting, examples include group and graph LASSO (Jacob et al., 2009b). From a probabilistic perspective sparsity can be encouraged through the use of *sparsity-promoting priors*. A non-exhaustive list of sparsity-promoting priors includes the Laplace prior (Park and Casella, 2008), Automatic Relevance Determination prior (Neal, 1996), the horseshoe prior (Carvalho et al., 2009) and the spike-and-slab prior (Mitchell and Beauchamp, 1988). All of these were originally designed to enforce simply sparsity, but they have all been generalized to the structured sparsity setting. The general strategy is to extend univariate densities to correlated multivariate densities by augmenting the models with a latent multivariate variable, where the correlation structure can be controlled explicitly, for example, using Markov Random Fields (Cevher et al., 2009; Hernandez-Lobato et al., 2011) or multivariate Gaussian distributions (Engelhardt and Adams, 2014). Here we limit ourselves to consider the latter.

From a probabilistic perspective, optimizing with an ℓ_1 regularization term can be interpreted as maximum a posteriori (MAP) inference under an i.i.d. Laplace prior distribution on

the regression weights (Park and Casella, 2008). The univariate Laplace prior has been generalized to the multivariate Laplace (MVL) distribution, which couples the prior variance of the regression weights through a scale mixture formulation (Gerven et al., 2009).

Another approach is Automatic Relevance Determination (ARD) (Neal, 1996), which works by imposing independent zero mean Gaussian priors with individual precision parameters on the regression weights. These precision parameters are then optimized using a maximum likelihood type II and the idea is then that the precision parameters of irrelevant features will approach infinity and thereby forcing the weights of the irrelevant features to zero. Wu et al. (2014b) extend the ARD framework to promote spatial sparsity by introducing a latent multivariate Gaussian distribution to impose spatial structure onto the precision parameters of ARD giving rise to *dependent relevance determination priors*.

The horseshoe prior is defined as a scale mixture of Gaussians, where a half-Cauchy distribution is used as prior for the standard deviation of the Gaussian density (Carvalho et al., 2009). The resulting density has two very appealing properties for promoting sparsity, namely heavy tails and an infinitely large spike at zero. A generalization to the multivariate case has been proposed by Hernández-Lobato and Hernández-Lobato (2013).

The spike-and-slab prior is an increasingly popular choice of sparsity promoting prior and is given by a binary mixture of two components: a Dirac delta distribution (spike) at zero and Gaussian distribution (slab) (Mitchell and Beauchamp, 1988; Carbonetto and Stephens, 2012). The spike-and-slab prior has been generalized to the group setting by Hernández-Lobato et al. (2013), to clustered sparsity setting by Yu et al. (2012) and spatial structures by Andersen et al. (2014), Nathoo et al. (2014), and Engelhardt and Adams (2014). Nathoo et al. (2014) induce the spatial structure using basis functions and Andersen et al. (2014) impose the structure using a multivariate Gaussian density. The latter is the starting point of this work.

Our work is closely related to the work on the multivariate Laplace prior (MVL) (Gerven et al., 2009) as mentioned above and the work on the network-based sparse Bayesian classification algorithm (NBSBC) (Hernandez-Lobato et al., 2011). The former also uses EP for approximating the posterior distribution of a Gaussian linear model with the MVL prior, where the structure of the support is encoded into the model using a sparse precision matrix. The NBSBC method also uses EP to approximate the posterior distribution of linear model with coupled spike-and-slab priors, but the structure of the support is encoded in a network using a Markov Random Field (MRF) prior. In contrast, we can inject a priori knowledge of the structure into the model using generic covariance functions rather than clique potentials as in the MRF-based models, which makes it easier to interpret interesting quantities like the characteristic lengthscale etc.

1.2 Structure of paper

This paper is organized as follows. In section 2 we review the structured spike-and-slab prior and in section 3 we discuss different ways of extending the model to include the temporal

structure as well. After introducing the models we propose an algorithm for approximate inference based on the expectation propagation (EP) framework. We review the basics of EP and describe the proposed algorithm in section 4. In section 5 we introduce three simple approximation schemes to speed of the inference process and discuss their properties. Finally, in section 7 we demonstrate the proposed method using synthetic and real data sets.

1.3 Notation

We use bold uppercase letters to denote matrices and bold lowercase letters to denote vectors. Unless stated otherwise, all vectors are column vectors. Furthermore, we use the notation $\boldsymbol{a}_{n,\cdot} \in \mathbb{R}^{1\times D}$ and $\boldsymbol{a}_{\cdot,i} \in \mathbb{R}^{N\times 1}$ for the n'th row and i'th column in the matrix $\boldsymbol{A} \in \mathbb{R}^{N\times D}$, respectively. $[K]$ denotes the set of integers from 1 to K, that is $[K] = \{1, 2, .., K\}$. We use the notation $\boldsymbol{a} \circ \boldsymbol{b}$ to denote the element-wise Hadamard product of $\boldsymbol{a}$ and $\boldsymbol{b}$ and $\boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{R}^{MN\times MN}$ for the Kronecker product of matrices $\boldsymbol{A} \in \mathbf{R}^{M\times M}$ and $\boldsymbol{B} \in \mathbf{R}^{N\times N}$. We use $\mathcal{N}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{V})$ to denote a multivariate Gaussian density over $\boldsymbol{x}$ with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{V}$ and $\mathrm{Ber}(z|p)$ denotes a Bernoulli distribution on z with probability of $p(z = 1) = p$.

2. The structured spike-and-slab prior

The purpose of this section is to describe the *structured spike-and-slab prior* (Andersen et al., 2014), but first we briefly review the conventional spike-and-slab prior (Mitchell and Beauchamp, 1988). For $\boldsymbol{x} \in \mathbb{R}^D$, the spike-and-slab prior distribution is given by

$$p(\boldsymbol{x}|p_0, \rho_0, \tau_0) = \prod_{i=1}^{D} \left[(1 - p_0)\delta(x_i) + p_0\mathcal{N}(x_i|\rho_0, \tau_0)\right], \tag{3}$$

where $\delta(x)$ is the Dirac delta function and p_0, ρ_0 and τ_0 are hyperparameters. In particular, p_0 is the prior probability of a given variable being active, that is $p(x_i \neq 0) = p_0$, and ρ_0, τ_0 are the prior mean and variance, respectively, of the active variables. The spike-and-slab prior in eq. (3) is also known as the Bernoulli-Gaussian prior since the prior can decomposed as

$$p(\boldsymbol{x}\big|p_0, \rho_0, \tau_0) = \sum_{\boldsymbol{z}} \prod_{i=1}^{D} \left[(1 - z_i)\delta(x_i) + z_i\mathcal{N}(x_i|\rho_0, \tau_0)\right] \prod_{i=1}^{D} \mathrm{Ber}(z_i|p_0), \tag{4}$$

where the sum is over all the binary variables z_i for $i \in [D]$. Thus, the latent binary variable $z_i \in \{0, 1\}$ can interpreted as an indicator variable for the event $x_i \neq 0$. We will refer to $\boldsymbol{z}$ as the *sparsity pattern* or the *support* of $\boldsymbol{x}$. In eq. (3) and (4) we condition explicitly on the hyperparameters p_0, ρ_0, τ_0, but to ease the notation we will omit this in the remainder of this paper.

The variables x_i and x_j are assumed to be independent for $i \neq j$ as seen in eq. (3) and (4). This implies that the number of active variables follows a binomial distribution and hence, the marginal probability of x_i and x_j being jointly active, is given by $p(x_i \neq 0, x_j \neq 0) = p_0^2$ for all $i \neq j$. However, in many applications the variables $\{x_k\}_{k=1}^D$ might a priori have an

underlying topographic relationship such as a spatial or temporal structure. Without loss of generality we will assume a spatial relationship, where $\boldsymbol{d}_i$ denotes the spatial coordinates of x_i. For such models, it is often a reasonable assumption that $p(x_i \neq 0, x_j \neq 0)$ should depend on $\|\boldsymbol{d}_i - \boldsymbol{d}_j\|$. For instance, neighboring voxels in functional magnetic resonance imaging (fMRI) analysis (Penny et al., 2005) are often more likely to be active simultaneously compared to two voxels far apart. Such a priori knowledge is neglected by the conventional spike-and-slab prior in eq. (3).

The structured spike-and-slab model is capable of modeling such structure and is given in terms of a hierarchical model

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{D} \left[(1 - z_i)\, \delta(x_i) + z_i \mathcal{N}\left(x_i | \rho_0, \tau_0\right) \right], \tag{5}$$

$$p(\boldsymbol{z}|\boldsymbol{\gamma}) = \prod_{i=1}^{D} \mathrm{Ber}\left(z_i | \phi(\gamma_i)\right), \qquad \phi : \mathbb{R} \to (0,1), \tag{6}$$

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right), \tag{7}$$

where $\boldsymbol{\gamma}$ is a latent variable controlling the structure of the sparsity pattern. Using this model prior knowledge of the structure of the sparsity pattern can be encoded using $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. The mean value $\boldsymbol{\mu}_0$ controls the expected degree of sparsity and the covariance matrix $\boldsymbol{\Sigma}_0$ determines the prior correlation of the support. The map $\phi : \mathbb{R} \to (0,1)$ serves the purpose of squeezing γ_i into the unit interval and thereby $\phi(\gamma_i)$ represents the probability of $z_i = 1$. Here we choose ϕ to be the standard normal cumulative distribution function (CDF), but other choices, such as the logistic function, are also possible.

Using this formulation, the marginal prior probability of the i'th variable being active is given by

$$p(z_i = 1) = \int p(z_i = 1|\gamma_i) p(\gamma_i) \mathrm{d}\gamma_i = \int \phi(\gamma_i) \mathcal{N}\left(\gamma_i | \mu_i, \Sigma_{0,ii}\right) \mathrm{d}\gamma_i = \phi\left(\frac{\mu_i}{\sqrt{1 + \Sigma_{0,ii}}}\right). \tag{8}$$

From this expression it is seen that when $\mu_i = 0$, the prior belief of z_i is unbiased and $p(z_i = 1) = 0.5$, but when $\mu_i < 0$ the variable z_i is biased toward zero and vice versa. If a subset of features $\{x_j | j \in \mathcal{J} \subset [D]\}$ is a priori more likely to explain the observed data $\boldsymbol{y}$, then this information can be encoded in the prior distribution by assigning the prior mean of $\boldsymbol{\gamma}$ such that $\mu_j > \mu_i$ for all $j \in \mathcal{J}$ and for all $i \notin \mathcal{J}$. However, in the remainder of this paper we will assume that the prior mean is constant, that is $\mu_i = \nu_0$ for some $\nu_0 \in \mathbb{R}$. For more details on the prior distribution, see Appendix D.

The prior probability of two variables, x_i and x_j, being jointly active is

$$p(z_i = 1, z_j = 1) = \int \phi(\gamma_i) \phi(\gamma_j)\, \mathcal{N}\left(\boldsymbol{\gamma} | \boldsymbol{\mu}, \boldsymbol{\Sigma}_0\right) \mathrm{d}\boldsymbol{\gamma}. \tag{9}$$

If $\boldsymbol{\Sigma}_0$ is a diagonal matrix, γ_i and γ_j become independent and we recover the conventional spike-and-slab prior. On the other hand, if we choose $\boldsymbol{\Sigma}_0$ to be a covariance matrix of

the form $\Sigma_{0,ij} = g\left(\|\boldsymbol{d}_i - \boldsymbol{d}_j\|\right)$, we see that the joint activation probabilities indeed depend on the spatial distance as desired. Finally, we emphasize that this parametrization it not limited to nearest neighbors-type structures. In fact, this parametrization supports general structures that can be modeled using generic covariance functions.

3. The spatio-temporal spike-and-slab prior

In the following we will extend the structured spike-and-slab prior distribution to model temporal smoothness of the sparsity pattern as well. Let $t \in [T]$ be the time index, then $\boldsymbol{x}_t$, $\boldsymbol{z}_t$ and $\boldsymbol{\gamma}_t$ are the signal coefficients, the sparsity patterns and the latent structure variable at time t. Furthermore, we define the corresponding matrix quantities $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_T\end{bmatrix}$, $\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{z}_1 & \boldsymbol{z}_2 & \boldsymbol{z}_T\end{bmatrix}$ and $\boldsymbol{\Gamma} = \begin{bmatrix}\gamma_1 & \gamma_2 & \dots & \gamma_T\end{bmatrix}$.

There are several natural temporal extensions of the model. The simplest extension is to assume that $\{\boldsymbol{\gamma}_t\}_{t=1}^T$ is independent in time, so that $p(\boldsymbol{Z}, \boldsymbol{\Gamma}) = \prod_{t=1}^T p(\boldsymbol{z}_t|\boldsymbol{\gamma}_t) \prod_{t=1}^T p(\boldsymbol{\gamma}_t)$, which is equivalent to solving each of the T regressions problems in eq. (1) independently. Another simple extension is to use the so-called *joint sparsity* assumption (Cotter et al., 2005; Zhang and Rao, 2011; Ziniel and Schniter, 2013b) and assume that the sparsity pattern is static across time, and thus all $\{\boldsymbol{x}_t\}_{t=1}^T$ vectors share a common binary support vector $\boldsymbol{z}$, and $p(\boldsymbol{X}|\boldsymbol{z}) = \prod_{t=1}^T \prod_{i=1}^D \left[(1 - z_i)\,\delta\left(x_{i,t}\right) + z_i \mathcal{N}\left(x_{i,t}|\rho_0, \tau_0\right)\right]$. A more interesting and flexible model is to assume that the support is slowly changing in time, by modelling the temporal evolution of $\boldsymbol{\gamma}_t$ using a first order Gauss-Markov process of the form $p\left(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\gamma}_t|\,(1-\alpha)\,\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right)$, where the hyperparameters $\alpha \in [0, 1]$ and $\beta > 0$ control the temporal correlation and the "innovation" of the process, respectively.

The first order model has the advantage that it factorizes across time, which makes the resulting inference problem much easier. On the other hand, first order Markovian dynamics is often not sufficient for capturing long range correlations. Imposing a Gaussian process distribution on $\boldsymbol{\Gamma}$ with arbitrary covariance structure would facilitate modeling of long range correlations in both time and space. Therefore, the hierarchical prior distribution for $\boldsymbol{X}$ becomes

$$p(\boldsymbol{X}|\boldsymbol{Z}) = \prod_{t=1}^{T} \prod_{i=1}^{D} \left[(1 - z_{i,t})\,\delta\left(x_{i,t}\right) + z_{i,t}\mathcal{N}\left(x_{i,t}|\rho_0, \tau_0\right)\right] \tag{10}$$

$$p(\boldsymbol{Z}|\boldsymbol{\Gamma}) = \prod_{t=1}^{T} \mathrm{Ber}\left(\boldsymbol{z}_t|\phi\left(\boldsymbol{\gamma}_t\right)\right) \tag{11}$$

$$p(\boldsymbol{\Gamma}) = \mathcal{N}\left(\boldsymbol{\Gamma}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right), \tag{12}$$

where the mean $\boldsymbol{\mu}_0 \in \mathbb{R}^{TD\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{TD\times TD}$ are now defined for the full $\boldsymbol{\Gamma}$-space. This model is more expressive, but the resulting inference problem becomes infeasible for even moderate sizes of D and T. But if we assume that the underlying spatio-temporal grid can be written in Cartesian product form, then covariance matrix simplifies to a Kronecker product

$$p(\boldsymbol{\Gamma}) = \mathcal{N}\left(\boldsymbol{\Gamma}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{\text{temporal}} \otimes \boldsymbol{\Sigma}_{\text{spatial}}\right), \tag{13}$$

where $\mathbf{\Sigma}_{\text{temporal}} \in \mathbb{R}^{T\times T}$ and $\mathbf{\Sigma}_{\text{spatial}} \in \mathbb{R}^{D\times D}$. This decomposition leads to more efficient inference schemes as we will discuss in section 5. In the remainder of the paper, we will focus on the model with Kronecker structure, but we refer to (Andersen et al., 2015) for more details on the first order model and joint sparsity model.

The coefficients $\{x_{i,t}\}$ are conditionally independent given the support $\{z_{i,t}\}$. For some applications it could be desirable to impose either spatial smoothness, temporal smoothness or both on the non-zero coefficients themselves (Wu et al., 2014a; Ziniel and Schniter, 2013a), but in this work we only assume a priori knowledge of the structure of the support. Although temporal smoothness of $x_{i,t}$ could easily be incorporated into the models described above.

4. Inference using spatiotemporal priors

In the previous sections we have described the structured spike-and-slab prior and how to extend it to model temporal smoothness as well. We now turn our attention on how to perform inference using these models. We focus our discussion on the most general formulation using as given in eq. (10)-(12). Let $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{y}_1 & \boldsymbol{y}_2 & \dots & \boldsymbol{y}_T\end{bmatrix}$ be an observation matrix, where $\boldsymbol{y}_t \in \mathbb{R}^N$ is an observation vector for time t. We assume that the distribution on $\boldsymbol{Y}$ factors over time and is given by

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{x}_t). \tag{14}$$

We consider two different noise models: an isotropic Gaussian noise model and a probit noise model. The Gaussian noise model $p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{y}_t|\boldsymbol{A}\boldsymbol{x}_t, \sigma_2\boldsymbol{I}\right)$ is suitable for linear inverse problems with forward model $\boldsymbol{A} \in \mathbb{R}^{N\times D}$ or equivalently sparse linear regression problems with design matrix $\boldsymbol{A} \in \mathbb{R}^{N\times D}$. On the other hand, the probit model is suitable for modeling binary observations, with $y_{t,n} \in \{-1, 1\}$, and is given by $p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \prod_{n=1}^{N} \phi\left(y_{t,n}\boldsymbol{a}_{n,\cdot}\boldsymbol{x}_t\right)$, where $\boldsymbol{a}_{n,\cdot}$ is the n'th row of $\boldsymbol{A}$. For both models we further assume that the matrix $\boldsymbol{A}$ is constant across time. However, this assumption can be easily relaxed to have $\boldsymbol{A}$ depend on t.

For both noise models the resulting joint distribution becomes

$$\begin{aligned} p\left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \mathbf{\Gamma}\right) &= p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z}|\mathbf{\Gamma})p(\mathbf{\Gamma}) && (15)\\ &= \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{x}_t) \prod_{t=1}^{T} \left[(1-\boldsymbol{z}_t)\circ\delta\left(\boldsymbol{x}_t\right) + \boldsymbol{z}_t \circ \mathcal{N}\left(\boldsymbol{x}_t|0, \tau\boldsymbol{I}\right)\right] \\ &\quad \prod_{t=1}^{T} \mathrm{Ber}\left(\boldsymbol{z}_t|\phi\left(\boldsymbol{\gamma}_t\right)\right) \mathcal{N}\left(\mathbf{\Gamma}|\boldsymbol{\mu}_0, \mathbf{\Sigma}_0\right). && (16)\end{aligned}$$

We seek the posterior distribution of the parameters $\boldsymbol{X}, \boldsymbol{Z}$ and $\mathbf{\Gamma}$ conditioned on the observations $\boldsymbol{Y}$, which is obtained by applying Bayes's Theorem to the joint distribution in

eq. (15)

$$p\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma} \middle| \boldsymbol{Y}\right) = \frac{1}{Z} \prod_{t=1}^{T} p(\boldsymbol{y}_t | \boldsymbol{x}_t) \prod_{t=1}^{T} \left[(1 - \boldsymbol{z}_t) \circ \delta\left(\boldsymbol{x}_t\right) + \boldsymbol{z}_t \circ \mathcal{N}\left(\boldsymbol{x}_t \middle| 0, \tau \boldsymbol{I}\right)\right] \prod_{t=1}^{T} \mathrm{Ber}\left(\boldsymbol{z}_t \middle| \phi\left(\boldsymbol{\gamma}_t\right)\right) \mathcal{N}\left(\boldsymbol{\Gamma} \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right), \tag{17}$$

where $Z = p(\boldsymbol{Y})$ is the marginal likelihood of $\boldsymbol{Y}$. Due to the product of mixtures in the distribution $p(\boldsymbol{X}|\boldsymbol{Z})$, the expression for the marginal likelihood Z involves a sum over 2^{DT} terms. This renders the computation of the normalization constant Z intractable for even small D and T. Hence, the desired posterior distribution is also intractable and we have to resort to approximate inference.

In the literature researchers have applied a whole spectrum of approximate inference methods for spike-and-slab priors, for example, Monte Carlo-methods (Mitchell and Beauchamp, 1988), mean-field variational inference (Titsias and Lazaro-Gredilla, 2011), approximate message passing (Vila and Schniter, 2013) and expectation propagation (Hernández-Lobato et al., 2013; Andersen et al., 2014). We use the latter since expectation propagation has been shown to have good performance for linear models with spike-and-slab priors (Hernández-Lobato et al., 2015) and it has been shown to provide a much better approximation of the first and second moment posterior moment for spike-and-slab models (Peltola et al., 2014).

4.1 The Expectation Propagation Framework

In this section, we briefly review expectation propagation for completeness. Expectation propagation (EP) (Minka, 2001; Opper and Winther, 2000) is a deterministic framework for approximating probability distributions. Consider a probability distribution over the variable $\boldsymbol{x} \in \mathbb{R}^D$ that factorizes into N components

$$f(\boldsymbol{x}) = \prod_{i=1}^{N} f_i(\boldsymbol{x}_i), \tag{18}$$

where $\boldsymbol{x}_i$ is taken to be a subvector of $\boldsymbol{x}$. EP takes advantage of this factorization and approximates f with a distribution Q that shares the same factorization

$$Q(\boldsymbol{x}) = \prod_{i=1}^{N} \tilde{f}_i(\boldsymbol{x}_i). \tag{19}$$

EP approximates each *site term* f_i with a (scaled) distribution $\tilde{f}_i$ from the exponential family. Since the exponential family is closed under products, the *global approximation* Q will also be in the exponential family. Consider the product of all $\tilde{f}_i$ terms except the j'th term

$$Q^{\setminus j}(\boldsymbol{x}) = \prod_{i \neq j} \tilde{f}_i(\boldsymbol{x}_i) = \frac{Q(\boldsymbol{x})}{\tilde{f}_j(\boldsymbol{x}_j)}. \tag{20}$$

The core of the EP framework is to choose $\tilde{f}_j$ such that $\tilde{f}_j(\boldsymbol{x}_j)Q^{\setminus j}(\boldsymbol{x}_j) \approx f_j(\boldsymbol{x}_j)\,Q^{\setminus j}(\boldsymbol{x}_j)$. By approximating f_j with $\tilde{f}_j$ in the context of $Q^{\setminus j}$, we ensure that the approximation is most accurate in the region of high density according to the *cavity distribution* $Q^{\setminus j}$. This scheme is implemented by iteratively minimizing the Kullbach-Leibler divergence $\mathrm{KL}\Big(f_j(\boldsymbol{x}_j)\,Q^{\setminus j}(\boldsymbol{x})\big|\big|\tilde{f}_j(\boldsymbol{x}_j)\,Q^{\setminus j}(\boldsymbol{x})\Big)$. Since $\tilde{f}_j(\boldsymbol{x}_j)\,Q^{\setminus j}(\boldsymbol{x})$ belongs to the exponential family, the unique solution is obtained by matching the expected sufficient statistics (Bishop, 2006). Once the solution,

$$Q^* = \underset{q}{\operatorname{argmin}}\ \mathrm{KL}\Big(f_j(\boldsymbol{x}_j)\,Q^{\setminus j}(\boldsymbol{x})\big|\big|q\Big), \tag{21}$$

is obtained, the j'th site approximation is updated as

$$\tilde{f}_j^*(\boldsymbol{x}_j) \propto \frac{Q^*(\boldsymbol{x})}{Q^{\setminus j}(\boldsymbol{x})}. \tag{22}$$

The steps in eq. (20), (21) and (22) are repeated sequentially for all $j \in [D]$ until convergence is achieved.

4.2 The Expectation Propagation Approximation

The EP framework provides flexibility in the choice of the approximating factors. This choice is a trade-off between analytical tractability and sufficient flexibility for capturing the important characteristics of the true density. Consider the desired posterior density of interest

$$\begin{aligned} p\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Gamma}\middle|\boldsymbol{Y}\right) \propto & \underbrace{\prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{x}_t)}_{f_1(\boldsymbol{X})} \underbrace{\prod_{t=1}^{T}\left[(1-\boldsymbol{z}_t)\circ\delta\left(\boldsymbol{x}_t\right)+\boldsymbol{z}_t\circ\mathcal{N}\left(\boldsymbol{x}_t\middle|0,\tau\boldsymbol{I}\right)\right]}_{f_2(\boldsymbol{X},\boldsymbol{Z})} \\ & \underbrace{\prod_{t=1}^{T}\mathrm{Ber}\left(\boldsymbol{z}_t\middle|\phi\left(\boldsymbol{\gamma}_t\right)\right)}_{f_3(\boldsymbol{Z},\boldsymbol{\Gamma})} \underbrace{\mathcal{N}\left(\boldsymbol{\Gamma}\middle|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0\right)}_{f_4(\boldsymbol{\Gamma})}. \end{aligned} \tag{23}$$

This posterior density is decomposed into four terms f_i for $i = 1,..,4$, where the first three terms can be further decomposed. The term $f_1(\boldsymbol{X})$ is decomposed into T terms of the form $f_{1,t}(\boldsymbol{x}_t) = p(\boldsymbol{y}_t|\boldsymbol{x}_t)$, whereas the terms f_2 and f_3 are further decomposed as follows

$$f_1(\boldsymbol{X}) = \prod_{t=1}^{T}\tilde{f}_{1,}(\boldsymbol{x}_t) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{x}_t), \tag{24}$$

$$f_2(\boldsymbol{X},\boldsymbol{Z}) = \prod_{t=1}^{T}\prod_{i=1}^{D} f_{2,i,t}(x_{i,t},z_{i,t}) = \prod_{t=1}^{T}\prod_{i=1}^{D}\left[(1-z_{i,t})\circ\delta\left(x_{i,t}\right)+z_{i,t}\circ\mathcal{N}\left(x_{i,t}\middle|\rho,\tau\right)\right], \tag{25}$$

$$f_3(\boldsymbol{Z},\boldsymbol{\Gamma}) = \prod_{t=1}^{T}\prod_{i=1}^{D} f_{3,i,t}(z_{i,t},\gamma_{i,t}) = \prod_{t=1}^{T}\prod_{i=1}^{D}\mathrm{Ber}\left(z_{i,t}\middle|\phi\left(\gamma_{i,t}\right)\right). \tag{26}$$

Each $f_{1,t}$ term only depends on $\boldsymbol{x}_t$, $f_{2,i,t}$ only depends on $x_{i,t}$ and $z_{i,t}$ and $f_{3,j,t}$ only depends on $z_{i,t}$ and $\gamma_{i,t}$. Furthermore, the terms $f_{2,i,t}$ couple the variables $x_{i,t}$ and $z_{i,t}$, while $f_{3,i,t}$ couple the variables $z_{i,t}$ and $\gamma_{i,t}$. Based on these observations, we choose $\tilde{f_{1,t}}$, $\tilde{f_{2,i,t}}$ and $\tilde{f_{3,j,t}}$ to have the following forms

$$\tilde{f}_{1,t}\left(\boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_t \middle| \hat{\boldsymbol{m}}_{1,t}, \hat{\boldsymbol{V}}_{1,t}\right), \tag{27}$$

$$\tilde{f}_{2,i,t}\left(x_{i,t}, z_{i,t}\right) = \mathcal{N}\left(x_{i,t} \middle| \hat{m}_{2,i,t}, \hat{v}_{2,i,t}\right) \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\hat{\gamma}_{2,i,t}\right)\right) \tag{28}$$

$$\tilde{f}_{3,i,t}\left(z_{i,t}, \gamma_{i,t}\right) = \mathcal{N}\left(\gamma_{i,t} \middle| \hat{\mu}_{3,j,t}, \hat{\sigma}_{3,i,t}\right) \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\hat{\gamma}_{3,j,t}\right)\right). \tag{29}$$

The exact term f_1 is a distribution wrt. $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$, whereas the approximate term $\tilde{f}_1$ is a function of $\boldsymbol{x}$ that depends on the data $\boldsymbol{y}$ through $\hat{\boldsymbol{m}}_1$ and $\hat{\boldsymbol{V}}_1$ etc. Finally, f_4 already belongs to the exponential family and does therefore not have to be approximated by EP. That is, $\tilde{f}_4\left(\boldsymbol{\Gamma}\right) = f_4\left(\boldsymbol{\Gamma}\right) = \mathcal{N}\left(\boldsymbol{\Gamma} \middle| \mu_0, \boldsymbol{\Sigma}_0\right)$.

Define $\hat{\boldsymbol{m}}_{2,t} = \begin{bmatrix}\hat{m}_{2,t,1} & \hat{m}_{2,t,2} & \dots & \hat{m}_{2,t,D}\end{bmatrix}^T$, $\hat{\boldsymbol{V}}_{2,t} = \mathrm{diag}\begin{pmatrix}\hat{v}_{2,t,1} & \hat{v}_{2,t,2} & \dots & \hat{v}_{2,t,D}\end{pmatrix}^T$ and $\hat{\boldsymbol{\gamma}}_{2,t} = \begin{bmatrix}\hat{\gamma}_{2,t,1} & \hat{\gamma}_{2,t,2} & \dots & \hat{\gamma}_{2,t,D}\end{bmatrix}$ and similarly for $\hat{\boldsymbol{\mu}}_{3,t}$, $\hat{\boldsymbol{\Sigma}}_{3,t}$ and $\hat{\boldsymbol{\gamma}}_{3,t}$, then the resulting global approximation becomes

$$\begin{aligned}
Q\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}\right) \propto & \prod_{t=1}^{T} \underbrace{\mathcal{N}\left(\boldsymbol{x}_t \middle| \hat{\boldsymbol{m}}_{1,t}, \hat{\boldsymbol{V}}_{1,t}\right)}_{\tilde{f}_{1,t}} \prod_{t=1}^{T} \underbrace{\mathcal{N}\left(\boldsymbol{x}_t \middle| \hat{\boldsymbol{m}}_{2,t}, \hat{\boldsymbol{V}}_{2,t}\right) \mathrm{Ber}\left(\boldsymbol{z}_t \middle| \phi\left(\hat{\boldsymbol{\gamma}}_{2,t}\right)\right)}_{\tilde{f}_{2,t}} \\
& \prod_{t=1}^{T} \underbrace{\mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \hat{\boldsymbol{\mu}}_{3,t}, \hat{\boldsymbol{\Sigma}}_{3,t}\right) \mathrm{Ber}\left(\boldsymbol{z}_t \middle| \phi\left(\hat{\boldsymbol{\gamma}}_{3,t}\right)\right)}_{\tilde{f}_{3,t}} \underbrace{\mathcal{N}\left(\boldsymbol{\Gamma} \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)}_{\tilde{f}_4} \\
\propto & \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{x}_t \middle| \hat{\boldsymbol{m}}_t, \hat{\boldsymbol{V}}_t\right) \prod_{t=1}^{T} \mathrm{Ber}\left(\boldsymbol{z}_t \middle| \phi\left(\hat{\boldsymbol{\gamma}}_t\right)\right) \mathcal{N}\left(\boldsymbol{\Gamma} \middle| \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right),
\end{aligned} \tag{30}$$

where the parameters of the global approximation are obtained by summing the natural parameters. In terms of mean and variance, we get

$$\hat{\boldsymbol{V}}_t = \left[\hat{\boldsymbol{V}}_{1,t}^{-1} + \hat{\boldsymbol{V}}_{2,t}^{-1}\right]^{-1} \tag{31}$$

$$\hat{\boldsymbol{m}}_t = \hat{\boldsymbol{V}}_t \left[\hat{\boldsymbol{V}}_{1,t}^{-1} \hat{\boldsymbol{m}}_{1,t} + \hat{\boldsymbol{V}}_{2,t}^{-1} \hat{\boldsymbol{m}}_{2,t}\right] \tag{32}$$

$$\hat{\boldsymbol{\Sigma}} = \left[\boldsymbol{\Sigma}_0^{-1} + \hat{\boldsymbol{\Sigma}}_3^{-1}\right]^{-1} \tag{33}$$

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\Sigma}} \left[\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \hat{\boldsymbol{\Sigma}}_3^{-1} \hat{\boldsymbol{\mu}}_3\right] \tag{34}$$

$$\phi\left(\hat{\gamma}_{i,t}\right) = \frac{\phi\left(\hat{\gamma}_{2,i,t}\right) \phi\left(\hat{\gamma}_{3,i,t}\right)}{\left(1 - \phi\left(\hat{\gamma}_{2,i,t}\right)\right)\left(1 - \phi\left(\hat{\gamma}_{3,j,t}\right)\right) + \phi\left(\hat{\gamma}_{2,i,t}\right) \phi\left(\hat{\gamma}_{3,i,t}\right)}, \tag{35}$$

where $\hat{\boldsymbol{\Sigma}}_3 \in \mathbb{R}^{TD \times TD}$ is a diagonal matrix, whose the diagonal is obtained by stacking the site variances $\hat{\boldsymbol{\Sigma}}_{3,t}$ for each time point and $\hat{\boldsymbol{\mu}}_3 \in \mathbb{R}^{TD}$ is a vector obtained by stacking the site means $\hat{\boldsymbol{\mu}}_{3,t}$ for each time point. To compute the global approximation, we need to

estimate the parameters $\hat{\boldsymbol{m}}_{1,t}$, $\hat{\boldsymbol{V}}_{1,t}$, $\hat{\boldsymbol{m}}_{2,t}$, $\hat{\boldsymbol{V}}_{2,t}$, $\hat{\boldsymbol{\mu}}_{3,t}$, $\hat{\boldsymbol{\Sigma}}_{3,t}$, $\hat{\gamma}_{2,t}$ and $\hat{\gamma}_{3,t}$ for all $t \in [T]$ using EP. The estimation procedure of $\hat{\boldsymbol{m}}_{1,t}$ and $\hat{\boldsymbol{V}}_{1,t}$ depends on the observation model being used, whereas the estimation procedure of the remaining parameters are independent on the choice of observation model.

In principle, we could choose the approximate posterior distribution of $\boldsymbol{\Gamma}$ in eq. (30) from a family of distributions that factorizes across space, time or both to reduce the computational complexity. This choice would indeed reduce the computational burden, but in contrast to classical variational inference schemes, the correlation structure of the prior would be ignored in the EP scheme and thus, the resulting posterior approximation would be meaningless for this specific model.

In the conventional EP algorithm, the site approximations are updated in a sequential manner meaning that the global approximation is updated every time a single site approximation (Minka, 2001) is refined. In this work, we use the parallel update scheme to reduce the computational complexity of the algorithm. That is, we first update all the site approximations of the form $\tilde{f}_{2,i,t}$ for $i \in [D]$, $t \in [T]$, and then we update the global approximation w.r.t. $\boldsymbol{x}_t$ and similarly for the $\tilde{f}_{3,i,t}$ and the global approximation w.r.t. $\boldsymbol{\gamma}_t$. From a message passing perspective this can be interpreted as a particular scheduling of messages (Minka, 2005). The proposed algorithm is summarized in Algorithm 1.

4.3 Estimating parameters for $\tilde{f}_{1,t}$

The estimation procedure for $\tilde{f}_{1,t}$ depends on the choice of observation model. Here we consider two different observation models, namely the isotropic Gaussian and the probit models. Both of these models lead to closed form update rules, but this is not true for all choices of $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$. In general if $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ factorizes over n and each term only depends on $\boldsymbol{x}_t$ through $\boldsymbol{A}\boldsymbol{x}_t$, then the resulting moment integrals are 1-dimensional and can be solved relatively fast using numerical integration procedures (Jylänki et al., 2011) if no closed form solution exists.

Under the Gaussian noise model, we have

$$f_{1,t}\left(\boldsymbol{x}_t\right) = p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{y}_t\middle|\boldsymbol{A}\boldsymbol{x}_t, \sigma^2\boldsymbol{I}\right). \tag{36}$$

Thus, $f_{1,t}$ is already in the exponential family for all $t \in [T]$ and does therefore not have to be approximated using EP. In particular, the parameters for $\tilde{f}_{1,t}$ are determined by the relations $\hat{\boldsymbol{V}}_{1,t}^{-1} = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{A}$ and $\hat{\boldsymbol{V}}_{1,t}^{-1}\hat{\boldsymbol{m}}_{1,t} = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{y}_t$. For simplicity we also assume that the noise variance is constant for all t.

Under the probit likelihood the term $f_{1,t}$ decompose to $f_{1,t} = \prod_{n=1}^{N} f_{1,t,n}$. In this case, the update of each site approximation $\tilde{f}_{1,t,n}$ resembles the updates for Gaussian process classification using EP, see appendix C for details.

- Initialize approximation terms $\tilde{f}_a$ for $a = 1, 2, 3, 4$ and Q
- Repeat until stopping criteria
  - For each $\tilde{f}_{1,n,t}$ (*For non-Gaussian likelihoods only*):
    - Compute cavity distribution: $Q^{\setminus 1,n,t} \propto \frac{Q}{\tilde{f}_{1,n,t}}$
    - Minimize: $\mathrm{KL}\big(f_{1,n,t} Q^{\setminus 1,n,t} || Q^{1,t,\mathrm{new}}\big)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{1,n,t} \propto \frac{Q^{1,t,\mathrm{new}}}{Q^{\setminus 1,n,t}}$ to update parameters $\hat{m}_{1,n,t}, \hat{v}_{1,n,t}$ and $\hat{\gamma}_{1,n,t}$.
  - For each $\tilde{f}_{2,i,t}$:
    - Compute cavity distribution: $Q^{\setminus 2,i,t} \propto \frac{Q}{\tilde{f}_{2,i,t}}$
    - Minimize: $\mathrm{KL}\big(f_{2,i,t} Q^{\setminus 2,i,t} || Q^{2,t,\mathrm{new}}\big)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{2,i,t} \propto \frac{Q^{2,t,\mathrm{new}}}{Q^{\setminus 2,i,t}}$ to update parameters $\hat{m}_{2,i,t}, \hat{v}_{2,i,t}$ and $\hat{\gamma}_{2,i,t}$.
  - Update joint approximation parameters: $\hat{\boldsymbol{m}}, \hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{\gamma}}$
  - For each $\tilde{f}_{3,i,t}$:
    - Compute cavity distribution: $Q^{\setminus 3,i,t} \propto \frac{Q}{\tilde{f}_{3,i,t}}$
    - Minimize: $\mathrm{KL}\big(f_{3,i,t} Q^{\setminus 3,i,t} || Q^{3,t,\mathrm{new}}\big)$ w.r.t. $Q^{3,t,\mathrm{new}}$
    - Compute: $\tilde{f}_{3,i,t} \propto \frac{Q^{3,t,\mathrm{new}}}{Q^{\setminus 3,i,t}}$ to update parameters $\hat{\mu}_{3,i,t}, \hat{\sigma}_{3,i,t}$ and $\hat{\gamma}_{3,i,t}$
  - Update joint approximation parameters: $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\gamma}}$
- Compute marginal likelihood approximation

Algorithm 1: Proposed algorithm for approximating the joint posterior distribution over $\boldsymbol{X}, \boldsymbol{Z}$ and $\boldsymbol{\Gamma}$ conditioned on $\boldsymbol{Y}$ using parallel EP.

4.4 Estimating parameters for $\tilde{f}_{2,t}$

The terms $\tilde{f}_{2,t} = \prod_{i=1}^{D} \tilde{f}_{2,i,t}\left(x_{i,t}, z_{i,t}\right)$ factor over i, which implies that we only need the marginal cavity distributions of each pair of $x_{i,t}$ and $z_{i,t}$. Consider the update of the j'th term at time t, that is $\tilde{f}_{2,j,t}\left(x_{j,t}, z_{j,t}\right)$. The first step is to compute the marginal cavity distributions by removing the contribution of $\tilde{f}_{2,j,t}\left(x_{j,t}, z_{j,t}\right)$ from the marginal of the global approximation Q using eq. (20)

$$Q^{\setminus 2,j,t}\left(x_{j,t}, z_{j,t}\right) = \frac{Q^{\setminus 2,j,t}\left(x_{j,t}, z_{j,t}\right)}{\tilde{f}_{2,j,t}\left(x_{j,t}, z_{j,t}\right)} \propto \mathcal{N}\left(x_{j,t} \middle| \mu^{\setminus 2,j,t}, \Sigma^{\setminus 2,j,t}\right) \mathrm{Ber}\left(z_{j,t} \middle| \phi\left(\gamma^{\setminus 2,j,t}\right)\right). \quad (37)$$

When the approximate distribution belongs to the exponential family, the cavity distribution is simply obtained by computing the differences in natural parameters. Expressed in terms of mean and variance, we get

$$\hat{v}^{\setminus 2,j,t} = \left[\hat{V}_{t,jj}^{-1} - \hat{v}_{2,j,t}^{-1}\right]^{-1}, \quad (38)$$

$$\hat{m}^{\setminus 2,j,t} = \hat{v}^{\setminus 2,j,t}\left[\hat{V}_{t,jj}^{-1}\hat{m}_{j,t} - \hat{v}_{2,j,t}^{-1}\hat{m}_{2,j,t}\right], \quad (39)$$

$$\hat{\gamma}^{\setminus 2,j,t} = \hat{\gamma}_{3,j,t}. \quad (40)$$

The cavity parameter for $\gamma_{j,t}$ in $f_{2,j,t}$ is simply equal to $\hat{\gamma}_{3,j,t}$ (and vice versa) since $\hat{\gamma}_{2,j,t}$ and $\hat{\gamma}_{3,j,t}$ are the only two terms contributing to the distribution over $\boldsymbol{z}_{j,t}$. Next, we form the *tilted* distribution $f_{2,j,t}Q^{\setminus 2,j,t}$ and compute the solution to the KL minimization problem in eq. (21) by matching the expected sufficient statistics. This amounts to computing the zeroth, first and second moments w.r.t. $x_{j,t}$

$$X_m = \sum_{z_{j,t}} \int x_{j,t}^m \cdot f_{2,j,t}\left(x_{j,t}, z_{j,t}\right) Q^{\setminus 2,j,t}\left(x_{j,t}, z_{j,t}\right) \mathrm{d}x_{j,t} \quad \text{for} \quad m = 0, 1, 2, \tag{41}$$

and the first moment of $z_{j,t}$

$$Z_1 = \sum_{z_{j,t}} \int z_{j,t} \cdot f_{2,j,t}\left(x_{j,t}, z_{j,t}\right) Q^{\setminus 2,j,t}\left(x_{j,t}, z_{j,t}\right) \mathrm{d}x_{j,t}. \tag{42}$$

For notational convenience we have dropped the dependencies of X_m and Z_1 on the indices t and j. Alternatively, the moments could be obtained by computing the partial derivatives of the log normalizer of the tilted distribution.

The central moments of Q^* in eq.(21) are given by

$$E\left[x_{j,t}\right] = \frac{X_1}{X_0}, \qquad V\left[x_{j,t}\right] = \frac{X_2}{X_0} - \frac{X_1^2}{X_0^2}, \qquad E\left[z_{j,t}\right] = \frac{Z_1}{X_0}. \tag{43}$$

Refer to Appendix A for analytical expressions for these moments. Once Q^* has been obtained, we can compute the new update site approximation for $\tilde{f}_{2,j,t}$ using eq. (22) as follows

$$\tilde{f}_{2,j,t}^*\left(x_{j,t}, z_{j,t}\right) = \frac{Q^*\left(x_{j,t}, z_{j,t}\right)}{Q^{\setminus 2,j,t}\left(x_{j,t}, z_{j,t}\right)} \propto \mathcal{N}\left(x_{j,t} \middle| \hat{m}_{2,j,t}^*, \hat{v}_{2,j,t}^*\right) \mathrm{Ber}\left(z_{j,t} \middle| \phi\left(\hat{\gamma}_{2,j,t}^*\right)\right), \tag{44}$$

where the new site parameters $\hat{m}_{2,j,t}^*$ and $\hat{v}_{2,j,t}^*$, are obtained by computing differences in natural parameters in the same manner as for the cavity parameters in eq. (38) - (40)

$$\hat{v}_{2,j,t}^* = \left[V\left[x_{j,t}\right]^{-1} - \left(\hat{v}^{\setminus 2,j,t}\right)^{-1}\right]^{-1} \tag{45}$$

$$\hat{m}_{2,j,t}^* = \hat{v}_{2,j,t}^* \left[V\left[x_{j,t}\right]^{-1} E\left[x_{j,t}\right] - \left(\hat{v}^{\setminus 2,j,t}\right)^{-1} \hat{m}^{\setminus 2,j,t}\right] \tag{46}$$

The new site parameters for $z_{j,t}$ are obtained as (see Appendix A for details)

$$\phi\left(\hat{\gamma}_{2,j,t}^*\right) \overset{(a)}{=} \frac{\frac{E[z_{j,t}]}{\phi\left(\hat{\gamma}^{\setminus 2,j,t}\right)}}{\frac{1-E[z_{j,t}]}{1-\phi\left(\hat{\gamma}^{\setminus 2,j,t}\right)} + \frac{E[z_{j,t}]}{\phi\left(\hat{\gamma}^{\setminus 2,j,t}\right)}} \overset{(b)}{=} \frac{\mathcal{N}\left(0 \middle| \hat{m}^{\setminus 2,i} - \rho_0, \hat{V}^{\setminus 2,j,t} + \tau_0\right)}{\mathcal{N}\left(0 \middle| \hat{m}^{\setminus 2,i}, \hat{V}^{\setminus 2,i}\right) + \mathcal{N}\left(0 \middle| \hat{m}^{\setminus 2,i} - \rho_0, \hat{V}^{\setminus 2,j,t} + \tau_0\right)}, \tag{47}$$

where (a) follows from forming the quotient of the two Bernoulli distributions and (b) follows from straightforward algebraic reduction after substituting in the expression for the expectation of $z_{j,t}$.

4.5 Estimating parameters for $\tilde{f}_{3,t}$

The procedure for updating $\tilde{f}_{3,t} = \prod_{i=1}^{D} \tilde{f}_{3,j,t}$ is completely analogously to the procedure for $\tilde{f}_{2,t}$. Consider the update for the j'th term at time t, that is $\tilde{f}_{3,j,t}$. After computing the cavity distribution in the same manner as in eq. (38)-(40), we now compute the moments w.r.t. $\gamma_{j,t}$ and $z_{j,t}$ of the (unnormalized) tilted distribution

$$G_m = \sum_{z_{j,t}} \int \gamma_{j,t}^m \cdot f_{3,j,t}\left(z_{j,t}, \gamma_{j,t}\right) Q^{\setminus 3,j,t}\left(z_{j,t}, \gamma_{j,t}\right) \mathrm{d}\gamma_{j,t} \quad \text{for} \quad m = 0, 1, 2, \tag{48}$$

$$Z_1 = \sum_{z_{j,t}} \int z_{j,t} \cdot f_{3,j,t}\left(z_{j,t}, \gamma_{j,t}\right) Q^{\setminus 3,j,t}\left(z_{j,t}, \gamma_{j,t}\right) \mathrm{d}\gamma_{j,t}. \tag{49}$$

Given these moments, we can obtain the central moments for Q^* in eq. (21)

$$E\left[\gamma_{j,t}\right] = \frac{G_1}{G_0}, \qquad V\left[\gamma_{j,t}\right] = \frac{G_2}{G_0} - \frac{G_1^2}{G_0^2}, \qquad E\left[z_{j,t}\right] = \frac{Z_1}{G_0}. \tag{50}$$

Refer to Appendix B for analytical expression of the moments. These moments completely determine Q^* and the j'th site update at the t is computed analogous to $\tilde{f}_{2,j,t}$ in eq. (44) using eq. (45), (46) and (47).

4.6 The computational details

In the previous sections, we have described how to use EP for approximate inference for the proposed model, and in this section, we discuss some of the computational details of the resulting EP algorithm.

4.6.1 Updating the global covariance matrices

Given a set of updated site approximations, $\tilde{f}_{2,t} = \prod_j \tilde{f}_{2,j,t}$, we can compute the parameters for the global approximate distribution of $\boldsymbol{x}_t$ using eq. (31) and (32). Direct evaluation of eq. (31) results in a computational complexity of $\mathcal{O}\left(D^3\right)$. Recall, that N is assumed to be smaller than D. This implies that $\hat{\boldsymbol{V}}_{1,t}^{-1} = \frac{1}{\sigma_0^2}\boldsymbol{A}^T\boldsymbol{A}$ has low rank. Furthermore, the matrix $\hat{\boldsymbol{V}_{2,t}}$ is diagonal, and therefore we can apply the matrix inversion lemma as follows

$$\hat{\boldsymbol{V}}_t = \hat{\boldsymbol{V}}_{2,t} - \hat{\boldsymbol{V}}_{2,t}\boldsymbol{A}^T\left(\sigma_0^2\boldsymbol{I} + \boldsymbol{A}\hat{\boldsymbol{V}}_{2,t}\boldsymbol{A}^T\right)^{-1}\boldsymbol{A}\hat{\boldsymbol{V}}_{2,t}. \tag{51}$$

The inverse of $\sigma_0^2\boldsymbol{I} + \boldsymbol{A}\hat{\boldsymbol{V}}_{2,t}\boldsymbol{A}^T = \boldsymbol{L}_t\boldsymbol{L}_t^T$ can be computed in $\mathcal{O}\left(N^3\right)$ using a Cholesky decomposition. Thus, for $N < D$ eq. (51) scales as $\mathcal{O}\left(ND^2\right)$. Moreover, eq. (38) shows that we only require the diagonal elements of $\tilde{\boldsymbol{V}}_t$ in order to update the site approximation parameters for $\tilde{f}_{2,t}$. Hence, we can further reduce the computational complexity by only computing the diagonal of $\hat{\boldsymbol{V}}_t$ as follows

$$\begin{aligned}
\operatorname{diag}\left[\hat{\boldsymbol{V}}_t\right] &= \operatorname{diag}\left[\hat{\boldsymbol{V}}_{2,t}\right] - \operatorname{diag}\left[\hat{\boldsymbol{V}}_{2,t}\boldsymbol{A}^T\boldsymbol{L}_t^{-T}\boldsymbol{L}_t^{-1}\boldsymbol{A}\hat{\boldsymbol{V}}_{2,t}\right] \\
&= \operatorname{diag}\left[\hat{\boldsymbol{V}}_{2,t}\right] - \operatorname{diag}\left[\hat{\boldsymbol{V}}_{2,t}^2\right] \circ \left(\boldsymbol{1}^T\left(\boldsymbol{R}_t \circ \boldsymbol{R}_t\right)\right),
\end{aligned} \tag{52}$$

where $\boldsymbol{R}_t \in \mathbb{R}^{N \times D}$ is defined as $\boldsymbol{R}_t = \boldsymbol{L}_t^{-1}\boldsymbol{A}$ and $\mathbf{1}$ is a column vector of ones. The resulting computational cost is $\mathcal{O}\left(N^2 D\right)$. Similarly, the mean of the global approximate distribution of $\boldsymbol{x}_t$, can be efficiently evaluated as

$$\hat{\boldsymbol{m}}_t = \hat{\boldsymbol{V}}_{2,t}\boldsymbol{\eta}_t - \hat{\boldsymbol{V}}_{2,t}\boldsymbol{R}_t^T \boldsymbol{R}_t \hat{\boldsymbol{V}}_{2,t}\boldsymbol{\eta}_t, \tag{53}$$

where $\boldsymbol{\eta}_t = \hat{\boldsymbol{V}}_{1,t}^{-1}\hat{\boldsymbol{m}}_{1,t} + \hat{\boldsymbol{V}}_{2,t}^{-1}\hat{\boldsymbol{m}}_{2,t}$. The total cost of updating the posterior distribution for $\boldsymbol{x}_t$ for all $t \in [T]$ is therefore $\mathcal{O}\left(TN^2 D\right)$.

Unfortunately, we cannot get the same speed up for the refinement of the global approximation of $\boldsymbol{\Gamma}$ since the prior covariance matrix $\boldsymbol{\Sigma}_0$ in general is full rank. However, we still only require the diagonal elements of the approximate covariance matrix $\hat{\boldsymbol{\Sigma}}$. We implement the update as advocated by Rasmussen and Williams (2006), that is,

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= \left[\boldsymbol{\Sigma}_0^{-1} + \hat{\boldsymbol{\Sigma}}_3^{-1}\right]^{-1} \\ &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\hat{\boldsymbol{\Sigma}}_3^{-\frac{1}{2}}\left(\hat{\boldsymbol{\Sigma}}_3^{-\frac{1}{2}}\boldsymbol{\Sigma}_0\hat{\boldsymbol{\Sigma}}_3^{-\frac{1}{2}} + \boldsymbol{I}\right)^{-1}\hat{\boldsymbol{\Sigma}}_3^{-\frac{1}{2}}\boldsymbol{\Sigma}_0,\end{aligned} \tag{54}$$

where the second equality follows from the matrix inverse lemma. Again, we compute the required inverse matrix using the Cholesky decomposition, so that the total cost is $\mathcal{O}\left(D^3T^3\right)$.

4.6.2 Initialization, Convergence and negative variances

We initialize all the site terms to be rather uninformative, that is $\hat{m}_{2,i,t} = 0$, $\hat{v}_{2,i,t} = 10^4$, $\hat{\gamma}_{2,i,t} = 0$, $\hat{\mu}_{3,i,t} = 0$, $\hat{\sigma}_{3,i,t} = 10^4$, $\hat{\gamma}_{3,i,t} = 0$ for all $i \in [D]$ and $t \in [T]$ assuming standard scaling of the forward model $\boldsymbol{A}$.

There are in general no convergence guarantees for EP and the parallel version in particular can suffer from convergence problems (Seeger, 2005). The standard procedure to overcome this problem is to use "damping" when updating the site parameters

$$\tilde{f}^* = \tilde{f}_{\text{old}}^{1-\alpha}\tilde{f}_{\text{new}}^{\alpha}, \tag{55}$$

where $\alpha \in [0, 1]$ is the damping parameter and $\tilde{f}_{\text{old}}$ is the site approximation at the previous iteration. Since both $\tilde{f}_{\text{old}}$ and $\tilde{f}_{\text{new}}$ belongs to the exponential family, the update in eq. (55) corresponds to taking a convex combination of the previous and the new natural parameters of the site approximation.

Negative variances occur "naturally" in EP (Bishop, 2006) when updating the site approximations. However, this can lead to instabilities of the algorithm, non-positive semi-definitiveness of the posterior covariance matrices and convergence problems. We therefore take measures to prevent negative site variances. One way to circumvent this is to change a negative variance to $+\infty$, which corresponds to minimizing the KL divergence in eq. (21) with the site variance constrained to be positive (Hernández-Lobato et al., 2013). In practice, when encountering a negative variance after updating a given site we use $v_\infty = 10^2$ and $\sigma_\infty = 10^6$ for $\tilde{f}_{2,i,t}$ and $\tilde{f}_{3,i,t}$, respectively.

5. Further Approximations

As mentioned earlier, the updates of the global parameters for $\boldsymbol{x}_t$ and $\boldsymbol{\Gamma}$ are the dominating operations scaling as $\mathcal{O}\left(TN^2D\right)$ and $\mathcal{O}\left(D^3T^3\right)$, respectively. The latter term becomes prohibitive for moderate sizes of D and T and calls for further approximations. In this section, we introduce three simple approximations to reduce the computational complexity of the refinement of the posterior distribution for $\boldsymbol{\Gamma}$. The approximations and their computational complexities are summarized in table 1.

Approximation		**Complexity**	**Storage**
Full EP	(EP)	$\mathcal{O}\left(T^3D^3\right)$	$\mathcal{O}\left(T^2D^2\right)$
Low rank	(LR)	$\mathcal{O}\left(K^2TD\right)$	$\mathcal{O}\left(KTD\right)$
Common precision	(CP)	$\mathcal{O}\left(TD^2 + DT^2\right)$	$\mathcal{O}\left(D^2 + T^2\right)$
Group	(G)	$\mathcal{O}\left(T_g^3D_g^3\right)$	$\mathcal{O}\left(T_g^2D_g^2\right)$

Table 1: Summary of approximation schemes for updating the global parameters for $\boldsymbol{\Gamma}$.

5.1 The low rank approximation

The eigenvalue spectrum of many prior covariance structures of interest, for example simple neighborhoods, decay relatively fast. Therefore, we can approximate $\boldsymbol{\Sigma}_0$ with a low rank approximation plus a diagonal matrix $\boldsymbol{\Sigma}_0 \approx \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T + \boldsymbol{\Lambda}$, where $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is a diagonal matrix containing K largest eigenvalues, and $\boldsymbol{U} \in \mathbb{R}^{DT\times K}$ is a matrix containing the corresponding eigenvectors (Riihimäki et al., 2014). The diagonal matrix $\boldsymbol{\Lambda}$ is chosen such that the diagonal in the exact prior covariance matrix $\boldsymbol{\Sigma}_0$ is preserved. This allows us to apply the matrix inversion lemma to compute the update of the posterior covariance matrix for $\boldsymbol{\Gamma}$ (see section 4.6.1).

Computing the eigendecomposition of $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{DT\times DT}$ scales in general as $\mathcal{O}\left(D^3T^3\right)$. However, when the prior covariance has Kronecker structure, the eigendecompositions of $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_t \otimes \boldsymbol{\Sigma}_s$ can be efficiently obtained from the eigendecompositions of $\boldsymbol{\Sigma}_t \in \mathbb{R}^{T\times T}$ and $\boldsymbol{\Sigma}_s \in \mathbb{R}^{D\times D}$. In this case, the eigendecomposition of $\boldsymbol{\Sigma}_0$ can be obtained in $\mathcal{O}\left(D^3 + T^3\right)$.

Using a K-rank approximation, the computational cost of refining the covariance matrix for $\boldsymbol{\Gamma}$ becomes $\mathcal{O}\left(K^2DT\right)$ and the memory footprint is $\mathcal{O}\left(TDK\right)$. For a fixed value of K this scales linearly in both D and T. However, to maintain a sufficiently good approximation K can scale with both D and T.

5.2 The common precision approximation

Rather than approximating the prior covariance matrix as done in the low rank approximation, we now approximate the EP approximation scheme itself. If the prior covariance matrix for $\boldsymbol{\Gamma}$ can be written in terms of Kronecker products, we can significantly speed up the computation of the posterior covariance matrix of $\boldsymbol{\Gamma}$ by approximating the site precisions with a single common parameter. Let $\tilde{\boldsymbol{\theta}}_3 \in \mathbb{R}^{DT\times 1}$ be a vector containing the site precisions (inverse variances) for the site approximations $\{f_{3,i,t}\}$ for all $i \in [D]$ and for all $t \in [T]$, then

we make the following approximation

$$\tilde{\boldsymbol{\Sigma}}_3 \approx \bar{\theta}^{-1}\boldsymbol{I}, \tag{56}$$

where $\bar{\theta}$ is the mean of value of $\tilde{\boldsymbol{\theta}}_3$. Assume the prior covariance matrix for $\boldsymbol{\Gamma}$ can be decomposed into a temporal part and a spatial part as follows $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_t \otimes \boldsymbol{\Sigma}_s$. Let $\boldsymbol{U}_t$, $\boldsymbol{U}_s$ and $\boldsymbol{S}_t$, $\boldsymbol{S}_s$ be eigenvectors and eigenvalues for $\boldsymbol{\Sigma}_t \in \mathbb{R}^{T\times T}$ and $\boldsymbol{\Sigma}_s \in \mathbb{R}^{D\times D}$, respectively. The global covariance matrix is updated as $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \tilde{\boldsymbol{\Sigma}}_3\right)^{-1} \tilde{\boldsymbol{\Sigma}}_3$. We now use the properties of eigendecompositions for Kronecker products to compute the inverse matrix

$$\begin{aligned}
\left(\boldsymbol{\Sigma}_t \otimes \boldsymbol{\Sigma}_s + \tilde{\boldsymbol{\Sigma}}_3\right)^{-1} &\approx \left(\boldsymbol{\Sigma}_t \otimes \boldsymbol{\Sigma}_s + \bar{\Sigma}_3\boldsymbol{I}\right)^{-1} \\
&= \left[\left(\boldsymbol{U}_t \otimes \boldsymbol{U}_s\right)\left(\boldsymbol{S}_t \otimes \boldsymbol{S}_s\right)\left(\boldsymbol{U}_t^T \otimes \boldsymbol{U}_s^T\right) + \bar{\Sigma}_3\boldsymbol{I}\right]^{-1} \\
&= \left(\boldsymbol{U}_t \otimes \boldsymbol{U}_s\right)\left(\boldsymbol{S}_t \otimes \boldsymbol{S}_s + \bar{\Sigma}_3\boldsymbol{I}\right)^{-1}\left(\boldsymbol{U}_t^T \otimes \boldsymbol{U}_s^T\right),
\end{aligned} \tag{57}$$

where $\left(\boldsymbol{S}_t \otimes \boldsymbol{S}_s + \bar{\Sigma}_3\boldsymbol{I}\right)$ is diagonal and therefore fast to invert. The *common precision* approximation $\hat{\boldsymbol{\Sigma}}_{CP}$ is then obtained as

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{CP} &= \left(\boldsymbol{\Sigma}_t \otimes \boldsymbol{\Sigma}_s\right)\left(\boldsymbol{\Sigma}_t \otimes \boldsymbol{\Sigma}_s + \bar{\Sigma}_3\boldsymbol{I}\right)^{-1}\bar{\Sigma}_3\boldsymbol{I} \\
&= \left(\boldsymbol{U}_t \otimes \boldsymbol{U}_s\right)\left(\boldsymbol{S}_t \otimes \boldsymbol{S}_s\right)\left(\boldsymbol{S}_t \otimes \boldsymbol{S}_s + \bar{\Sigma}_3\boldsymbol{I}\right)^{-1}\left(\boldsymbol{U}_t^T \otimes \boldsymbol{U}_s^T\right)\bar{\Sigma}_3.
\end{aligned} \tag{58}$$

Let $\boldsymbol{M} \in \mathbb{R}^{TD\times 1}$ denote the diagonal of $\left(\boldsymbol{S}_t \otimes \boldsymbol{S}_s\right)\left(\boldsymbol{S}_t \otimes \boldsymbol{S}_s + \bar{\Sigma}_3\boldsymbol{I}\right)^{-1}$, then we can compute the diagonal of $\hat{\boldsymbol{\Sigma}}_{CP}$ as follows

$$\begin{aligned}
\operatorname{diag}\left[\hat{\boldsymbol{\Sigma}}_{CP}\right]_i &= \bar{\Sigma}_3 \sum_k \left(\boldsymbol{U}_t \otimes \boldsymbol{U}_s\right)_{ik} M_k \left(\boldsymbol{U}_t^T \otimes \boldsymbol{U}_s^T\right)_{ki} \\
&= \bar{\Sigma}_3 \sum_k \left(\boldsymbol{U}_t \otimes \boldsymbol{U}_s\right)_{ik}^2 M_k \\
\Rightarrow \quad \operatorname{diag}\left[\hat{\boldsymbol{\Sigma}}_{CP}\right] &= \bar{\Sigma}_3 \left(\boldsymbol{U}_t \circ \boldsymbol{U}_t \otimes \boldsymbol{U}_s \circ \boldsymbol{U}_s\right)\boldsymbol{M},
\end{aligned} \tag{59}$$

where $\circ$ is the Hadamard-product. We now see that the desired diagonal can be obtained by multiplying a Kronecker product with a vector and this can be computed efficiently using the identity

$$\operatorname{vec}\left[\boldsymbol{ABC}\right] = \left(\boldsymbol{C}^T \otimes \boldsymbol{A}\right)\operatorname{vec}\left[\boldsymbol{B}\right]. \tag{60}$$

Therefore,

$$\operatorname{diag}\left[\hat{\Sigma}_{CP}\right] = \bar{\Sigma}_3 \cdot \operatorname{vec}\left[\left(\boldsymbol{U}_s \circ \boldsymbol{U}_s\right)\operatorname{vec}^{-1}\left[\boldsymbol{M}\right]\left(\boldsymbol{U}_t \circ \boldsymbol{U}_t\right)^T\right]. \tag{61}$$

Since the Hadamard products can be precomputed, this scales as $\mathcal{O}\left(D^2T + T^2D\right)$. During the EP iterations we only need to store $\boldsymbol{U}_s \in \mathbb{R}^{D\times D}$ and $\boldsymbol{U}_t \in \mathbb{R}^{T\times T}$, so the resulting memory footprint is $\mathcal{O}\left(D^2 + T^2\right)$. The posterior mean vector can also be computed efficiently by iteratively applying the result from eq. (60)

$$\hat{\boldsymbol{\Sigma}}_{CP}\boldsymbol{\eta} = \left(\boldsymbol{U}_t \otimes \boldsymbol{U}_s\right)\operatorname{diag}\left[M\right]\left(\boldsymbol{U}_t^T \otimes \boldsymbol{U}_s^T\right)\boldsymbol{\eta}, \tag{62}$$

where $\boldsymbol{\eta} = \hat{\boldsymbol{\Sigma}}_3^{-1}\hat{\boldsymbol{\mu}}_3 + \hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{\mu}}_0$.

The proposed approximation reduces the cost from $\mathcal{O}\left(D^3T^3\right)$ to $\mathcal{O}\left(D^2T + T^2D\right)$. If the spatial covariance matrix is a Kronecker product itself, for example, $\boldsymbol{\Sigma}_s = \boldsymbol{\Sigma}_x \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}}$ or $\boldsymbol{\Sigma}_s = \boldsymbol{\Sigma}_x \otimes \boldsymbol{\Sigma}_y \otimes \boldsymbol{\Sigma}_z$, the computational complexity can be further reduced. Such covariance structures could occur in image application or in analysis of fMRI data.

This common precision approximation is closely related to the recently proposed *Stochastic Expectation Propagation* (SEP) (Li et al., 2015), where both the means and variances of the site approximation terms have been tied together. Tying both means and variances is reasonable when the site terms are approximating likelihood terms and $N \gg D$. In case of the present model, we expect positive values of $\Gamma_{i,t}$ for $z_{i,t} = 1$ and negative values of $\Gamma_{i,t}$ for $z_{i,t} = 0$, and thus enforcing a common mean for the site approximation terms $\tilde{f}_{3,i,t}$ would not make sense.

From experiments we have observed that this common precision approach significantly increases the number of iterations until convergence. However, this problem can be mitigated by repeating the updates for the site approximations $\tilde{f}_{3,i,t}$ and the global approximation for $\boldsymbol{\Gamma}$ a few times before moving on to update the site approximations for $\tilde{f}_{2,i,t}$. Specifically, within each EP iteration we repeat the updates for posterior distribution of $\boldsymbol{\Gamma}$ 5 times. The added computational workload is still negligible compared to full EP. Furthermore, for some problem instances CP-EP can oscillate. The oscillation can be alleviated heuristically by decreasing the damping parameter α by 10% if the approximate log likelihood decreases from one iteration to another after the first 100 iterations.

5.3 Grouping the latent structure variables

Consider a problem, where the spatial coordinates $\boldsymbol{d}_i$ for each $x_{i,\cdot}$ form a uniformly spaced grid. Assume the characteristic length-scale of the sparsity pattern is large relative to the grid size, then support variables $\{z_i\}$ in a neighborhood could "share" the same γ-variable with a little loss of accuracy (Jacob et al., 2009a; Hernández-Lobato et al., 2013). This *grouping* of the latent variables could either be in the spatial, temporal or both dimensions. Let G be the number of groups and $g : [D] \times [T] \to [G]$ be a grouping function that maps from a spatial and temporal index to a group index, then the grouped version of the prior is given by

$$p(\boldsymbol{Z}|\boldsymbol{\gamma}) = \prod_{t=1}^{T}\prod_{i=1}^{D} \mathrm{Ber}\left(z_{t,i}|\phi\left(\gamma_{g(i,t)}\right)\right), \tag{63}$$

$$p(\boldsymbol{\gamma}) = \mathcal{N}\left(\boldsymbol{\gamma}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right), \tag{64}$$

where $\boldsymbol{\mu}_0 \in \mathbb{R}^G$ and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{G\times G}$ are the prior mean and covariance for the new grouped model. The resulting computational complexity is indeed determined by the size of the groups. For example, assume that the support variable for a given problem have been grouped in groups of 2 in both the spatial dimension and temporal dimension, then the total number of groups becomes $G = \frac{1}{2}D\frac{1}{2}T = \frac{1}{4}DT$ and the resulting computational cost is reduced to a fraction of $\left(\frac{1}{4}\right)^3$ of the cost of the full EP scheme. Furthermore, if necessary

both the low rank and the common precision approximation can be applied on top of this approximation.

6. The marginal likelihood approximation and model selection

The model contains several hyperparameters $\boldsymbol{\Omega} \in \mathbb{R}^L$, which include, for example, the hyperparameters of the kernel for $\boldsymbol{\Gamma}$. In a fully Bayesian setting, the natural approach to handle hyperparameters is to impose prior distributions and marginalize over the hyperparameters. The exact, but generally intractable marginalization integral is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}) = \int p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}) p\left(\boldsymbol{\Omega}\middle|\boldsymbol{Y}\right) \mathrm{d}\boldsymbol{\Omega}, \tag{65}$$

where $p\left(\boldsymbol{\Omega}\middle|\boldsymbol{Y}\right) \propto p(\boldsymbol{Y}\middle|\boldsymbol{\Omega}) p(\boldsymbol{\Omega})$ for some prior distribution $p(\boldsymbol{\Omega})$. The true marginal likelihood $p(\boldsymbol{Y}\middle|\boldsymbol{\Omega})$ is given by the following marginalization

$$p(\boldsymbol{Y}\middle|\boldsymbol{\Omega}) = \int f_1\left(\boldsymbol{X}\middle|\boldsymbol{\Omega}\right) \sum_{\boldsymbol{Z}} f_2\left(\boldsymbol{X}, \boldsymbol{Z}\middle|\boldsymbol{\Omega}\right) \mathrm{d}\boldsymbol{X} \int f_3\left(\boldsymbol{Z}, \boldsymbol{\Gamma}\middle|\boldsymbol{\Omega}\right) f_4\left(\boldsymbol{\Gamma}\middle|\boldsymbol{\Omega}\right) \mathrm{d}\boldsymbol{\Gamma}. \tag{66}$$

The exact quantity is intractable, but the EP framework provides an approximation to the marginal likelihood conditioned on the hyperparameters, $p(\boldsymbol{Y}\middle|\boldsymbol{\Omega}) \approx Q(\boldsymbol{Y}\middle|\boldsymbol{\Omega})$. The approximation $Q(\boldsymbol{Y}\middle|\boldsymbol{\Omega})$ is obtained by substituting the exact site terms, for example, $f_{2,i,t}$, with a scaled version of the corresponding site approximation, for example, $s_{2,i,t}\tilde{f}_{2,i,t}$, and then carrying out the marginalization analytically. The scaling constants, for example, $s_{2,i,t}$, are chosen such that

$$\mathbb{E}_{Q^{\setminus 2,i,t}}\left[f_{2,i,t}\left(x_{i,t}, z_{i,t}\right)\right] = s_{2,i,t}\mathbb{E}_{Q^{\setminus 2,i,t}}\left[\tilde{f}_{2,i,t}\left(x_{i,t}, z_{i,t}\right)\right] \tag{67}$$

and similarly for all the site terms $f_{a,i,t}$ for $a \in [4]$, $i \in [D]$, $t \in [T]$. In the following, we will describe three different approximation strategies based on the marginal likelihood approximation.

6.1 Maximum Likelihood and MAP estimation

This simplest and most crude approximation is to use a point estimate of $\boldsymbol{\Omega}$ instead of integrating over the uncertainty. Specifically, we aim to locate the maximum a posteriori (MAP) value by maximizing $\ln Q\left(\boldsymbol{\Omega}|\boldsymbol{Y}\right) = \ln Q\left(\boldsymbol{Y}\middle|\boldsymbol{\Omega}\right) + \ln p\left(\boldsymbol{\Omega}\right) + \text{constant}$ using gradient-based methods. A maximum likelihood type II estimate is obtained by choosing an (improper) flat prior $p\left(\boldsymbol{\Omega}\right) \propto 1$. For severely ill-posed problems, the marginal likelihood approximation can be completely non-informative with regard to one or more hyperparameters and thus, the maximum likelihood estimate can lead to suboptimal and unstable results for some problems. For some problem instances, it can also happen that the marginal likelihood solution with regard to the prior mean and variance of $\boldsymbol{\Gamma}$ is not in the interior of $\mathbb{R}^2$ and thus, gradient-based optimization with regard to these parameters will diverge. However, this problem is easily solved by imposing a weakly informative prior on the prior variance of $\boldsymbol{\Gamma}$ with little influence on the result (see Appendix D for more details).

The marginal likelihood approximation, $Q\left(\boldsymbol{Y}\middle|\boldsymbol{\Omega}\right)$, depends on the hyperparameters $\boldsymbol{\Omega}$ directly as well as through the site parameters, but the latter dependency can be ignored in gradients computations when the EP fixed point conditions hold (Seeger, 2005). The hyperparameter optimization procedure proceeds in an iterative two-stage fashion, where we first run EP until convergence and then we take a gradient step with regard to the hyperparameters and then repeat.

6.2 Approximate marginalization using numerical integration

As a better approximation of eq. (65), we propose to approximate the marginalization integral using numerical integration with a finite sum using a central composite design (CCD) grid (Rue et al., 2009). This method has previously been successfully applied for marginalizing over hyperparameters in Gaussian process based models and the accuracy is reported to be between empirical Bayes and full marginalization using a dense grid (Vanhatalo et al., 2010). We approximate the marginal posterior distribution as follows

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}) \approx \int Q(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}) Q\left(\boldsymbol{\Omega}\middle|\boldsymbol{Y}\right) \mathrm{d}\boldsymbol{\Omega} \tag{68}$$

$$\approx \sum_{m=1}^{M} Q(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_m) Q\left(\boldsymbol{\Omega}_m\middle|\boldsymbol{Y}\right) w_m, \tag{69}$$

for a set of points $\{\boldsymbol{\Omega}_m\}_{m=1}^M$, a set of integration weights $\{w_m\}_{m=1}^M$. Thus, the resulting approximate marginal posterior distribution becomes a Gaussian mixture model with mixing weights $\pi_m = Q\left(\boldsymbol{\Omega}_m\middle|\boldsymbol{Y}\right) w_m$ and components $Q(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Gamma}|\boldsymbol{Y}, \boldsymbol{\Omega}_m)$.

To keep the computational burden to a minimum, we use a so-called Central Composite Design (CCD) to choose the points and weights. Most of the hyperparameters are variance or scale parameters and hence constrained to be positive. Therefore, we first transform these parameters into an unconstrained space using a log transformation, $\lambda_i = \ln \Omega_i$. Next, we locate the mode in the transformed parameter space, $\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}}$, by optimizing $Q\left(\boldsymbol{Y}\middle|\boldsymbol{\lambda}\right)$ with regard to $\boldsymbol{\lambda}$ using gradient-based optimization methods and numerically estimate the inverse Hessian, $\hat{\boldsymbol{S}} = \boldsymbol{H}^{-1}$, at the mode $\hat{\boldsymbol{\lambda}}_{\mathrm{MAP}}$.

The CCD integration points are then obtained as $\boldsymbol{\lambda}_m = \hat{\boldsymbol{S}}^{\frac{1}{2}} \boldsymbol{p}_m + \hat{\boldsymbol{\lambda}_{\mathrm{MAP}}}$, where $\{\boldsymbol{p}_m\}_{m=1}^M$ is a CCD design grid (Rue et al., 2009) in L-dimensions. The points on the CCD grid consist of a fractional factorial design as well as $2K$ star points and a center point. All points, except for the center point, lives on the surface of a L-dimensional ball with radius $\sqrt{L}$. This specific design choice requires a much smaller number of points compared to a dense grid. For example, for $L = 2, 3, 4, 5$ parameters, the number of CCD points are $M = 9, 15, 25, 43$, respectively. The integration weights $\{w_m\}_{m=1}^M$ are chosen such that the integral match for the first three moments of a L-dimensional standardized Gaussian random variable, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_L)$, and $\mathbb{E}\left[1\right] = 1$, $\mathbb{E}\left[\boldsymbol{z}\right] = 0$, $\mathbb{E}\left[\boldsymbol{z}^T\boldsymbol{z}\right] = L$.

6.3 Bayesian Optimization

There can be some challenges with gradient-based optimization of the marginal likelihood approximation. Firstly, the optimization problem is in general non-convex and thus, the results can suffer from poor local optima. Secondly, for some problem instances the marginal likelihood approximate can exhibit discontinuities (as discussed in experiment 7.1).

To counter these issues, we consider Bayesian optimization (Shahriari et al., 2016; Snoek et al., 2012) as a third strategy to model selection as it does not depend on gradient information. As indicated by the name, Bayesian optimization is a probabilistic approach to optimization, where the objective function is modelled as a random function. Thus, the approach allows us to model the potential discontinuities. Specifically, we use a Gaussian process to model log posterior density as follows

$$\ln Q(\mathbf{\Omega}|Y) \sim \mathcal{GP}\left(\mu\left(\mathbf{\Omega}\right), k\left(\mathbf{\Omega}, \mathbf{\Omega}'\right)\right), \tag{70}$$

where $\mu : \mathbb{R}^K \to \mathbb{R}$ is a mean function and $k : \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$ is the kernel function. Rather than following the direction of the gradient, Bayesian optimization works by exploring values of $\mathbf{\Omega}$, that are likely to improve the value of the objective function as measured by a so-called acquisition function. For more details on Bayesian optimization, we refer to (Shahriari et al., 2016; Snoek et al., 2012; Brochu et al., 2010) .

7. Numerical Experiments

In this section, we conduct a series of experiments designed to investigate the properties of the proposed model and the associated EP inference scheme.

We describe seven experiments with a Gaussian observation model and one experiment with a probit observation model. In the first five experiments, we focus on problem instances with a single measurement vector. Experiment 1 investigates the effect of the prior by analyzing a synthetic data set with a range of different values for the hyperparameters. In the second experiment, we compare the three different approximation schemes (low rank, common precision, group) to standard EP. Specifically, we analyze a synthetic data set with all four methods and compare the results. Experiment 3 is designed to investigate how the EP algorithms perform as a function of the undersampling ratio N/D giving rise to the so-called *phase transition curves* (Donoho and Tanner, 2010). In experiment 4, we apply the proposed model to a compressed sensing problem and in experiment 5, we apply our model to a binary classification task, where the goal is to discriminate between utterances of two different vowels using log-periodograms as features.

In Experiment 6-8, we turn our attention to problems with multiple measurement vectors. In the sixth experiment, we qualitatively study the properties of the proposed methods in the multiple measurement vector setting. We demonstrate the benefits of modeling both the spatial and temporal structure of the support and discuss the marginal likelihood approximation for hyperparameter tuning. Experiment 7 studies the performance of the EP algorithms as a function of the undersampling ratio when multiple measurement vectors are available and compare the results to competing methods. Finally, in Experiment 8 we apply

the proposed method to an EEG source localization problem (Baillet et al., 2001).

For the subset of experiments, where the ground truth solutions are available, we use the *Normalized Mean Square Error (NMSE)* and the *F-measure* (Rijsbergen, 1979) to quantify the performance of the algorithms. In particular, we compute the NMSE between the posterior mean $\hat{\boldsymbol{X}} = \mathbb{E}_{Q(\boldsymbol{X}|\boldsymbol{Y})}[\boldsymbol{X}]$ and the true solution $\boldsymbol{X}_0$ to quantify the algorithms abilities to reconstruct the true signal $\boldsymbol{X}_0$

$$\text{NMSE} = \frac{\|\hat{\boldsymbol{X}} - \boldsymbol{X}_0\|_{\text{F}}^2}{\|\boldsymbol{X}_0\|_{\text{F}}^2}, \tag{71}$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm. We use the F-measure to quantify the algorithms' abilities to recover the true support set

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{72}$$

where *precision* (positive predictive value) is the fraction of non-zero weights found by the algorithm that are also non-zero in the true model, while recall *sensitivity* is the fraction of non-zeros in the true model that have been identified by the algorithm. Here a given weight $x_{i,t}$ is identified as being non-zero if the posterior mean of $z_{i,t}$ is above 0.5.

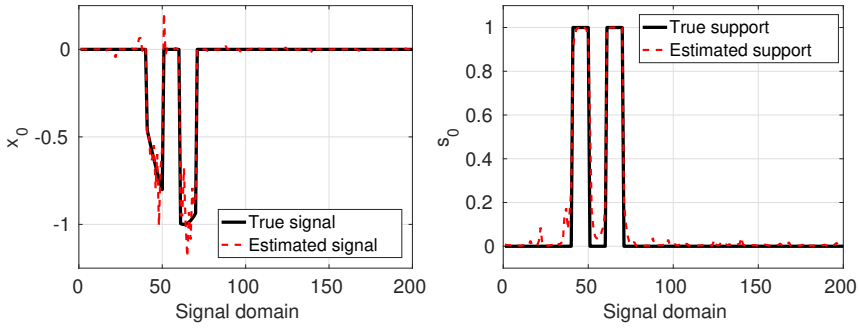
The code is available at `https://github.com/MichaelRiis/SSAS`.

7.1 Experiment 1: The effect of the prior

In this experiment, we investigate the effect of the structured spike-and-slab prior on the reconstructed support set. For simplicity we only consider spatial structure, $T = 1$, and we further assume that the spatial coordinates are on a regular 1D grid. We construct a sparse 1D test signal $\boldsymbol{x}_0 \in \mathbb{R}^{200}$, where the active coefficients are sampled from a cosine function, see Figure 1(a)–(b). Based on this signal we generate a synthetic data set using the linear model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}$, where $A_{ij} \sim \mathcal{N}(0, 1)$, $\boldsymbol{e} \sim \mathcal{N}(0, 5\boldsymbol{I})$ is isotropic Gaussian noise (SNR $\approx$ 5dB) and the number of samples is $N = 0.5D$. The prior on $\boldsymbol{\gamma}$ is of the form $p(\boldsymbol{\gamma}) = \mathcal{N}(\boldsymbol{\gamma}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \nu \cdot \boldsymbol{1} \quad \text{and} \quad \boldsymbol{\Sigma}_{ij} = \kappa_1^2 \exp\left(-\frac{D_{ij}^2}{2\kappa_2^2}\right).$$

We sample the length-scale κ_2 equidistantly 100 times in $\left[10^{-3}, 50\right]$ and run the algorithm on the synthetic data set for each value of κ_2. For this experiment we use the standard EP scheme with no further approximations. The noise variance is fixed to the true value and the remaining hyperparameters are fixed $\nu = 0$, $\tau = 1, \kappa_1^2 = 5$. The posterior results are shown in the panels in leftmost column in Figure 2. The topmost panel shows the marginal likelihood approximation as a function of the spatial length scale κ_2. The panel in the middle shows the posterior mean $\mathbb{E}_{Q(z_i|\boldsymbol{y})}[\gamma_i]$, as a function of the scale κ_2. That is, each column in the image corresponds to the posterior mean of $\boldsymbol{\gamma}$ for a specific value of κ_2. The panel in the bottom shows a similar plot for the posterior support probabilities $\mathbb{E}_{Q(z_i|\boldsymbol{y})}[z_i]$.



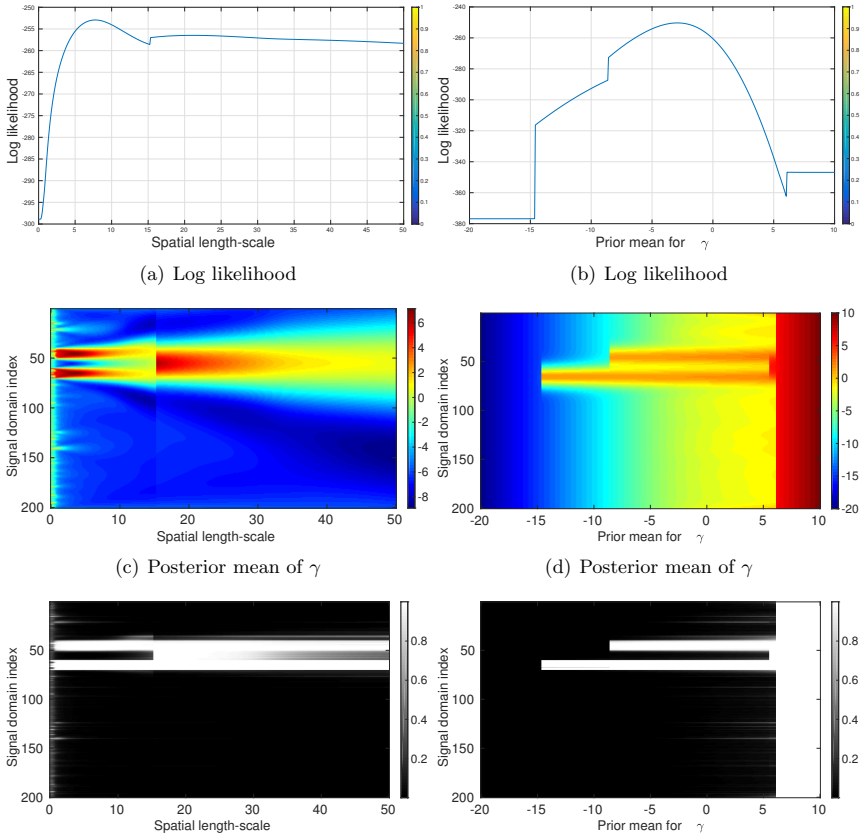
(a) Signal (b) Support

Figure 1: (a) Synthetic test signal $\boldsymbol{x}_0$ superimposed with the posterior mean of the test signal. The active coefficients are sampled from a cosine function (b) The support of the test signal superimposed with the posterior support probabilities.

When κ_2 is close to zero the posterior mean vectors for both $\boldsymbol{\gamma}$ and $\boldsymbol{z}$ are very irregular and resemble the solution of an independent spike-and-slab prior. As the length-scale increases the posterior mean vector $\boldsymbol{\gamma}$ becomes more and more smooth and eventually give rise to well-defined clusters in the support. The algorithm recovers the correct support for $\kappa_2 \in [3, 15]$. However, at $\kappa_2 \approx 15$ a discontinuity is seen. Since the prior distribution on $\boldsymbol{z}$ does not exhibit any phase transitions with regard to κ_2, this is likely to be an effect of a unimodal approximation to a highly multimodal distribution. The discontinuity is also present in the marginal likelihood approximation as seen in the top panel and therefore one should be cautious when optimizing the marginal likelihood using line search based methods. We repeated this experiment for multiple realizations of the noise and the discontinuity was only present occasionally.

The rightmost column in Figure 2 shows equivalent figures for a sweep over ν, which is the prior mean of γ_i, where it is seen that the algorithm recovers the correct support for $\nu \in [-15, 0]$. It is seen that when ν is below some threshold ν_{lower}, the posterior mean of z_i is close to zero for all $i \in [D]$. The total number of active variables increases with ν, until ν surpasses an upper threshold ν_{upper}, where all variables are included in the support set. It is also seen that variables are included cluster-wise rather than individually, which gives rise to discontinuities in the marginal likelihood in the topmost panel.

Figure 1 shows the estimated signal and the estimated support probabilities for the optimal hyperparameter values in the top row of Figure 2, that is the prior mean $\nu = -2.93$ and lengthscale $\kappa_2 = 7.72$, where it is seen that both the estimated coefficients $\hat{\boldsymbol{x}}$ and the estimated support $\hat{\boldsymbol{s}}$ are high-quality approximations of the true quantities. We will make these relationships more quantitative in experiment 3.



(a) Log likelihood (b) Log likelihood

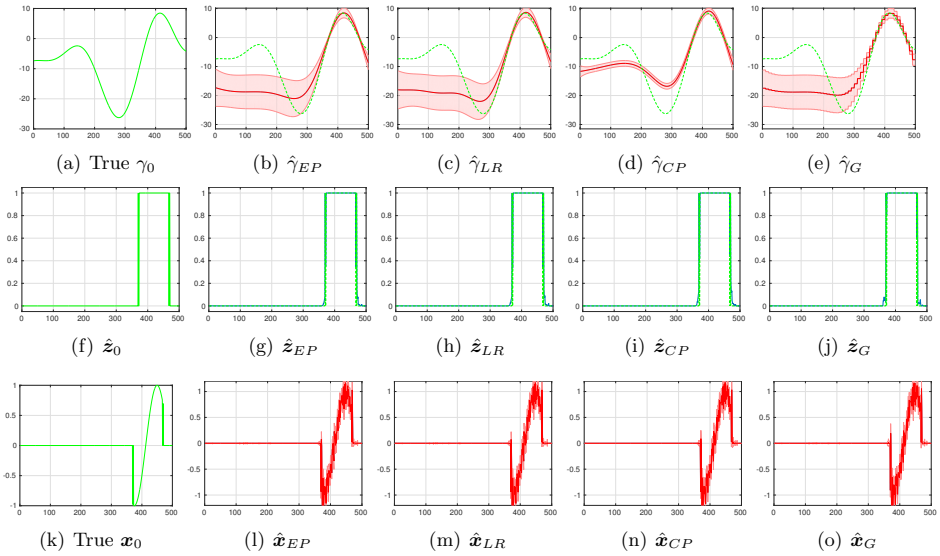
(c) Posterior mean of γ (d) Posterior mean of γ

(e) Posterior mean of z (f) Posterior mean of z

Figure 2: The effect of the spatial prior distribution. The left-most column shows the approximate marginal log likelihood, posterior mean of $\boldsymbol{\gamma}$ and posterior mean of z as a function of the prior length-scale of $\boldsymbol{\gamma}$. The right-most column shows similar plots as a function of the prior mean ν_0 of $\boldsymbol{\gamma}$.

7.2 Experiment 2: Comparison of approximation schemes

In this experiment, we investigate the properties of the proposed algorithm and the three approximation schemes: standard EP (EP), the low rank approximation (LR-EP), the common precision approximation (CP-EP) and the group approximation (G-EP). Using a similar setup as in Experiment 1, we generated a sample of $\boldsymbol{\gamma}_0$, $\boldsymbol{z}_0$ and $\boldsymbol{x}_0$ from the prior distribution specified in eq. (5)-(7) with $\rho_0 = 0, \tau_0 = 1$ and a squared exponential kernel with variance $\kappa_1^2 = 100$ and lengthscale $\kappa_2 = 75$. The generated sample is shown in the leftmost panels in Figure 3. We generated observations from a linear measurement model



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

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

(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 3: Comparison of approximation schemes. The panels in the first column shows a realization $\boldsymbol{x}, \boldsymbol{z}$, and $\boldsymbol{\gamma}$ from the prior distribution in eq. (5)-(7). The columns 2-5 show the posterior mean quantities for EP, the low rank approximation (LR-EP), the common precision approximation (CP-EP) and the group approximation (G-EP), respectively. The pink shaded areas depict $\pm$ standard deviation.

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}$, where $A_{ij} \sim \mathcal{N}(0,1)$ and the noise variance σ^2 is chosen such that the signal-to-noise is 20dB. Next, we computed the posterior distributions of $\boldsymbol{x}_0$, $\boldsymbol{z}_0$ and $\boldsymbol{\gamma}_0$ from the observed measurements $\boldsymbol{y}$ using standard EP and the three approximation schemes. For the low rank approximation we included 7 eigenvectors corresponding to 99% percent of the variance and for G-EP we used a group size of 10 variables. Columns 2-5 in Figure 3 show the posterior mean values for $\boldsymbol{x}, \boldsymbol{z}$, and $\boldsymbol{\gamma}$ for EP, LR-EP, CP-EP and G-EP, respectively. Consider the posterior mean and standard deviation for $\boldsymbol{\gamma}$ for standard EP (topmost row, second column). In the region where $\boldsymbol{\gamma}_0$ is positive the posterior mean accurately recovers $\boldsymbol{\gamma}_0$ with high precision, whereas both the accuracy and the precision is lower in regions where $\boldsymbol{\gamma}_0$ is negative. The reason for the additional uncertainty is that negative values of γ_i are in general associated with a small value of $|x_i|$, but $|x_i|$ can be small for two reason. Recall that each x_i can be considered as a product $x_i = z_i \cdot c_i$, where $z_i \in \{0, 1\}$ and $c_i \in \mathbb{R}$. If $z_i = 0$, then clearly $x_i = 0$, but we can still have that $x_i \approx 0$ even if $z_i = 1$ and $c_i \approx 0$ and thus the increased uncertainty.

We can now compare the posterior distribution of $\boldsymbol{\gamma}_i$ for standard EP with the three approximations. Based on visual inspection one cannot tell the difference between the standard EP and EP with the low rank approximation, but the results for CP-EP and G-EP

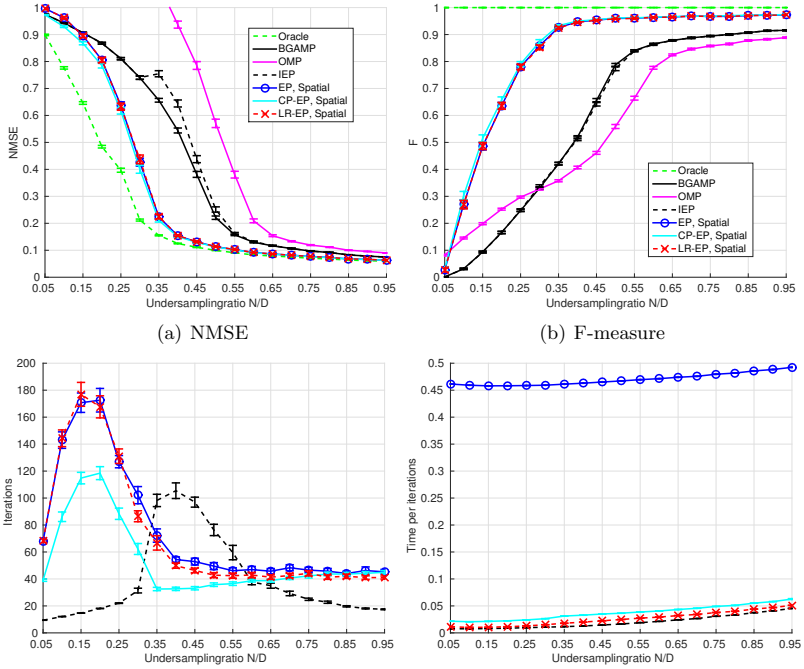
are quite different. For CP-EP it is seen that the posterior mean in the positive region is accurate, but the CP-EP approximation underestimates the uncertainty in general. The grouping effect for G-EP is clearly seen in the topmost panel in the last column, but despite the staircase pattern the posterior mean and variance are accurately recovered. The second and the third row in Figure 3 show the reconstructions of $\boldsymbol{x}$ and $\boldsymbol{z}$. We see that all of the four approaches accurately reconstruct the true quantities despite the approximation of the posterior distribution of $\boldsymbol{\gamma}$.

7.3 Experiment 3: Phase transitions for a single measurement vector

The purpose of this experiment is two-fold. The experiment serves to validate the inference algorithm, but it also serves to quantify the relationship between the recovery performance of the algorithm as a function of the undersampling ratio. It is well-known that the quality of the inferred solutions strongly depend on both the undersampling ratio $\delta = N/D$ and the number of non-zeros $K = ||\boldsymbol{x}||_0$ and that linear inverse problems exhibit a phase transition from almost perfect recovery to no recovery of solution $\boldsymbol{x}$ in the (δ, K)-space (Donoho and Tanner, 2010; Donoho et al., 2011). We hypothesize that the phase transition curves for signals with structured support can be improved, so that we can recover structured sparse signals using fewer measurements for a given number of non-zero coefficients K by taking advantage of the structure. We investigate this hypothesis by measuring the recovery performance of the EP algorithms as a function of the undersampling ratio N/D and compare with state-of-the-art probabilistic methods that ignore the structure of the support.

Using a squared exponential kernel for $\boldsymbol{\gamma}$ with variance $\kappa_1^2 = 50$ and lengthscale $\kappa_2 = 10$, we generated 100 realizations of $\boldsymbol{x}_0$ from the prior for $D = 500$. We fixed the expected sparsity to $K = \frac{1}{4}D = 125$ by choosing the prior mean of $\boldsymbol{\gamma}$ to $\nu = \phi^{-1}\left(\frac{1}{4}\right)(1 + \kappa_1^2)$. As the recovery performance is very sensitive to the number of non-zero coefficients, we conditioned each sample of $\boldsymbol{x}$ on $||\boldsymbol{x}||_0 = K$ by discarding samples where $||\boldsymbol{x}||_0 \neq K$ to reduce the variance of the resulting curves for NMSE and F-measure. For each of the samples, we generated measurements $\boldsymbol{y} \in \mathbb{R}^N$ through the linear observation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}$ for a range of values for N. The forward model $\boldsymbol{A}$ is a Gaussian i.i.d. ensemble, where the column have been scaled to unit ℓ_2-norm. The noise $e \sim \mathcal{N}\left(0, \sigma^2\right)$ is zero-mean Gaussian noise, where the noise variance σ^2 is chosen such that the signal-to-noise (SNR) ratio is fixed to 20dB. We choose values of N such that $\frac{N}{D} \in [0.05, 0.10, ..., 0.95]$.

We compare our methods with Bernoulli-Gaussian Approximate Message Passing (BG-AMP) (Vila and Schniter, 2013), Orthogonal Matching Pursuit (OMP) (Needell and Tropp, 2010) and an "oracle" estimator, which computes a ridge regression estimate based on knowledge of the true support. In this work, we use the BG-AMP method as baseline. It uses a (generalized) approximate message passing algorithm (Sundeep, 2010) for inference in a probabilistic model with i.i.d. spike-and-slab priors and a Gaussian likelihood. The AMP algorithm is closely related to EP algorithm (Meng et al., 2015), and the phase transition curve for BG-AMP is state of the art to the best of our knowledge. The OMP algorithm is a non-probabilistic greedy method, that iteratively select the column of $\boldsymbol{A}$ that correlate best with the current residuals until a pre-specified number of columns have been selected. The



(a) NMSE (b) F-measure

(c) Iterations (d) Time per iterations

Figure 4: Performance of the methods as a function of undersampling ratio N/D for $T = 1$. We compare full EP (EP, Spatial), EP with diagonal prior covariance (IEP), the common precision approximation (CP-EP) and the low rank approximation (LR-EP). The results are averaged over 100 realization.

regularization parameters for the ridge regression is fixed to $\lambda = 10^{-3}$ for all runs. Finally, for comparison we also apply the proposed EP algorithm with a diagonal prior covariance matrix, which correspond to the conventional independent spike-and-slab prior (IEP). We provide BG-AMP and OMP with prior knowledge of the true number of non-zero variables in $\boldsymbol{x}_0$ and the noise variance used to generate the observations. The results are shown in Figure 4.

The two black curves in Figure 4 show the results for BG-AMP (black, solid) and EP with diagonal prior covariance (black, dashed). Both of these methods are based on conventional independent spike-and-slab priors. It is seen that the methods with prior correlation, that is EP (blue), CP-EP (cyan), and LR-EP (red, dashed), are uniformly better than the methods with independent priors both in terms of NMSE and F-measure. In fact, these methods achieve as good performance as the support-aware oracle estimator around $N/D = 0.6$ in terms of NMSE. Furthermore, it is also seen that the two approximations CP-EP and LR-EP are indistinguishable from the full EP algorithm in terms of accuracy.

Panel (c) and (d) show the number iterations and the run time per iteration for the EP-based methods. Here it is seen that IEP has the lowest computational complexity per iteration, but the CP-EP and LR-EP are almost as fast.

7.4 Experiment 4: Compressed sensing of optical characters

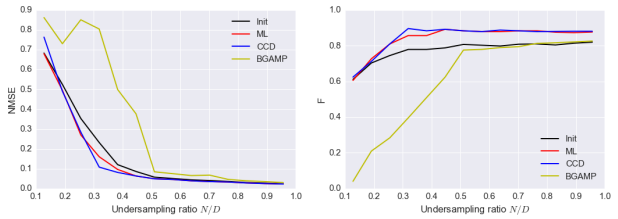
In this experiment, we apply the structured spike-and-slab model with Gaussian likelihood to an application of compressed sensing (Donoho, 2006) of numerical characters from the MNIST data set (LeCun et al., 1998; Hernández-Lobato et al., 2013). The images of the numerical digits are 28 pixels $\times$ 28 pixels and they are represented as vectors $\boldsymbol{x}_0 \in \mathbb{R}^{784}$. The objective is to reconstruct $\boldsymbol{x}_0$ from a small set of linear and noisy measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{\epsilon}$. The sensing matrix $\boldsymbol{A}$ is sampled independently from a standardized Gaussian distribution, that is $A_{ni} \sim \mathcal{N}(0, 1)$ and the noise variance is scaled such that the SNR is fixed 10dB.

We use a squared exponential kernel with a single lengthscale defined on the 2D image grid to encourage the neighbourhood structure expected in the images. We impose a Gaussian prior distribution on ν_0 with zero mean and variance κ_1^2, that is $\nu_0 \sim \mathcal{N}(0, \kappa_1^2)$ and integrate over ν_0 analytically to get the kernel function

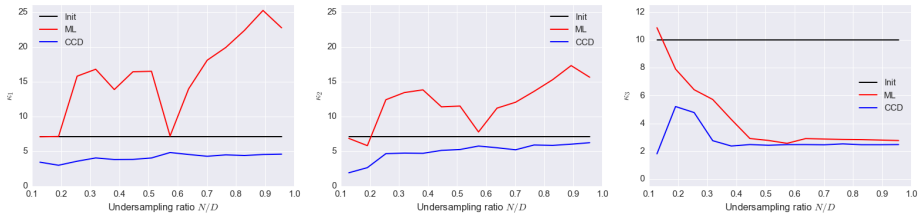
$$k(i, j) = \kappa_1^2 + \kappa_2^2 \exp\left(-\frac{\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2}{2\kappa_3^2}\right), \tag{73}$$

where $\boldsymbol{d}_i$ is the image grid coordinates of γ_i. We assume that the noise variance is known and we fix the prior mean and variance of the 'slab' component to a standardized Gaussian with $\rho_0 = 0$ and $\tau_0 = 1$. Thus, the hyperparameters to be learned are $\boldsymbol{\Omega} = \{\kappa_1, \kappa_2, \kappa_3\}$. For the CCD procedure, we have to choose prior distributions for the hyperparameters. For the lengthscale parameter, we can use the fact that the 'pen' is roughly a few pixels wide on average and choose a log-normal prior with mean 4 and standard deviation 2, that is $\kappa_3 \sim \mathcal{LN}\left(4, 2^2\right)$. The 10'th and 90'th percentiles for this distribution are approximately 2 and 7, respectively. For the remaining two hyperparameters, we will use the same prior distribution, that is $\kappa_1, \kappa_2 \sim \mathcal{LN}\left(4, 2^2\right)$, but for a different reason than for the lengthscale parameter κ_3. The mode of the distribution $\mathcal{LN}\left(4, 2^2\right)$ is approximate 2.9 and then the 10'th and 90'th percentiles of the distribution of $\phi(\gamma)$ for $\gamma \sim \mathcal{N}(0, 2.9^2)$ are approximately 0.0001 and 0.9999, respectively. Furthermore, the choice of lognormal priors generally works well for the CCD scheme, which can yield poor performance if the distributions have too heavy tails.

We use the low-rank approximation for all computations in this experiment. Figure 5(a)–(b) show the NMSE reconstruction error and F measure as a function of the undersampling ratio $\frac{N}{D}$. In this experiment, we also compare with the BGAMP method, which is informed about the noise level. We also use a standardized Gaussian as slab distribution for BGAMP. The black curves in panels (a) and (b) show the performance for the model when the hyperparameter are fixed to the initial values. It is seen that for small undersampling rates $N/D < 0.5$, we obtain slightly better results in terms of NMSE when adapting the hyperparameters, but we get a uniform improvement in terms of the F measure. Figure 5(c)–(e) show the estimated values for the hyperparameters as a function of the undersampling ratio. It is seen



(a) NMSE (b) F-measure



(c) Bias κ_1 (d) Magnitude κ_2 (e) Lengthscale κ_3

Figure 5: Performance of compressed sensing of numerical digits as a function of undersampling ratio. Panels (a) and (b) show the NMSE and F-measure, while panels (c)–(e) show the estimated values of the hyperparameters as a function of the undersampling ratio. For the CCD method, the panels (c)–(e) show the CCD-weighted average of the hyperparameters.

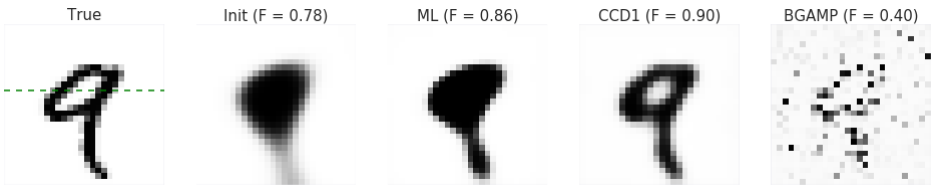
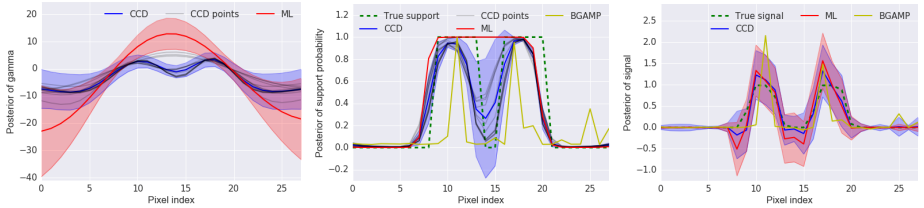


Figure 6: Posterior mean of $\boldsymbol{z}$ for compressed sensing of numerical digits, where $N/D \approx 0.3$. The panels in the top row show the posterior mean of the support, while the panels in the bottom row show the posterior mean of signal. Figure 7 shows the posterior distributions of the row indicated the dashed line.

that the ML solution tends to overestimate the lengthscale for small sample sizes. In this case, only weak information are propagated from the observations to the prior of $\boldsymbol{\gamma}$ and thus the model becomes over-regularized. It is also seen that the bias and magnitude parameters are correlated as expected from the relationship in eq. 8, (see Appendix D for more details).



(a) Posterior $\boldsymbol{\gamma}$ (b) Posterior support (c) Posterior signal

Figure 7: Comparison of the posterior distribution of $\boldsymbol{\gamma}$, $\boldsymbol{z}$ and $\boldsymbol{x}$ for the row shown by the green dashed line in Figure 6 for $N/D \approx 0.3$.

Method	**Training error**	**Test error**	**Train LPPD**	**Test LPPD**
NBSBC	9.7 ($\pm$0.3)	19.5 ($\pm$0.1)	-42.7 ($\pm$0.8)	-698.6 ($\pm$3.3)
LR-EP (ML)	13.3 ($\pm$0.3)	19.4 ($\pm$0.1)	-50.6 ($\pm$0.8)	-673.8 ($\pm$2.2)
LR-EP (CCD)	13.4 ($\pm$0.3)	**19.2** ($\pm$0.1)	-50.7 ($\pm$0.8)	$\mathbf{-665.5}$ ($\pm$1.5)

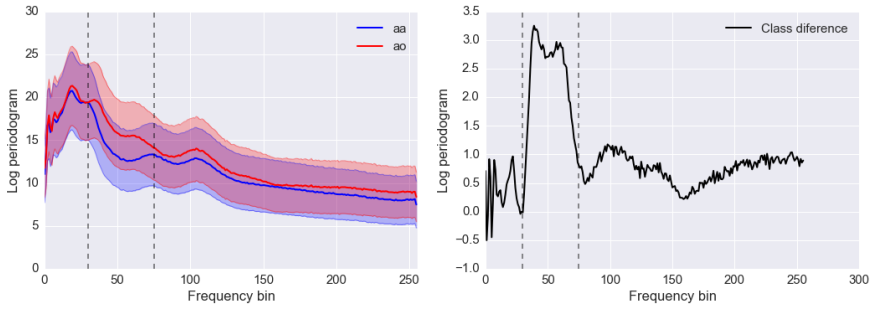
Table 2: Results for phoneme classification experiment

Figure 6 shows the posterior mean of the support for each method for $N/D \approx 0.3$, where it is seen that we obtain a qualitative and quantitative improvement of the support estimate by taking a priori knowledge into account and integrating over the uncertainty. Figure 7 shows the posterior distribution for $\boldsymbol{\gamma}$, $\boldsymbol{z}$ and $\boldsymbol{x}$ for the line indicated by the green dashed line in Figure 6(a). Recall, that the posterior distributions obtained using CCD are finite mixture models. The thin gray lines in left and center columns show the posterior mean of the individual mixture components, while the solid colored lines and the shaded areas show the mean and variance of the mixture distributions, respectively. From the center panel, it is seen all methods fail to capture the true support perfectly, but the mean of the support of the CCD solution are significantly improved compared to the ML solution and more interestingly, the CCD solution also has high variance in the region, where it is wrong. These uncertainties are not properly reflected in the NMSE and F metrics, but the log posterior density of the true support of the ML solution is -181.654, while the same quantity for the CCD method and BG-AMP evaluate to -74.181 and -339.065, respectively.

7.5 Experiment 5: Phoneme recognition

In this experiment, we consider the task of phoneme recognition (Hastie et al., 2001; Hernandez-Lobato et al., 2011). In particular, we consider the problem of discriminating between the spoken vowels "aa" and "ao" using their log-periodograms as features. The data set consists of 695 and 1022 utterances of the vowels "aa" and "ao", respectively, along with their corresponding labels. The response variable in this experiment is binary and therefore we use the probit model rather than the Gaussian observation model.

Each log-periodogram has been sampled at 256 uniformly spaced frequencies. The left-most



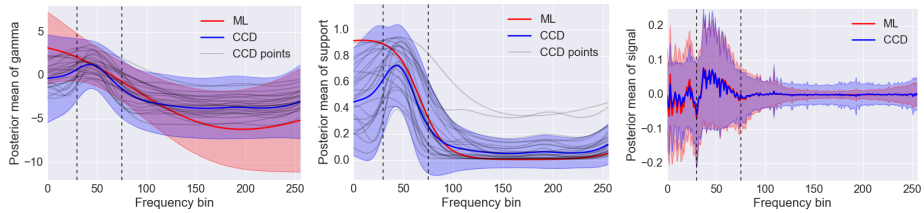
(a) The two classes (b) Difference in mean

Figure 8: a) The frequency-wise mean and standard deviation of the log-periodogram of the two spoken vowels "aa" and "ao". b) The difference of the two mean signals.

panel in Figure 2 shows the frequency-wise mean and standard deviation of the two classes and the right-most panel shows the difference of the two mean signals. We choose a squared exponential kernel for $\boldsymbol{\gamma}$ since it is assumed that frequency bands rather than single frequencies are relevant for discriminating between the two classes. The total number of observations is 1717 and we use $N = 150$ examples for training and the remaining 1567 examples for testing. We repeat the experiment 100 times using different partitions into training and test sets. The training and test sets are generated such that the prior odds of the two classes are the same for both training and test. The number of input features is $D = 257$ (256 frequencies + bias).

We use the low rank approximation for this experiment, and we choose the number of eigenvectors such that 99% of the total variance in $\boldsymbol{\Sigma}_0$ is explained. We use the maximum likelihood method and the CCD marginalization method for the hyperparameter inference. We choose the prior mean of $\boldsymbol{x}$ as $\rho_0 = 0$ to reflect our ignorance on the sign of the active weights and we impose a half Student's t distribution on the prior standard deviation of the $\boldsymbol{x}$, that is $\sqrt{\tau_0} \sim t^+(\mathrm{df} = 4)$, which is considered to be weakly informative.

As in the compressed sensing experiment, we impose a zero-mean normal distribution on the prior mean of $\boldsymbol{\gamma}$ and integrate it out analytically to obtain a kernel of the form given in eq. (73). Compared to the compressed sensing example, our a priori knowledge of the structure of the support are more diffuse, but we expect that the lengthscale is significantly smaller than the number of frequency bins. Therefore, we choose a log-normal prior with mean 40 and standard deviation 30, that is $\kappa_3 \sim \mathcal{LN}\left(40, 30^2\right)$. The 5'th and 95'th percentiles for this distribution is approximately 10 and 100, respectively. For the remaining two hyperparameters, we use the same two prior distributions as in the previous experiment, that is $\kappa_1, \kappa_2 \sim \mathcal{LN}\left(4, 2^2\right)$. To predict the label of a new observation, we compute the predictive distribution by integrating the probit likelihood with respect to the approximate posterior distribution of the weights.



(a) Posterior $\boldsymbol{\gamma}$ (b) Posterior support (c) Posterior signal

Figure 9: Comparison of the posterior distribution of $\boldsymbol{\gamma}$, $\boldsymbol{z}$ and $\boldsymbol{x}$ for ML and CCD hyper-parameter inference. The posterior distribution for the CCD approximation is a mixture model and the thin black solid lines show the posterior mean of each individual component. These are omitted for the posterior of the signal to improve visual clarity.

We compare our method against the network-based sparse Bayesian classification (NBSBC) method, which also uses EP to approximate the posterior distribution of linear model with coupled spike-and-slab priors. Instead of using a Gaussian process to encode the structure of the support, the NBSBC model encodes the structure in a network using a Markov random field prior. This method has been shown to outperform competing method on this specific problems (Hernandez-Lobato et al., 2011).

Table 2 summarizes the performance of the methods based on the average number of misclassifications and average log posterior predictive density (LPPD). It is seen that the LR-EP (ML) method achieves similar performance in terms test error as the NBSBC method, but it performs marginally better in terms of test LPPD. On the other hand, it is seen that the LR-EP (CCD) method outperforms both other methods. The panels in Figure 9 shows the posterior distributions of $\boldsymbol{\gamma}$, $\boldsymbol{z}$ and $\boldsymbol{x}$. The posterior distribution for the CCD approximation is a finite mixture model and each of the thin black lines shows the posterior mean for each individual mixture component. However, these are omitted for the posterior of $\boldsymbol{x}$ to improve the visual clarity of the figure. Based on Figure 2(b), we expect the weights for the frequencies between bin 35 and bin 70 to most discriminative of the two classes and it is seen that both the ML method and the CCD method have high posterior probabilities for the support in the region.

7.6 Experiment 6: Spatio-temporal example

In the previous experiments the focus was on problems with only one measurement vector, whereas in this and the following experiments we consider problems with multiple measurement vectors. Specifically, in this experiment we qualitatively study the properties of the proposed algorithm in the spatio-temporal setting using simulated data. We have synthesized a signal, where the support set satisfies the following three properties: 1) non-stationarity, 2) spatiotemporal correlation, and 3) the number of active coefficients change over time. The support of the signal is shown in panel (a) in figure 10. Based on the support set, we sample

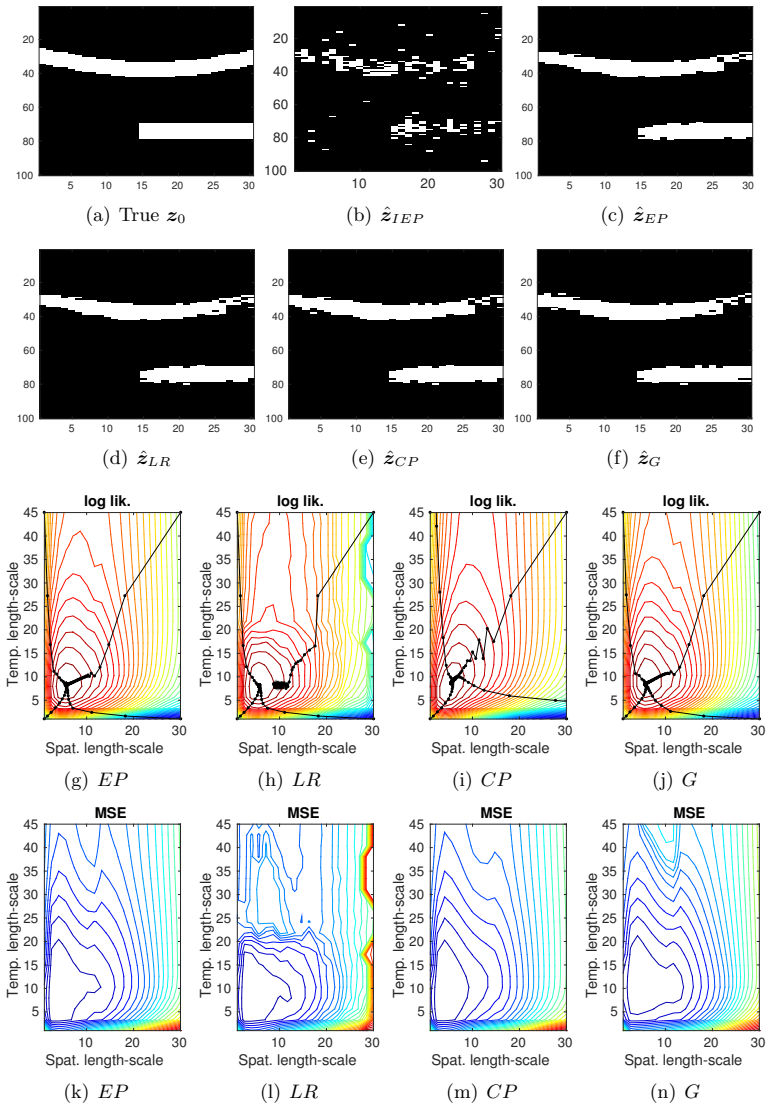
the active coefficients from a zero-mean isotropic Gaussian distribution. We then observe the desired signal $\boldsymbol{X}$ through linear measurements $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}$, where both the forward model $\boldsymbol{A}$ and the noise $\boldsymbol{E}$ is sampled from a zero-mean isotropic Gaussian distribution. The noise variance is scaled such that the SNR is 5dB. We apply our proposed method to estimate $\boldsymbol{X}$ given the forward model $\boldsymbol{A}$ and the observations $\boldsymbol{Y}$.

Panel (b) in figure 10 shows the reconstructed support $\boldsymbol{Z}$ using the proposed EP algorithm with a diagonal prior covariance matrix on $\boldsymbol{\Gamma}$, which implies no prior correlation in the support. The panels (c)–(f) shows the reconstructed support for full EP, low rank EP, common precision EP and group EP, which all assumes that the prior covariance matrix for $\boldsymbol{\Gamma}$ is a Kronecker product of two squared exponential components. For the group approximation the group size is chosen to 5 and 10 in the spatial and temporal dimension, respectively, and for the low-rank approximation the rank is chosen such that the minimum number of eigenvectors explain 99% of the variance in the prior. All hyperparameters are chosen by maximizing the approximate marginal likelihood. By inspecting the panels (a)–(f) it is seen that the reconstructed support is qualitatively improved by modeling the additional structure. Furthermore, the reconstructions using the approximation schemes do not differ significantly from the result using full EP.

Panels (g)–(j) shows the marginal likelihood approximation as a function of the spatial and temporal length scale of the prior covariance matrix for the proposed methods, while the panels (k)–(g) show the corresponding NMSE between the reconstructed coefficients $\hat{\boldsymbol{X}}$ and the true coefficient $\boldsymbol{X}$. The black curves superimposed on the marginal likelihood plots show the trajectories of the optimization path for the length-scales of the prior distribution starting from four different initial values. It is seen that the marginal likelihood approximation is unimodal and correlates strongly with the NMSE surface, which suggests that it is reasonable to tune the length-scales of the prior covariance using the marginal likelihood approximation. However, we emphasize that this is not always the case and for some problems this indeed leads to suboptimal results.

7.7 Experiment 7: Phase transitions for multiple measurement vectors

The multiple measurement vector problem also exhibits a phase transition analogously to the single measurement vector problem described in Experiment 3 (Cotter et al., 2005; Ziniel and Schniter, 2013a; Andersen et al., 2015). In this experiment, we investigate how the location of the phase transition of the EP algorithms improves when the sparsity pattern of the underlying signal is smooth both in space and time and multiple measurement vectors are available. Using a similar setup as in Experiment 3, we generate 100 realizations of $\boldsymbol{X}$ from the prior specified in eq. (10)–(12) such that the total number of active components is fixed to $K = \frac{1}{4}DT = 2500$. The covariance structure is of the form $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\text{temporal}} \otimes \boldsymbol{\Sigma}_{\text{spatial}}$, where both the temporal and spatial components are chosen to be squared exponential kernels. Figure 11 shows an example of a sample realization of $\boldsymbol{\Gamma}, \boldsymbol{Z}$ and $\boldsymbol{X}$ from the prior distribution. For each of the realizations of $\boldsymbol{X}$, we generate a set of linear observations $Y = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{E}$, where the forward model $\boldsymbol{A}$ is Gaussian i.i.d. and $E_{nt} \sim \mathcal{N}\left(0, \sigma^2\right)$ is zero-mean Gaussian scaled such that the SNR is fixed to 20dB. For reference we compare our



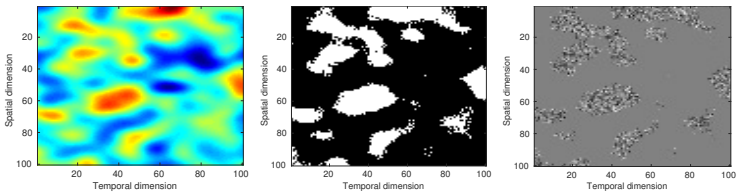
(a) True $\boldsymbol{z}_0$ (b) $\hat{\boldsymbol{z}}_{IEP}$ (c) $\hat{\boldsymbol{z}}_{EP}$

(d) $\hat{\boldsymbol{z}}_{LR}$ (e) $\hat{\boldsymbol{z}}_{CP}$ (f) $\hat{\boldsymbol{z}}_{G}$

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

(k) EP (l) LR (m) CP (n) G

Figure 10: Results for a simulated spatio-temporal example with $D = 100$, $T = 30$, $N = 33$ and SNR $= 5dB$. Panels (a)–(f) compare the true sparsity pattern and the reconstructed sparsity patterns and panels (g)–(n) show the approximate marginal likelihood and the MSE error metric as a function of the spatial and temporal length-scale. For the low rank approximation (LR-EP), the number of eigenvectors is chosen to explain 99% of the variance and the group size for group approximation (G-EP) is chosen to 5 and 10 in the spatial and temporal dimensions, respectively.



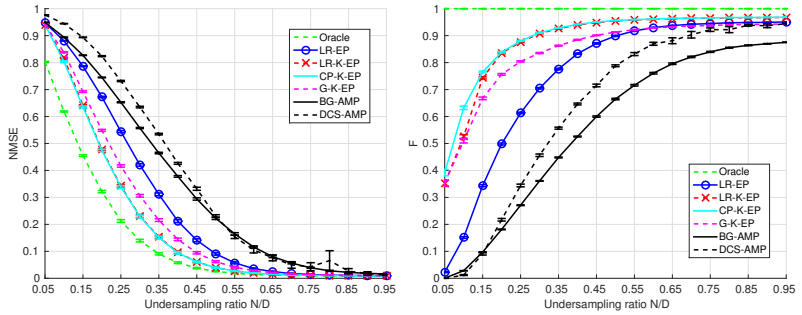
(a) Sample of $\mathbf{\Gamma}$ (b) Sample of $\boldsymbol{Z}$ (c) Sample of $\boldsymbol{X}$

Figure 11: Example of a realization of the synthetic signals in Experiment 6.

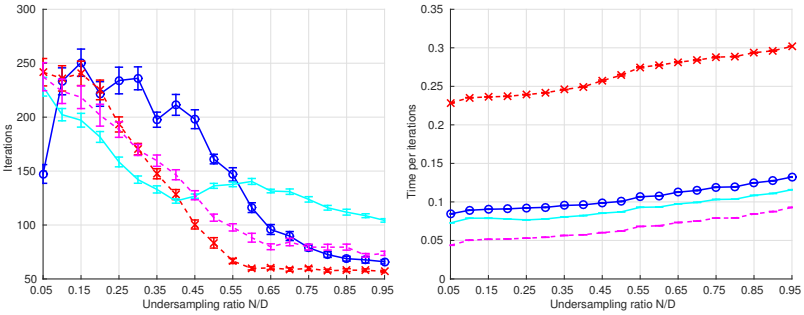
methods against BG-AMP (Vila and Schniter, 2013) and DCS-AMP (Ziniel and Schniter, 2013a). The DCS-AMP method is a temporal extension to the BG-AMP method (see Experiment 3 for a brief description), and it uses approximate message passing inference based on spatially i.i.d. spike-and-slab priors, but assumes that the binary support variables evolve in time according to a first order Markov process. Both BG-AMP and DCS-AMP methods are informed about the true number of active coefficients and the true noise level. The results are shown in figure 12.

The method LR-EP (blue) assumes that the sparsity pattern is spatially correlated, but independent in time. The method LR-K-EP (red, dashed) applies the low rank approximation to EP and assumes that the sparsity pattern is spatio-temporally correlated and that the prior covariance for $\mathbf{\Gamma}$ is described by a Kronecker product (hence the prefix "-K"). Similarly, the methods CP-K-EP (cyan) and G-K-EP (magenta) have the same assumptions about the sparsity pattern, but use the common precision approximation and the group approximation, respectively. For G-K-EP we use groups of 5 in both the spatial dimension and temporal dimension. In this experiment, we do not run full EP with the spatiotemporal prior because it would be prohibitively slow.

On panel (a) in figure 12 it is seen that as the number of measurements increase, all methods eventually reach the NMSE level of the support-aware oracle estimator, but the general picture is that the more structure a method takes into account (i.i.d. sparsity vs. spatial sparsity vs. spatio-temporal sparsity), the better it performs in terms of NMSE. In particular, at $N/D \approx 0.3$ BG-AMP achieves NMSE ≈ 0.63 and LR-EP achieves NMSE ≈ 0.44 while LR-K-EP and CP-K-EP achieve NMSE ≈ 0.24. Panel (b) shows a similar picture for F-measure. Furthermore, it is seen that the performance of LR-K-EP and CP-K-EP are similar and slightly better than the performance of G-K-EP both in terms of NMSE and F-measure. However, the G-K-EP approximation has the lowest computational complexity per iteration as seen in panel (d). In terms of run time the EP-methods are slower compared to the AMP-based methods, which have linear time complexity in all dimensions. However, the EP methods are not limited to Gaussian i.i.d. ensembles as the AMP-based methods are.



(a) NMSE (b) F-measure score



(c) Iterations (d) Time per iterations

Figure 12: Performance of the methods as a function of undersampling ratio N/D for $T = 100$. We compare low-rank EP with spatial structure only (LR-EP, Spatial), low rank EP spatio-temporal kronecker structure (LR-L-EP), the common precision EP with spatio-temporal kronecker structure (CP-K-EP), group EP with spatio-temporal kronecker structure (G-K-EP), the low rank approximation (LR-EP) with BGAMP and DCSAMP. The results are averaged over 100 realization.

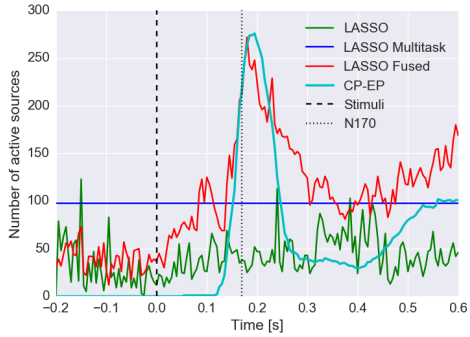


Figure 13: The number of active dipoles sources as a function of the time for the common precision approximation (CP-EP) and the LASSO, the Multi-task LASSO and the fused LASSO for a face perception experiment. The stimuli was presented at time $t = 0$.

7.8 Experiment 8: EEG Source Localization

In the final experiment, we apply the proposed method to an EEG source localization problem (Baillet et al., 2001), where the objective is to infer the locations of the active sources on the cortical surface of the human brain based on electroencephalogram (EEG) measurements. The brain is modelled using a discrete set of current dipoles distributed on the cortical surface and Maxwell's equations then describe how the magnitudes of the dipole sources relate to the EEG signals measured at the scalp. We apply the proposed method to an EEG data set, where the subjects are presented with pictures of faces and scrambled faces. The data set is publicly available and the experimental paradigm is described in (Henson et al., 2003). The data set has $N = 128$ electrodes and contains a total of 304 epochs with roughly 150 epochs of each of the two conditions: face or scrambled face. Each epoch has a duration of roughly $800ms$ corresponding to $T = 161$ samples in time and the stimuli is presented at $t = 0s$. We generated a forward model[1] with 5124 dipole sources, that is $\boldsymbol{A} \in \mathbb{R}^{128\times5124}$. To encourage spatio-temporal coherence of the sources, we choose the covariance matrix for $\boldsymbol{\Gamma}$ to be of the form $\boldsymbol{\Sigma}_0 = \kappa^2 \cdot \boldsymbol{\Sigma}_{\text{temporal}} \otimes \boldsymbol{\Sigma}_{\text{spatial}}$, where both the temporal component and spatial component are squared exponential kernels with individual length-scales. For simplicity, we use the Euclidean distance to compute the pairwise distances among the dipole sources as opposed to the more advances approach, where the distances are computed within the manifold defined by the cortex.

The resulting inverse problem has $N = 128$ measurements, $T = 161$ measurement vectors and $D = 5124$ unknowns per time point and a total of $DT = 5124 \cdot 151 = 824964$ unknowns. The forward model has a condition number of $\text{cond}(\boldsymbol{A}) = 3.1099 \cdot 10^{15}$. Thus, the problem instance is both heavily ill-posed and ill-conditioned. Because of the dimensions

1. We used the SPM8 software (Ashburner et al., 2010).

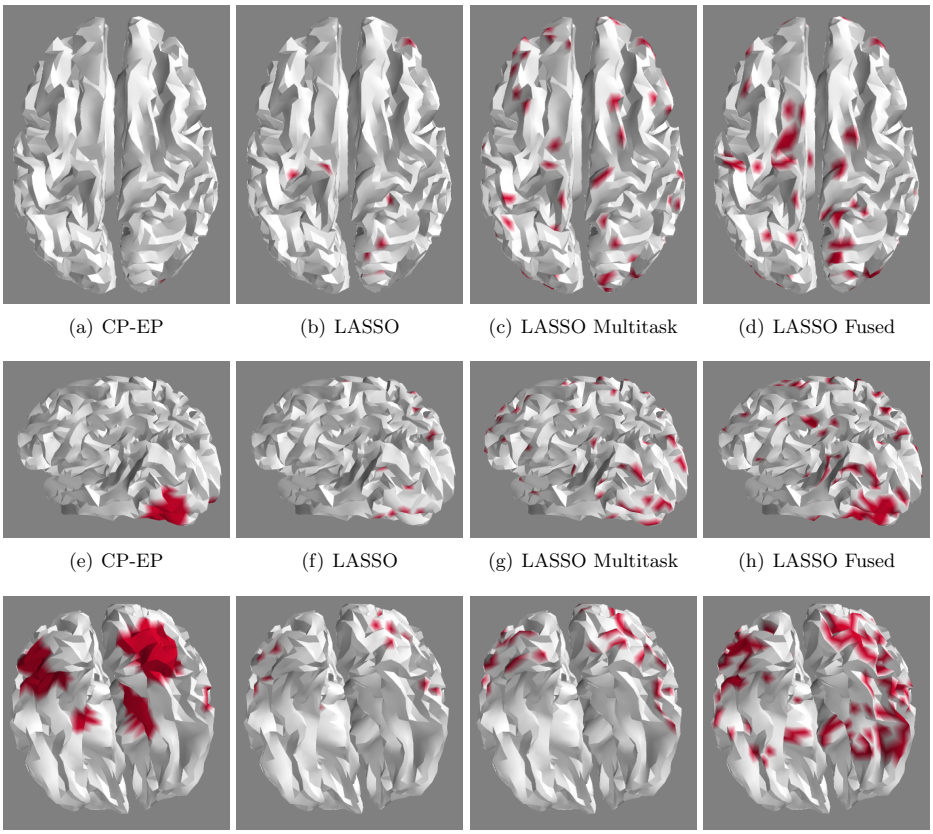
of this problem we use the common precision approximation for this data set. In fact, a low rank approximation of the prior covariance matrix $\mathbf{\Sigma}_0$ will require 3961 eigenvectors to explain 90% of the variance and the matrix low rank eigenvector matrix $\boldsymbol{U} \in \mathbb{R}^{824964 \times 3691}$ would then require more than $20GB$ of memory to store in 64 bit double precision.

Tuning the hyperparameters using the approximate marginal likelihood estimate leads poor solutions for this data set. In particular, the length-scales were significantly overestimated, which is consistent with what we observed in the compressed sensing experiment for very small samples sizes (see Figure 5(e)). However, manually specifying the hyperparameters using prior knowledge (spatial lengthscale $10mm$, temporal lengthscale $50ms$ and magnitude $\kappa^2 = 10$, prior mean 0), yields a posterior approximation with several interesting aspects. Ideally, we would compare the findings with the same posterior quantities for the BG-AMP and DCS-AMP methods as discussed earlier, but the highly correlated columns of the forward model make the AMP-approximations break down as they assume that the entries of the forward model are sampled from an Gaussian i.i.d. distribution. Instead, we compare with the LASSO[2] (Tibshirani, 1994), the multi-task LASSO[2] (Obozinski et al., 2006) and the fused LASSO[3] (Tibshirani et al., 2005). The LASSO, the Multi-task LASSO and the fused LASSO all minimize a quadratic reconstruction error subject to an ℓ_1 constraint, but the Multi-task LASSO also assumes that the sparsity pattern is constant in time (joint sparsity) and the fused LASSO has an additional constraint on the temporal first-order difference of the solution $\sum_{i,t} |x_{i,t} - x_{i,t-1}|$. The regularizing parameters are chosen using cross-validation.

Figure 13 shows the number of active dipole sources as a function of time for each method. The reconstructed support for both CP-EP and the fused LASSO are well-localized in time, whereas the distribution of active sources for LASSO are very diffuse in time. For the CP-EP method, it is seen that the number of active sources is zero until roughly time $t \approx 150ms$, where the number of active sources increase and peaks at $t \approx 180ms$, which is roughly consistent with the known time delay of 170ms for the face perception, that is the so-called N170 ERP component (Itier and Taylor, 2004). Figure 14 shows a visualization of the estimated sets of active sources for time $t = 180ms$ from a top view, a side view and a bottom view, respectively. Interestingly, CP-EP detects four spatially coherent areas: left and right occipital and fusiform face areas that are associated with the face perception (Henson et al., 2009). The LASSO, the Multi-task LASSO and the fused LASSO also detect several active dipoles in the left and right occipital areas, but they also detect active sources distributed over the entire cortex as seen in the top row.

Thus, from this experiment we conclude that this problem is too ill-posed for learning the hyperparameters of the model, but we can still extract meaningful information from the data using the model if we have access to additional a priori information. However, learning regularization parameters for neuroimaging problems are in general a difficult (Varoquaux et al., 2017).

2. We used the implementation in scikit-learn toolbox (Pedregosa et al., 2011)
3. We used the implementation in SPAMS toolbox (Jenatton et al., 2010; Mairal et al., 2010).



(a) CP-EP (b) LASSO (c) LASSO Multitask (d) LASSO Fused

(e) CP-EP (f) LASSO (g) LASSO Multitask (h) LASSO Fused

(i) CP-EP (j) LASSO (k) LASSO Multitask (l) LASSO Fused

Figure 14: Estimate support sets for each method at time $t \approx 180ms$ for the face perception experiment. The top, middle and bottom rows show the brain from the top, side and bottom respectively.

8. Summary and Outlook

In this work, we have addressed the problem of solving multiple underdetermined linear inverse problems subject to a sparsity constraint. We have proposed a new generalization of the spike-and-slab prior distribution to encode a priori correlation of the support of the solution in both space and time by imposing a transformed Gaussian process on the spike-and-slab probabilities. An expectation propagation (EP) algorithm for posterior inference under the proposed model has been derived. Computations involved in EP updates scale like $\mathcal{O}\left(D^3T^3\right)$ where D is the number of features and T is number of inverse problems, hence for large scale problems, the standard EP algorithm can be prohibitively slow. We therefore introduced three different approximation schemes for the covariance structure to reduce the computational complexity. First, assuming that the prior has a Kronecker decomposition brings complexity $\mathcal{O}\left(D^3 + T^3\right)$, based on this decomposition, a further K-rank approximation brings a reduction of complexity to $\mathcal{O}\left(K^2DT\right)$, we also proposed a common precision approximation of complexity $\mathcal{O}\left(D^2T + T^2D\right)$, and finally a scheme based on spatio-temporal grouping of variables effectively reducing D and T by the grouping factor. We also discussed several ways to handle unknown hyperparameters, including maximum likelihood estimates, maximum a posterior estimates and efficient numerical integration using central composite design (CCD) approach.

We investigated the role of the spatio-temporal prior and the approximation schemes in a series of experiments. First we studied a simple 1D problem with spatial, translational invariant smoothness of the support (single measurement case, $T = 1$). For a signal with two small connected components in the support, we illustrate the solutions for variable smoothness of the prior. For a wide range of prior parameters the correct form of the support is recovered, while the two support regions were found to merge in a single region as the smoothness length scale approaches the distance between the two regions. In the second experiment we investigated the role of the three covariance function approximations, also in a single measurement setup ($T = 1$). We found that all approaches accurately reconstruct the true simulated support and inverse problem solutions despite the approximation of the Gaussian process posterior. It is well-known that the quality of the inferred solutions strongly depends on both the undersampling ratio and the sparsity level of the true solution. We investigated how the location of the phase transition is improved by invoking the smoothness prior. We found that the methods based on assumed prior correlation, were uniformly better than the methods with independent priors both in terms of the quality of normalized mean square error and in terms of their accuracy of support recovery (F-measure). The covariance approximation schemes are almost as fast as the scheme without smoothness, while yielding greatly improved performance.

In the experiments 4 and 5, we investigated two applications: compressed sensing of numerical characters and phoneme recognition, respectively. In the former, we demonstrated how the quality of the reconstructed digits was improved using the structured prior. We also found that for the severely undersampled problems, maximum likelihood learning tends to overestimate the lengthscale of the kernel, which in turn lead to poor estimates of the support of and the weight. However, we also demonstrated how this could be alleviated by

imposing a proper prior distribution to the lengthscale parameter and integrating over the uncertainty using CCD. In the latter experiment, we demonstrated how the probit likelihood can be used to extend to model binary sparse classification problems and we found that our algorithm compare well with published benchmarks.

In a sixth experiment we studied the properties of the proposed algorithms in the spatio-temporal setting using simulated data. Signals were synthesized so that the support set showed non-stationarity, spatio-temporal correlation and so that the cardinality of the support set changed over time. We estimated prior hyperparameters by optimizing the approximate marginal likelihood and found they converged to optimal settings in all cases. We found that there was a good correspondence between the approximate marginal likelihood and the solution's quantitative performance measure (NMSE).

Also for the multiple measurement vector problem it is known that there is a phase transition-like dependence of the solution quality on undersampling ratio. In the sixth experiment we investigated how the location of the phase transition of the EP algorithms improved when the sparsity pattern of the underlying signal is smooth in both space and time for the multiple measurements case. We compare our various approximate solvers with the state-of-the-art tools based on approximate message passing: BG-AMP, DCS-AMP, both of which were informed about the true number of active coefficients and the true noise level. The full EP was too demanding to run for this problem. Significant improvement were found for the methods that exploited sparsity structure. Comparing performance with AMP methods, the EP methods performed best both in terms of identifying the support (F measure) and in terms of NMSE. Run times for the EP-methods were longer compared to the AMP-based methods, which have linear complexity in all dimensions. We noted importantly that the EP methods also can be used for more general forward model ensembles (A), while the AMP-based methods assume a Gaussian i.i.d. ensemble.

In the final experiment, we applied the proposed methods to the hard problem of EEG source localization; data for this experiment was derived from a publicly available brain imaging data set designed to detect brain areas involved in face perception (Henson et al., 2003). This was a larger scale application with $N = 128$ measurements and a total number of 824964 unknowns, hence, only the common precision approximation was feasible. Furthermore, the forward model was very ill-conditioned in contrast to the well-conditioned i.i.d. ensembles considered in the simulations. For this data set, the hyperparameters of the kernel, for example, spatial and temporal lengthscale, could not be estimated from the data and thus, additional prior knowledge was required to perform inference. In spite of these challenges highly interesting results were obtained: All four main foci of activation as earlier detected by fMRI, but not in these EEG data by other inference schemes, were here found to have well-defined and spatially extended support by the new approximate inference scheme. In contrast to fMRI EEG allowed us to monitor the dynamics in these areas in high temporal resolution.

This work has led to several interesting lines of research. First of all, from the the compressed sensing experiment as well as the source localization experiment, we concluded that the

lengthscale parameter of the kernel cannot be learned from the data if the problem is too ill-posed. Thus, in future work we will extend the model to handle EEG data for multiple subjects simultaneously in a hierarchical manner, which allows us to use much more data to estimate the hyperparameters. Future studies also include an analysis of the phase transitions of the approximate log marginal likelihood in the hyperparameter space of the spatiotemporal prior as discussed in Experiment 1. Furthermore, we also plan to apply the proposed algorithms to brain decoding problems, for example, in classification of fMRI task pattern data sets. Finally, we also plan to investigate the use of spatio-temporal sparsity priors for factor models like PCA and ICA.

Appendix A. Moments computations for $f_{2,t,j}$

In this section, we consider the update for the terms $\tilde{f}_{2,t,j}\left(x_{t,j}, z_{t,j}\right)$. First we compute the so-called cavity distribution $Q^{\backslash 2,t,j}\left(x_{t,j}, z_{t,j}\right)$ by removing the contribution of $f_{2,t,j}(x_{t,j}, t, j)$ from the marginals of the joint approximation $Q\left(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}\right)$

$$\begin{aligned} Q^{\backslash 2,t,j}(x_{t,j}, z_{t,j}) &= \frac{Q\left(x_{t,j}, z_{t,j}\right)}{\tilde{f}_{2,t,j}\left(x_{t,j}, z_{t,j}\right)} = \frac{\mathcal{N}\left(x_{t,j} \middle| \hat{m}_{t,j}, \hat{V}_{t,j}\right) \mathrm{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}_{t,j}\right)\right)}{\mathcal{N}\left(x_{t,j} \middle| \hat{m}_{2,t,j}, \hat{V}_{2,t,j}\right) \mathrm{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}_{2,t,j}\right)\right)} \\ &= K^{\backslash 2,t,j} \cdot \mathcal{N}\left(x_{t,j} \middle| \hat{m}^{\backslash 2,t,j}, \hat{V}^{\backslash 2,t,j}\right) \mathrm{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right), \end{aligned} \tag{74}$$

where

$$\hat{v}^{\backslash 2,t,j} = \left[\hat{V}_{jj}^{-1} - \hat{v}_{2,t,j}^{-1}\right]^{-1}, \tag{75}$$

$$\hat{m}^{\backslash 2,t,j} = \hat{v}^{\backslash 2,t,j}\left[\hat{V}_{t,jj}^{-1}\hat{m}_{t,j} - \hat{v}_{2,t,j}^{-1}\hat{m}_{2,t,j}\right], \tag{76}$$

$$\hat{\gamma}^{\backslash 2,t,j} = \hat{\gamma}_{3,t,j}. \tag{77}$$

Note that the cavity parameter for γ for $f_{2,t,j}$ is simply equal to $\hat{\gamma}_{3,t,j}$ (and vice versa) since $\hat{\gamma}_{2,t,j}$ and $\hat{\gamma}_{3,t,j}$ are the only two terms contributing to $\gamma_{t,j}$.

Next, we minimize the KL-divergence between $f_{2,t,j}Q^{\backslash 2,t,j}$ and q or equivalently matching the moments between the two distributions. Following the latter approach we first compute the (unnormalized) moment w.r.t. $z_{t,j}$

$$\begin{aligned} Z_1 &= \sum_{z_{t,j}} \int z_{t,j} f_{2,t,j}\left(x_{t,j}, z_{t,j}\right) Q^{\backslash 2,t,j}\left(x_{t,j}, z_{t,j}\right) \mathrm{d}x_{t,j} \\ &= \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right) \mathcal{N}\left(0 \middle| \hat{m}^{\backslash 2,i} - \rho_0, \hat{V}^{\backslash 2,t,j} + \tau_0\right). \end{aligned} \tag{78}$$

Next, the zeroth moment w.r.t $x_{t,i}$ or the normalization constant of $f_{2,t,j}Q^{\backslash 2,t,j}$

$$X_0 = \sum_{z_{t,j}} \int f_{2,t,j}\left(x_{t,j}, z_{t,j}\right) Q^{\backslash 2,t,j}\left(x_{t,j}, z_{t,j}\right) \mathrm{d}x_{t,j} \tag{79}$$

$$\begin{aligned}
&= \sum_{z_{t,j}} \int \left[\left(1 - z_{t,j}\right)\delta(x_{t,j}) + z_{t,j}\mathcal{N}\left(x_{t,j}|\rho_0, \tau_0\right)\right] \mathcal{N}\left(x_{t,j}|\hat{m}^{\backslash 2,t,j}, \hat{V}^{\backslash 2,t,j}\right) \mathrm{Ber}\left(z_{t,j}|\phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right) \mathrm{d}x_{t,j} \\
&= \left(1 - \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right) \mathcal{N}\left(0|\hat{m}^{\backslash 2,i}, \hat{V}^{\backslash 2,i}\right) + \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right) \mathcal{N}\left(0|\hat{m}^{\backslash 2,t,j} - \rho_0, \hat{V}^{\backslash 2,t,j} + \tau_0\right) \\
&= \left(1 - \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right) \mathcal{N}\left(0|\hat{m}^{\backslash 2,i}, \hat{V}^{\backslash 2,i}\right) + Z_1
\end{aligned} \tag{80}$$

We now compute the (unnormalized) first moment w.r.t. $x_{t,j}$

$$\begin{aligned}
X_1 &= \sum_{z_{t,j}} \int x_{t,j} f_{2,t,j}\left(x_{t,j}, z_{t,j}\right) Q^{\backslash 2,i}\left(x_{t,j}, z_{t,j}\right) \mathrm{d}x_{t,j} \\
&= \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right) \mathcal{N}\left(0|\hat{m}^{\backslash 2,t,j} - \rho_0, \hat{V}^{\backslash 2,t,j} + \tau_0\right) \frac{\frac{\hat{m}^{\backslash 2,t,j}}{\hat{V}^{\backslash 2,t,j}} + \frac{\rho_0}{\tau_0}}{\frac{1}{\tau_0} + \frac{1}{\hat{V}^{\backslash 2,t,j}}} \\
&= Z_1 \frac{\hat{m}^{\backslash 2,t,j}\tau_0 + \rho_0 \hat{V}^{\backslash 2,t,j}}{\tau_0 + \hat{V}^{\backslash 2,t,j}}
\end{aligned} \tag{81}$$

and the second (unnormalized) moment w.r.t. $x_{t,j}$

$$\begin{aligned}
X_2 &= \sum_{z_{t,j}} \int x_i^2 f_{2,i}\left(x_i, z_i\right) Q^{\backslash 2,i}\left(x_i, z_i\right) \mathrm{d}x_i \\
&= \phi\left(\hat{\gamma}^{\backslash 2,i}\right) \mathcal{N}\left(0|\hat{m}^{\backslash 2,i} - \rho_0, \hat{V}^{\backslash 2,i} + \tau_0\right) \left[\left(\frac{\frac{\hat{m}^{\backslash 2,i}}{\hat{V}^{\backslash 2,i}} + \frac{\rho_0}{\tau_0}}{\frac{1}{\tau_0} + \frac{1}{\hat{V}^{\backslash 2,i}}}\right)^2 + \frac{1}{\frac{1}{\tau_0} + \frac{1}{\hat{V}^{\backslash 2,i}}}\right] \\
&= Z_1 \left[\left(\frac{\hat{m}^{\backslash 2,t,j}\tau_0 + \rho_0 \hat{V}^{\backslash 2,t,j}}{\tau_0 + \hat{V}^{\backslash 2,t,j}}\right)^2 + \frac{\tau_0 \hat{V}^{\backslash 2,i}}{\hat{V}^{\backslash 2,i} + \tau_0}\right]
\end{aligned} \tag{82}$$

The central moments for Q^* in eq. (21) are given by

$$E\left[x_{t,j}\right] = \frac{X_1}{X_0}, \qquad V\left[x_{t,j}\right] = \frac{X_2}{X_0} - \frac{X_1^2}{X_0^2}, \qquad E\left[z_{t,j}\right] = \frac{Z}{X_0}. \tag{83}$$

Appendix B. Moment computations for $\tilde{f}_{3,t,j}$

The moments matching for $\tilde{f}_{3,t,j}$ is derived in a similar manner as for $\tilde{f}_{2,t,j}$ (see appendix A for details). First we compute the cavity distribution $Q^{\backslash 3,t,j}\left(z_{t,j}, \gamma_{t,j}\right)$ by removing the contribution of $f_{3_t,j}(z_{t,j}, \gamma_{t,j})$ from the marginals of the joint approximation Q

$$\begin{aligned}
Q^{\backslash 3,t,j}(z_{t,j}, \gamma_{t,j}) &= \frac{Q\left(z_{t,j}, \gamma_{t,j}\right)}{\tilde{f}_{3,t,j}\left(z_{t,j}, \gamma_{t,j}\right)} = \frac{\mathrm{Ber}\left(z_{t,j}|\phi\left(\hat{\gamma}_{t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j}, \hat{\mu}_{t,j}, \hat{\Sigma}_{t,jj}\right)}{\mathrm{Ber}\left(z_{t,j}|\phi\left(\hat{\gamma}_{3,t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j}, \hat{\mu}_{3,t,j}, \hat{\Sigma}_{3,t,j}\right)} \\
&= K^{\backslash 3,t,j} \cdot \mathrm{Ber}\left(z_{t,j}|\phi\left(\hat{\gamma}^{\backslash 3,t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j}|\hat{\mu}^{\backslash 3,t,j}, \hat{\Sigma}^{\backslash 3,t,j}\right),
\end{aligned} \tag{84}$$

where

$$\hat{\Sigma}^{\setminus 3,t,j} = \left(\hat{\Sigma}_{t,jj}^{-1} - \Sigma_{3,t,j}^{-1}\right)^{-1}, \tag{85}$$

$$\hat{\mu}^{\setminus 3,t,j} = \hat{\Sigma}^{\setminus 3,t,j}\left(\hat{\Sigma}_{t,jj}^{-1}\hat{\mu}_{t,j} - \hat{\Sigma}_{3,t,j}^{-1}\hat{\mu}_{3,t,j}\right), \tag{86}$$

$$\hat{\gamma}^{\setminus 3,t,j} = \hat{\gamma}_{2,t,j}. \tag{87}$$

Once again we minimize the KL-divergence between $f_{3,t,j}Q^{\setminus 3,t,j}$ and Q or equivalently matching the moments between the two distributions. We now compute the moments w.r.t. $\gamma_{j,t}$ and $z_{j,t}$ of the (unnormalized) tilted distribution

$$G_m = \sum_{z_{j,t}} \int \gamma_{j,t}^m \cdot f_{3,t,j}\left(z_{j,t}, \gamma_{j,t}\right) Q^{\setminus 3,t,j}\left(z_{j,t}, \gamma_{j,t}\right) \mathrm{d}\gamma_{j,t} \quad \text{for} \quad m = 0, 1, 2, \tag{88}$$

$$Z_1 = \sum_{z_{j,t}} \int z_{j,t} \cdot f_{3,t,j}\left(z_{j,t}, \gamma_{j,t}\right) Q^{\setminus 3,t,j}\left(z_{j,t}, \gamma_{j,t}\right) \mathrm{d}\gamma_{j,t} \tag{89}$$

We first compute the normalization constant of $f_{3,t,j}Q^{\setminus 3,t,j}$

$$\begin{aligned}
G_0 &= \sum_{z_{t,j}} \int f_{3,t,j}\left(z_{t,j}, \gamma_{t,j}\right) Q^{\setminus 3,t,j}\left(z_{t,j}, \gamma_i\right) \mathrm{d}\gamma_{t,j} \\
&= \sum_{z_{t,j}} \int \mathrm{Ber}\left(z_{t,j} \middle| \phi\left(\gamma_{t,j}\right)\right) \mathrm{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}^{\setminus 3,t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j} \middle| \hat{\mu}^{\setminus 3,t,j}, \hat{\Sigma}^{\setminus 3,t,j}\right) \mathrm{d}\gamma_{t,j} \\
&= \sum_{z_i} \int \left[\left(1 - z_i\right)\left(1 - \phi\left(\gamma_i\right)\right)\left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) + z_i \phi\left(\gamma_i\right) \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right] \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\
&= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \int \left(1 - \phi\left(\gamma_i\right)\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \int \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i
\end{aligned}$$

Integrals of the form $\int \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i$ can be solved analytically (Rasmussen and Williams, 2006),

$$\int \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i = \phi\left(c_{3,i}\right), \quad c_{3,i} \triangleq \frac{\hat{\mu}^{\setminus 3,i}}{\sqrt{1 + \hat{\Sigma}^{\setminus 3,i}}}. \tag{90}$$

Inserting this result back into the expression for G_0 yields

$$G_0 = \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right)\left(1 - \phi\left(c_{3,i}\right)\right) + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \phi\left(c_{3,i}\right). \tag{91}$$

We can now compute the moments of the unnormalized distribution

$$\begin{aligned}
Z_1 &= \sum_{z_i} \int z_i f_{3,i}\left(z_i, \gamma_i\right) Q^{\setminus 3,i}\left(z_i, \gamma_i\right) \mathrm{d}\gamma_i \\
&= \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \phi\left(c_{3,i}\right),
\end{aligned} \tag{92}$$

Then the first moment w.r.t. to $z_{i,t}$ is obtained as $E[z_{i,t}] = Z_1/G_0$.

For the moments w.r.t. γ_i, we get

$$\begin{aligned}
G_1 &= \sum_{z_i} \int \gamma_i f_{3,i}(z_i, \gamma_i) Q^{\setminus 3,i}(z_i, \gamma_i) \, \mathrm{d}\gamma_i \\
&= \sum_{z_i} \int \gamma_i \left[(1 - z_i)(1 - \phi(\gamma_i)) \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) + z_i \phi(\gamma_i) \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \right] \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\
&= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \int \gamma_i (1 - \phi(\gamma_i)) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\
&\quad + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \int \gamma_i \phi(\gamma_i) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\
&= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \left[\hat{\mu}^{\setminus 3,i} - \int \gamma_i \phi(\gamma_i) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \right] \\
&\quad + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \int \gamma_i \phi(\gamma_i) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i
\end{aligned} \tag{93}$$

Again we turn to (Rasmussen and Williams, 2006) for the analytical solution of the above integrals

$$\begin{aligned}
\int \gamma_i \phi(\gamma_i) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i &= \phi(c_{3,i}) \hat{\mu}^{\setminus 3,i} + \phi(c_{3,i}) \frac{\hat{\Sigma}^{\setminus 3,i} \mathcal{N}(c_{3,i} | 0, 1)}{\phi(c_{3,i}) \sqrt{1 + \hat{\Sigma}^{\setminus 3,i}}} \\
&= \phi(c_{3,i}) \hat{\mu}^{\setminus 3,i} + \phi(c_{3,i}) d_{3,i},
\end{aligned} \tag{94}$$

where we have defined

$$d_{3,i} \triangleq \frac{\hat{\Sigma}^{\setminus 3,i} \mathcal{N}(c_{3,i} | 0, 1)}{\phi(c_{3,i}) \sqrt{1 + \hat{\Sigma}^{\setminus 3,i}}}. \tag{95}$$

Plugging eq. (94) back into eq. (93) and simplifying yields

$$\begin{aligned}
G_1 &= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \left[(1 - \phi(c_{3,i})) \hat{\mu}^{\setminus 3,i} - \phi(c_{3,i}) d_{3,i} \right] + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \phi(c_{3,i}) \left[\hat{\mu}^{\setminus 3,i} + d_{3,i} \right] \\
&= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) (1 - \phi(c_{3,i})) \hat{\mu}^{\setminus 3,i} - \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \phi(c_{3,i}) d_{3,i} + Z_1 \left[\hat{\mu}^{\setminus 3,i} + d_{3,i} \right] \\
&= (G_0 - Z_1) \hat{\mu}^{\setminus 3,i} - \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \phi(c_{3,i}) d_{3,i} + Z_1 \left[\hat{\mu}^{\setminus 3,i} + d_{3,i} \right] \\
&= G_0 \hat{\mu}^{\setminus 3,i} + (2 Z_1 - \phi(c_{3,i})) d_{3,i}
\end{aligned} \tag{96}$$

Thus, the first moment w.r.t. $\gamma_{i,t}$ is given by $\mathbb{E}[\gamma_{i,t}] = G1/G0$.

Similarly, we compute the second moment w.r.t. γ_i

$$\begin{aligned} G_2 &= \sum_{z_i} \int \gamma_i^2 f_{3,i}(z_i, \gamma_i) Q^{\setminus 3,i}(z_i, \gamma_i) \, \mathrm{d}\gamma_i \\ &= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \int \gamma_i^2 (1 - \phi(\gamma_i)) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\ &\quad + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \int \gamma_i^2 \phi(\gamma_i) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \end{aligned} \tag{97}$$

The solution to the above integrals are given by (Rasmussen and Williams, 2006)

$$\begin{aligned} &\int \gamma_i^2 \phi(\gamma_i) \mathcal{N}\left(\gamma_i | \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\ &= \phi(c_{3,i}) \left[2\hat{\mu}^{\setminus 3,i}\left(\hat{\mu}^{\setminus 3,i} + d_{3,i}\right) + \left(\hat{\Sigma}^{\setminus 3,i} - \left(\hat{\mu}^{\setminus 3,i}\right)^2\right) - b_{3,i}\right] \end{aligned} \tag{98}$$

where

$$b_{3,i} \triangleq \frac{\left(\hat{\Sigma}^{\setminus 3,i}\right)^2 c_{3,i} \mathcal{N}(c_{3,i} | 0, 1)}{\phi(c_{3,i}) \left(1 + \hat{\Sigma}^{\setminus 3,i}\right)} \tag{99}$$

Furthermore, we define

$$w_{3,i} \triangleq 2\hat{\mu}^{\setminus 3,i}\left(\hat{\mu}^{\setminus 3,i} + d_{3,i}\right) + \left(\hat{\Sigma}^{\setminus 3,i} - \left(\hat{\mu}^{\setminus 3,i}\right)^2\right) - b_{3,i} \tag{100}$$

Substituting the above result back into eq. (97) and rearranging yields

$$\begin{aligned} G_2 &= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \left[\left(\hat{\mu}^{\setminus 3,i}\right)^2 + \hat{\Sigma}^{\setminus 3,i} - \phi(c_{3,i}) w_{3,i}\right] + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \phi(c_{3,i}) w_{3,i} \\ &= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \left[\left(\hat{\mu}^{\setminus 3,i}\right)^2 + \hat{\Sigma}^{\setminus 3,i} - \phi(c_{3,i}) w_{3,i}\right] + Z_1 w_{3,i} \end{aligned} \tag{101}$$

Thus, the second moment is given by $\mathbb{E}\left[\gamma_{i,t}^2\right] = G2/G0$. Finally, the central moments of Q^* then becomes

$$\mathbb{E}[\gamma_{j,t}] = \frac{G_1}{G_0}, \qquad \mathbb{V}[\gamma_{j,t}] = \frac{G_2}{G_0} - \frac{G_1^2}{G_0^2}, \qquad \mathbb{E}[z_{j,t}] = \frac{Z_1}{G_0}. \tag{102}$$

These moments completely determine the distribution $Q^{3,\text{new}}$ and thus, we compute the updates for $f_{3,i}$ as follows

$$\hat{\Sigma}_{3,i}^{\text{new}} = \left[\mathbb{V}[\gamma_i]^{-1} - \left(\hat{\Sigma}^{\setminus 3,i}\right)^{-1}\right]^{-1}, \tag{103}$$

$$\hat{\mu}_{3,i}^{\text{new}} = \hat{\Sigma}_{3,i}^{\text{new}} \left[\mathbb{V}[\gamma_i]^{-1} \mathbb{E}[\gamma_i] - \left(\hat{\Sigma}^{\setminus 3,i}\right)^{-1} \hat{\mu}^{\setminus 3,i}\right], \tag{104}$$

$$\hat{\gamma}_{3,i}^{\text{new}} = d\left(\phi(\mathbb{E}[z_i]), \hat{\gamma}^{\setminus 3,i}\right), \tag{105}$$

Appendix C. Moments computations for probit likelihood

The purpose of this section is to describe the details of the EP approximation of the structured spike-and-slab prior with a probit likelihood. Using the notation described in section 4, the probit likelihood term is given by

$$f_{1,t}\left(\boldsymbol{x}_t\right) = p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \prod_{n=1}^{N} \phi\left(y_{n,t}\boldsymbol{A}_{n,\cdot}\boldsymbol{x}_t\right). \tag{106}$$

First we compute the cavity distribution $Q^{\setminus 1,t,n}\left(\boldsymbol{x}\right)$ by removing the contribution of $\tilde{f}_{1_t,n}(\boldsymbol{x})$ from the marginals of the joint approximation Q

$$Q^{\setminus 1,t,n}(\boldsymbol{x}_t) = \frac{\mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}_t,\boldsymbol{V}_t\right)}{\tilde{f}_{1,t,n}(\boldsymbol{x}_t)} = K^{\setminus 1,t,n}\mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}^{\setminus 1,t,n},\boldsymbol{V}^{\setminus 1,t,n}\right), \tag{107}$$

where

$$\boldsymbol{V}^{\setminus 1,t,n} = \left(\hat{\boldsymbol{V}}_t^{-1} - \hat{\boldsymbol{V}}_{1,t,n}^{-1}\right)^{-1}, \tag{108}$$

$$\boldsymbol{m}^{\setminus 1,t,n} = \boldsymbol{V}^{\setminus 1,t,n}\left(\hat{\boldsymbol{V}}_t^{-1}\hat{\boldsymbol{m}}_t - \hat{\boldsymbol{V}}_{1,t}^{-1}\hat{\boldsymbol{m}}_{1,t,n}\right). \tag{109}$$

for diagonal matrices $\boldsymbol{V}_t$ and $\boldsymbol{V}^{\setminus 1,t,n}$. The tilted distribution then becomes

$$\hat{q}_{1,t,n}\left(\boldsymbol{x}_t\right) = \frac{1}{z_{1,t,n}}\phi\left(y_{n,t}\boldsymbol{A}_{n,\cdot}\boldsymbol{x}_t\right)\mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}^{\setminus 1,t,n},\boldsymbol{V}^{\setminus 1,t,n}\right),$$

First we compute the normalization constant, which is given by

$$z_{1,t,n} = \int \phi\left(y_{n,t}\boldsymbol{A}_{n,\cdot}\boldsymbol{x}_t\right)\mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}^{\setminus 1,t,n},\boldsymbol{V}^{\setminus 1,t,n}\right)\mathrm{d}\boldsymbol{x}_t \tag{110}$$

$$= \int \phi\left(u\right)\mathcal{N}\left(u|a_{1,t,n},b_{1,t,n}\right)\mathrm{d}u \tag{111}$$

$$= \phi\left(c_{1,t,n}\right), \tag{112}$$

where $a_{1,t,n} = y_{n,t}\boldsymbol{A}_{n,\cdot}\boldsymbol{m}^{\setminus 1,t,n}$, $b_{1,t,n} = \boldsymbol{A}_{n,\cdot}\boldsymbol{V}^{\setminus 1,t,n}\boldsymbol{A}_{n,\cdot}^T$ and $c_{1,t,n} = \frac{a_{1,t,n}}{\sqrt{1+b_{1,t,n}}}$. Since $y_{n,t} \in \{-1,1\}$, $y_{n,t}$ does not appear in the expression for $b_{1,t,n}$ due to square form. Define the row-vector $\tilde{\boldsymbol{A}}_n = y_{n,t}\boldsymbol{A}_{n,\cdot} \in \mathbb{R}^{1\times D}$, then the first moment w.r.t. $x_{t,j}$ is given by

$$\mathbb{E}\left[x_{t,j}\right] = \frac{1}{z_{1,t,n}}\int x_{t,j}\phi\left(y_{n,t}\boldsymbol{A}_{n,\cdot}\boldsymbol{x}_t\right)\mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}^{\setminus 1,t,n},\boldsymbol{V}^{\setminus 1,t,n}\right)\mathrm{d}\boldsymbol{x}_t \tag{113}$$

$$= \frac{1}{z_{1,t,n}}\int x_{t,j}\phi\left(\tilde{\boldsymbol{A}}_{n,\cdot}\boldsymbol{x}_t\right)\mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}^{\setminus 1,t,n},\boldsymbol{V}^{\setminus 1,t,n}\right)\mathrm{d}\boldsymbol{x}_t \tag{114}$$

$$= \frac{1}{z_{1,t,n}}\int x_{t,j}\int \phi\left(\tilde{\boldsymbol{A}}_{n,-j}\boldsymbol{x}_{-j} + \tilde{a}_{n,j}x_{t,j}\right)\mathcal{N}\left(\boldsymbol{x}_{t,-j}|\boldsymbol{m}_{-j}^{\setminus 1,t,n},\boldsymbol{V}_{-j}^{\setminus 1,t,n}\right)\mathrm{d}\boldsymbol{x}_{t,-j}\mathcal{N}\left(x_{t,j}|\boldsymbol{m}_j^{\setminus 1,t,n},\boldsymbol{V}_{jj}^{\setminus 1,t,n}\right)\mathrm{d}x_{t,j} \tag{115}$$

Performing a change of variable, $z = \tilde{A}_{n,-j}\boldsymbol{x}_{t,-j}$, reduces the inner integral to a one-dimensional integral and thus, the resulting two nested one-dimensional integrals can be solved using standard results for Gaussian integrals Rasmussen and Williams (2006). The resulting moment becomes:

$$\mathbb{E}\left[x_{t,j}\right] = m_j^{\setminus 1,t,n} + \alpha_{1,t,n}\tilde{a}_{n,j}V_{jj}^{\setminus 1,t,n}, \tag{116}$$

where we have defined $\alpha_{1,t,n} = \frac{\mathcal{N}(z)}{\sqrt{1+b_{1,t,n}}\phi(z)}$. Therefore,

$$\mathbb{E}\left[\boldsymbol{x}_t\right] = \boldsymbol{m}^{\setminus 1,t,n} + \alpha_{1,t,n} \cdot \left(\tilde{\boldsymbol{A}}_{n,\cdot} \circ \operatorname{diag}\left(\boldsymbol{V}^{\setminus 1,t,n}\right)\right). \tag{117}$$

Carrying out similar calculations for $\boldsymbol{x}\boldsymbol{x}^T$ yields

$$\mathbb{V}\left[\boldsymbol{x}_t\right] = \operatorname{diag}\left(\boldsymbol{V}^{\setminus 1,t,n}\right) - \alpha_{1,t,n} \cdot \frac{\left(\tilde{\boldsymbol{A}}_{n,\cdot}\mathbb{E}\left[\boldsymbol{x}_t\right] + \alpha_{1,t,n}\right)}{1 + b_{1,t,n}} \left(\tilde{\boldsymbol{A}}_{n,\cdot} \circ \operatorname{diag}\left(\boldsymbol{V}^{\setminus 1,t,n}\right)\right) \circ \left(\tilde{\boldsymbol{A}}_{n,\cdot} \circ \operatorname{diag}\left(\boldsymbol{V}^{\setminus 1,t,n}\right)\right). \tag{118}$$

Using these moments, we compute the updates for $\tilde{f}_{1,t,n}$ as follows

$$\hat{\boldsymbol{V}}_{1,t,n}^{\text{new}} = \left[\boldsymbol{V}\left[\boldsymbol{x}_t\right]^{-1} - \left(\boldsymbol{V}^{\setminus 1,t,n}\right)^{-1}\right]^{-1}, \tag{119}$$

$$\hat{\boldsymbol{m}}_{1,t,n}^{\text{new}} = \hat{\boldsymbol{V}}_{1,t,n}^{\text{new}}\left[\boldsymbol{V}\left[\boldsymbol{x}_t\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_t\right] - \left(\boldsymbol{V}^{\setminus 1,t,n}\right)^{-1}\boldsymbol{m}^{\setminus 1,t,n}\right]. \tag{120}$$

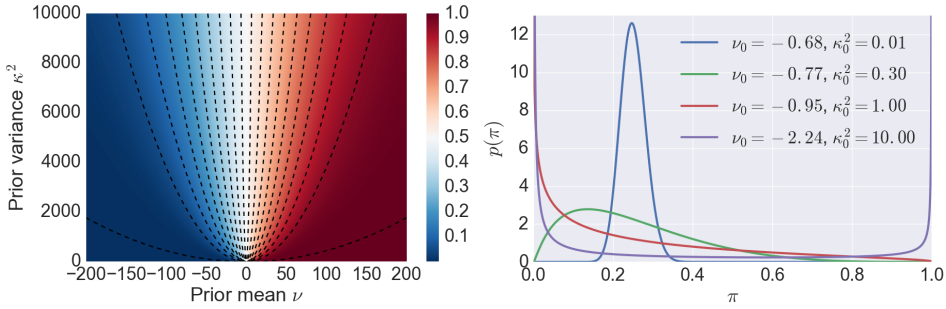
Appendix D. On the prior mean and variance of $\boldsymbol{\Gamma}$

The purpose of this appendix is to elaborate on the interplay between the prior mean and the prior variance of $\boldsymbol{\Gamma}$. For this analysis we will assume that the $\boldsymbol{\Gamma}$ has constant mean $\boldsymbol{\mu}_0 = \nu_0\mathbf{1}$ for $\nu_0 \in \mathbb{R}$, and covariance $\boldsymbol{\Sigma}_0 = \kappa_0^2\boldsymbol{R}_0$, where $\mathbf{1} \in \mathbb{R}^D$ is a column vector of ones and $\boldsymbol{R}_0 \in \mathbb{R}^{D\times D}$ is a correlation matrix. Recall from eq. (8) that the marginal prior probability of $z_i = 1$ is given by

$$\hat{p} = p(z_i = 1) = \int p(z_i = 1|\gamma_i)p(\gamma_i)\mathrm{d}\gamma_i = \int \phi(\gamma_i)\mathcal{N}\left(\gamma_i|\mu_i, \Sigma_{0,ii}\right)\mathrm{d}\gamma_i = \phi\left(\frac{\nu_0}{\sqrt{1+\kappa_0^2}}\right). \tag{121}$$

It is seen from the above expression that the marginal expected sparsity level is controlled by ν_0 and κ_0^2. Figure 15(a) shows the surface of $p(z_i = 1)$ as a function of ν_0 and κ_0^2, where the black dashed isocontours confirm that the same level of marginal expected sparsity can be obtained for any combination of (ν_0, κ_0^2) that satisfies the relationship in eq. (121) for some $\hat{p} \in (0, 1)$. Also, note that the prior probability $\hat{p}$ is by definition equal to the expectation of $\phi(\gamma_i)$, that is $\hat{p} = \mathbb{E}_{p(\gamma_i)}\left[\phi(\gamma_i)\right]$. However, as ϕ is a monotonic function, we can derive the full distribution of $\pi = \phi(\gamma)$ through a change of variable as follows

$$p(\pi) = p_\gamma(\phi^{-1}(\pi))\left|\frac{\mathrm{d}\phi^{-1}(\pi)}{\mathrm{d}\pi}\right| = \mathcal{N}\left(\phi^{-1}(\pi)|\nu_0, \kappa_0^2\right)\left|\frac{\mathrm{d}\phi^{-1}(\pi)}{\mathrm{d}\pi}\right| = \frac{\mathcal{N}\left(\phi^{-1}(\pi)|\nu_0, \kappa_0^2\right)}{\mathcal{N}\left(\phi^{-1}(\pi)|0, 1\right)}. \tag{122}$$

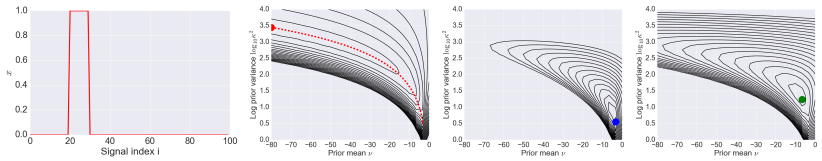


(a) Expected sparsity vs. prior mean and variance

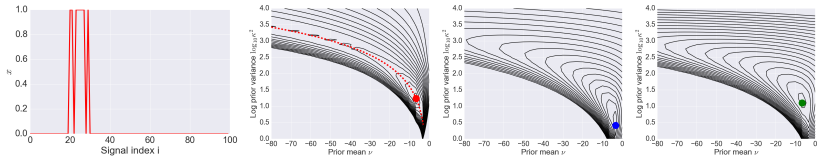
(b) Distribution of $\phi(\gamma_i)$ for $p(z_i = 1) = \frac{1}{4}$

Figure 15: Properties of the prior distribution. (a) Marginal prior probability $p(z_i = 1)$ as a function of (ν_0, κ_0^2). The black dashed lines are isocontours. (b) Distribution of $\pi = \phi(\gamma_i)$ for 4 different pairs of (ν_0, κ_0^2), but for fixed value of $p(z_i = 1)$.

Figure 15(b) shows a plot of the density of π for 4 pairs of (ν_0, κ_0^2) that all satisfy $\hat{p} = \mathbb{E}[\pi] = \frac{1}{4}$. Thus, increasing κ_0^2 while keeping $\mathbb{E}[\pi]$ fixed pushes the mass of $p(\pi)$ to the boundary values. Informally, the distribution of $p(\pi)$ will approach a mixture of two Dirac distributions at 0 and 1 with weights $1 - \mathbb{E}[\pi]$ and $\mathbb{E}[\pi]$, respectively, for very large values of κ_0^2 relative to $\nu_0 \neq 0$. In section 6, we discussed maximum likelihood among other methods for learning the hyperparameters of the structured spike-and-slab model. However, maximum likelihood learning of ν and κ can in some instances give rise to the similar problems as encountered in maximum likelihood learning of logistic regression models on data sets, that are completely separated in one or more dimensions (Gelman et al., 2008). The following small example illustrates the problem. Consider an instance of $\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{\epsilon}$, where $\boldsymbol{x}_1$ is the signal shown in Figure 16(a) and where the signal to noise ratio is such that the true support of the signal can be recovered exactly. The dimensions of the forward model is $\boldsymbol{A} \in \mathbb{R}^{50\times 100}$. Let $\mathbf{R}$ be the squared exponential kernel with lengthscale fixed to 8. Figure 16(c) shows the surface of the marginal likelihood approximation as a function of ν_0 and κ_0^2 while the remaining hyperparameters are kept fixed. The red dot indicates the maximum likelihood solution constrained to the domain shown in the figure. The red dashed line shows a plot of the implicit function $\hat{p}_{ML} = p(z_i = 1) = \phi\left(\frac{\nu_0}{\sqrt{1+\kappa_0^2}}\right)$ that intersects the maximum likelihood solution. It is clear that the likelihood surface has a ridge along the curve satisfying $\hat{p}_{ML} = \phi\left(\frac{\nu_0}{\sqrt{1+\kappa_0^2}}\right)$ and that the likelihood is increasing along that ridge as the magnitude of ν_0 and κ_0^2 increase. Thus, the maximum likelihood solutions pushes to magnitude of ν_0 and κ_0^2 to larger and larger values while keeping the sparsity level $\hat{p}_{ML}$ fixed and therefore, gradient-based optimization of the maximum likelihood w.r.t. (ν_0, κ_0^2) will never converge. However, this problem only occurs when the support is separated as in Figure 16(a). Figure 16(f) shows the marginal likelihood approximation surface for $\boldsymbol{y}_2 = \boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{\epsilon}$, where $\boldsymbol{x}_2$ in Figure 16(b). It is now seen that the maximum likelihood solution is well-defined within the



(a) $\boldsymbol{x}_1$ (b) LPD for $\boldsymbol{y}_1$ (flat prior) (c) LPD for $\boldsymbol{y}_1$ (student t prior) (d) LPD for $\boldsymbol{y}_1$ (log-normal prior)



(e) $\boldsymbol{x}_2$ (f) LPD for $\boldsymbol{y}_2$ (flat prior) (g) LPD for $\boldsymbol{y}_2$ (student t prior) (h) LPD for $\boldsymbol{y}_2$ (log-normal prior)

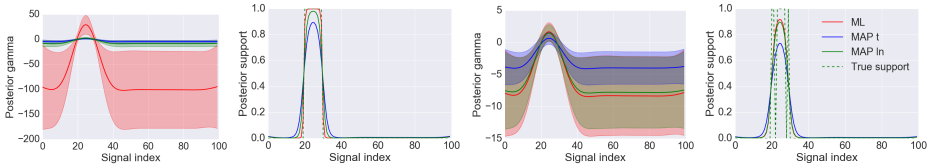
Figure 16: (a) Signal, where the support is contiguous. (b), (c), (d): Log posterior density for $\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{\epsilon}$ with a flat prior, half student t prior (df = 4) and a log normal prior (mean 6, std. dev 3) for κ_0, respectively. (e) Signal, where the support is not contiguous.(f), (g), (h): Log posterior density (LPD) for $\boldsymbol{y}_2 = \boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{\epsilon}$ with a flat prior, half student t prior (df = 4) and a log normal prior (mean 6, std. dev 3) for κ_0, respectively. The red dashed line shows a plot of the implicit function $\hat{p}_{ML} = p(z_i = 1) = \phi\left(\nu_0 \left(1 + \kappa_0^2\right)^{-\frac{1}{2}}\right)$ that intersects the maximum likelihood solution.

interior of $\mathbb{R}^2$. The problem is easily fixed by imposing a weakly informative prior on κ_0 to ensure that the solution is always well-defined. To illustrate this, we re-run this experiment shown in Figure 16(c) with two different priors on κ_0. Figures 16(d)-(e) show the results for a standardized half student t prior with 4 degrees of freedom and a log-normal prior with mean 6 and standard deviation 3, respectively. Figures 16(g)-(h) show the same plots for the signal $\boldsymbol{x}_2$. Figure 17 shows the resulting posterior distribution for both signals with and without priors distributions.

References

M. R. Andersen, O. Winther, and L. K. Hansen. Bayesian inference for structured spike and slab priors. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1745–1753. Curran Associates, Inc., 2014.

Michael R. Andersen, Ole Winther, and Lars Kai Hansen. Spatio-temporal spike and slab priors for MMV problems. *arXiv preprint arXiv:1508.04556*, 2015.



(a) Posterior of $\mathbf{\Gamma}$ for $\boldsymbol{y}_1$ (b) Posterior mean support for $\boldsymbol{y}_1$ (c) Posterior of $\mathbf{\Gamma}$ for $\boldsymbol{y}_2$ (d) Posterior mean support for $\boldsymbol{y}_2$

Figure 17: Posterior distributions for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ for hyperparameter values indicated by the colors dots in Figure 16.

J. Ashburner, G. Barnes, C. Chen, J. Daunizeau, G. Flandin, K. Friston, D. Gitelman, S. Kiebel, J. Kilner, V. Litvak, Rosalyn Moran, W. Penny, K. Stephan, D. Gitelman, R. Henson, C. Hutton, V. Glauche, J. Mattout, and C. Phillips. *SPM8 manual*, July 2010. URL `http://www.fil.ion.ucl.ac.uk/spm/doc/manual.pdf`.

S. Baillet, J. C. Mosher, J. C. Mosher, R. M. Leahy, and R. M. Leahy. Electromagnetic brain mapping. *IEEE Signal Processing Magazine, IEEE Signal Process Mag*, 18(6):14–30, 2001. ISSN 10535888. doi: 10.1109/79.962275.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 0387310738, 9780387310732.

E. Brochu, V. M. Cora, and N de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *CoRR*, abs/1012.2599, 2010. URL `http://arxiv.org/abs/1012.2599`.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

P. Carbonetto and M. Stephens. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.*, 7(1):73–108, March 2012.

C. M. Carvalho, N. G. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In David A. Van Dyk and Max Welling, editors, *AISTATS*, volume 5 of *JMLR Proceedings*, pages 73–80. JMLR.org, 2009.

V. Cevher, M. F. Duarte, C. Hegde, and R. Baraniuk. Sparse signal recovery using Markov random fields. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 257–264. Curran Associates, Inc., 2009.

S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *Signal Processing, IEEE Transactions on*, 53(7):2477–2488, 2005.

D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theor.*, 52(4):1289–1306, April 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.871582. URL http://dx.doi.org/10.1109/TIT.2006.871582.

D. L. Donoho and J. Tanner. Precise undersampling theorems. *Proceedings of the IEEE*, 98 (6):913–924, 2010.

D. L. Donoho, A. Maleki, and A. Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

B. E. Engelhardt and R. P. Adams. Bayesian structured sparsity from Gaussian fields. 8 July 2014.

A. Gelman, A. Jakulin, M. G. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383, December 2008.

M. V. Gerven, B. Cseke, R Oostenveld, and T. Heskes. Bayesian source localization with the multivariate Laplace prior. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1901–1909. Curran Associates, Inc., 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

R. N. Henson, Y. Goshen-Gottstein, T. Ganel, L. J. Otten, A. Quayle, and M. D. Rugg. Electrophysiological And Haemodynamic Correlates Of Face Perception, Recognition And Priming. *Cerebral cortex (New York, N.Y. : 1991)*, 13(7):793–805, July 2003. ISSN 1047-3211. doi: 10.1093/cercor/13.7.793. URL http://dx.doi.org/10.1093/cercor/13.7.793.

R. N. A. Henson, E. Mouchlianitis, and K. J. Friston. MEG and EEG data fusion: Simultaneous localisation of face-evoked responses. *NeuroImage*, 47(2):581–589, 2009.

D. Hernández-Lobato and J. M. Hernández-Lobato. Learning feature selection dependencies in multi-task learning. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 746–754. Curran Associates, Inc., 2013.

D. Hernandez-Lobato, J. M. Hernandez-Lobato, and A. Suarez. Expectation propagation for microarray data classification. *Pattern recognition letters*, 31(12):1618–1626, 2010. ISSN 01678655, 18727344. doi: 10.1016/j.patrec.2010.05.007.

D. Hernandez-Lobato, J. M. Hernandez-Lobato, and A. Suarez. Network-based sparse Bayesian classification. *Pattern recognition*, 44(4):886–900, 2011. ISSN 00313203, 18735142. doi: 10.1016/j.patcog.2010.10.016.

D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14:1891–1945, 2013.

J. M. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez. Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning*, 99(3):437–487, 2015. doi: 10.1007/s10994-014-5475-7. URL http://dx.doi.org/10.1007/s10994-014-5475-7.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 417–424, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553429. URL http://doi.acm.org/10.1145/1553374.1553429.

R. J. Itier and M. J. Taylor. N170 or N1? spatiotemporal differences between object and face processing using ERPs. *Cereb. Cortex*, 14(2):132–142, February 2004.

L. Jacob, G. Obozinski, and J. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, NY, USA, 2009a. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553431. URL http://doi.acm.org/10.1145/1553374.1553431.

L. Jacob, G. Obozinski, and J. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553431. URL http://doi.acm.org/10.1145/1553374.1553431.

R. Jenatton, J. Mairal, F. R. Bach, and G. R. Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 487–494, 2010.

P. Jylänki, J. Vanhatalo, and A. Vehtari. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12:3227–3257, 2011.

P. Jylänki, A. Nummenmaa, and A. Vehtari. Expectation propagation for neural networks with sparsity-promoting priors. *Journal of Machine Learning Research*, 15:1849–1901, 2014. URL http://jmlr.org/papers/v15/jylanki14a.html.

Y. LeCun, C. Cortes, and C. J. C. Burges. The MNIST database of handwritten digits, 1998.

Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2323–2331. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5760-stochastic-expectation-propagation.pdf.

J. Mairal, R. Jenatton, F. R. Bach, and G. R. Obozinski. Network flow algorithms for structured sparsity. In J D Lafferty, C K I Williams, J Shawe-Taylor, R S Zemel, and A Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1558–1566. Curran Associates, Inc., 2010.

X Meng, S Wu, L Kuang, and J Lu. An expectation propagation perspective on approximate message passing. *IEEE Signal Process. Lett.*, 22(8):1194–1197, August 2015.

T. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 362–369, San Francisco, CA, 2001. Morgan Kaufmann.

T. Minka. Divergence measures and message passing. Technical report, 2005.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear-regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988. ISSN 01621459, 1537274x.

F. S. Nathoo, A. Babul, A. Moiseev, N. Virji-Babul, and M. F. Beg. A variational Bayes spatiotemporal model for electromagnetic brain mapping. *Biometrics*, 70(1):132–143, 2014. ISSN 1541-0420. doi: 10.1111/biom.12126. URL `http://dx.doi.org/10.1111/biom.12126`.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996.

D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Commun. ACM*, 53(12):93–100, December 2010. ISSN 0001-0782. doi: 10.1145/1859204.1859229. URL `http://doi.acm.org/10.1145/1859204.1859229`.

G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley,*, 2006.

M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

T. Peltola, P. Jylänki, and A. Vehtari. Expectation propagation for likelihoods depending on an inner product of two multivariate random variables. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 769–777, 2014.

W. D. Penny, N. J. Trujillo-Barreto, and K. J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, January 2005. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.08.034. URL `http://dx.doi.org/10.1016/j.neuroimage.2004.08.034`.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN 0262256835, 1423769902, 9780262256834, 9781423769903.

J. Riihimäki, A. Vehtari, et al. Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448, 2014.

C. J. V. Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. ISBN 0408709294.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2): 319–392, 2009. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00700.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x`.

M. Seeger. Expectation propagation for exponential families. Technical report, 2005.

B Shahriari, K Swersky, Z Wang, R P Adams, and N de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE*, 104(1):148–175, January 2016.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc., 2012.

R. Sundeep. Generalized approximate message passing for estimation with random linear mixing. *CoRR*, abs/1010.5141, 2010. URL `http://arxiv.org/abs/1010.5141`.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 67(1):91–108, 1 February 2005.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, September 2001. ISSN 1532-4435. doi: 10.1162/15324430152748236. URL `http://dx.doi.org/10.1162/15324430152748236`.

M. K. Titsias and M. Lazaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, Nips 2011, Adv. Neural Inf. Process. Syst.: Annu. Conf. Neural Inf. Process. Syst., Nips*, 2011.

J. Vanhatalo, V. Pietiläinen, and A. Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Stat. Med.*, 29(15):1580–1607, 10 July 2010.

G. Varoquaux, P. R. Raamana, D. A. Engemann, Y. Hoyos-Idrobo, A.and Schwartz, and B. Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*, 145(Pt B):166–179, 15 January 2017.

J. P. P Vila and P. Schniter. Expectation-maximization Gaussian-mixture approximate message passing. *Signal Processing, IEEE Transactions on*, 61(19):4658–4672, 2013.

A. Wu, M. Park, O. O. Koyejo, and J. W. Pillow. Sparse Bayesian structure learning with dependent relevance determination priors. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1628–1636. Curran Associates, Inc., 2014a.

A. Wu, M. Park, O. O. Koyejo, and J. W. Pillow. Sparse Bayesian structure learning with dependent relevance determination priors. In *Advances in Neural Information Processing Systems*, pages 1628–1636, 2014b.

L. Yu, H. Sun, J. P. Barbot, and G. Zheng. Bayesian compressive sensing for cluster structured sparse signals. *Signal Processing*, 92(1):259 – 269, 2012. ISSN 0165-1684. doi: http://dx.doi.org/10.1016/j.sigpro.2011.07.015.

Z. Zhang and B. Rao. Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning. *IEEE Journal of Selected Topics in Signal Processing*, 5 (5):912–926, 2011.

J. Ziniel and P. Schniter. Dynamic compressive sensing of time-varying signals via approximate message passing. *IEEE Transactions on signal processing*, 2013a.

J. Ziniel and P. Schniter. Efficient high-dimensional inference in the multiple measurement vector problem. *IEEE Transactions on Signal Processing*, 61(2):340–354, 2013b.

Appendix D

A Hierarchical Model for Time-varying Functional Connectivity

---

D Andersen, M. R., Hansen, L. K., Winther, O., Koyejo, S. and Poldrack, R. (2017), 'A hierarchical model for time-varying functional connectivity'. *Submitted to the International Conference on Machine Learning (ICML) (24/2-2017), 10 pages*

A hierarchical model for time-varying functional connectivity

Michael Riis Andersen MIRI@DTU.DK
Department of Applied Mathematics and Computer Science, Technical University of Denmark

Lars Kai Hansen LKAI@DTU.DK
Department of Applied Mathematics and Computer Science, Technical University of Denmark

Ole Winther OLWI@DTU.DK
Department of Applied Mathematics and Computer Science, Technical University of Denmark

Sanmi Koyejo SANMI@ILLINOIS.EDU
Department of Computer Science, University of Illinois at Urbana-Champaign

Russell Poldrack POLDRACK@STANFORD.EDU
Department of Psychology, Stanford University

Abstract

We propose a probabilistic model for estimating time-varying covariances of a set of multivariate time series. The instantaneous covariance structure is modelled using time-varying and non-negative linear combinations of low rank matrices, where the time-varying weights are controlled by Gaussian processes. We derive a mean-field inference algorithm for the model and we demonstrate the performance of the model using numerical experiments with both synthetic and real data sets.

1. Introduction

Functional connectivity in the brain, defined as the temporal correlation between spatially remote neurophysiological events (Friston et al., 1993), has attracted a lot of interest from the scientific community during the past two decades (van den Heuvel & Hulshoff Pol, 2010; Bastos & Schoffelen, 2015; Fingelkurts et al.; Li et al., 2009; Friston, 2011). *Time-varying functional connectivity* (Calhoun et al., 2014) studies how the temporal correlation of spatially distant regions evolves over time, i.e. the change in coupling over time rather than change within a single region over time. We consider the problem of estimating the instantaneous covariance structure of a set of non-stationary multivariate time series, which is the underlying statistical problem in dynamic functional connectivity.

Researchers have proposed several analysis pipelines and methods for detection and characterization of dynamic functional connectivity in fMRI BOLD data (Hutchison et al., 2013; Damaraju et al., 2014; Calhoun et al., 2014; Gonzalez-Castillo et al., 2015) and in EEG data (Tagliazucchi & Laufs, 2015; Dimitriadis et al., 2009). The majority of these methods are based on the so-called *sliding window* approach, which is a two-step procedure, where the time series is first divided into overlapping windows after which some statistic of interest, e.g. correlation or covariance, is extracted independently for each window. The sliding window analysis is often followed by a clustering analysis of the estimated covariance matrices across time and thereby, implicitly assumes switching dynamics (Allen et al., 2014; Gonzalez-Castillo et al., 2015).

However, recent studies highlight some significant issues with the sliding window approach (Hindriks et al., 2015; Shakil et al., 2016). The window size of the sliding windows is often chosen as a trade-off between statistical power (large window) and time resolution (small windows), but in practice it can be difficult to choose the optimal window size. Furthermore, due to the high dimensional nature and the relatively slow sampling frequency in fMRI scans, the number of dimensions is usually much higher than the number of time points, which can lead to estimators with high variance. Furthermore, some evidence suggests that brain dynamics are continuous rather than discrete switching dynamics (Smith et al., 2012).

To mitigate the above issues, we propose a model-based approach to time-varying covariance estimation based on

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

continuous dynamics rather than discrete switching. The proposed model can be re-cast as a time-varying sparse factor model, where the prior variances of the factors are time-dependent and are modelled by Gaussian processes (Rasmussen & Williams, 2005). Model and inference are described in section 2, results on synthetic data sets with both continuous and discrete switching dynamics as well as real fMRI data sets are presented in section 3 and related work in section 4.

2. Model and Inference

We propose to model the fMRI BOLD time series using a multivariate normal distribution, where the instantaneous covariance matrix changes slowly as a function of time. Further, we propose a hierarchical model for simultaneous analysis of time-varying covariances for multiple subjects using a latent, shared representation of covariance matrices.

2.1. Model

Let $\boldsymbol{x}_t^n \in \mathbb{R}^D$ be the D-dimensional observed time series at time $t \in \{1, 2, \ldots, T\}$ for subject $n \in \{1, 2, \ldots, N\}$ and let $\mathcal{D}$ be the collection of the all observed time series for all subjects. The sampling distribution of $\boldsymbol{x}_t^n$ is assumed to be

$$\boldsymbol{x}_t^n \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_t^n\right), \tag{1}$$

where $\boldsymbol{\Sigma}_t^n$ is the instantaneous covariance matrix time at t for subject n. We propose to decompose $\boldsymbol{\Sigma}_t^n$ into a non-negative weighted sum of components $\boldsymbol{S}_k \in \mathbb{R}^{D \times D}$,

$$\boldsymbol{\Sigma}_t^n = \beta^{-1}\boldsymbol{I} + \sum_{k=1}^{K} \alpha_{k,t}^n \boldsymbol{S}_k, \tag{2}$$

where $\mathcal{S} = \{\boldsymbol{S}_k\}_{k=1}^K$ is a *dictionary of covariance matrix components*, the coefficients $\alpha_{k,t}^n \geq 0$ are a set of non-negative real mixing weights that govern the dynamics, i.e. $\alpha_{k,t}^n$ controls the contribution of $\boldsymbol{S}_k$ to instantaneous covariance $\boldsymbol{\Sigma}_t^n$ at time t for the n'th subject. The parameter $\beta^{-1} > 0$ is a positive real number controlling the amount of additive white noise. The covariance matrix components $\mathcal{S}$ and the noise precision β are assumed to be independent of time and hence, the second order dynamics of the time series $\boldsymbol{x}_t^n$ are governed solely by the set of coefficients $\mathcal{A}^n = \left\{\alpha_{k,t}^n\right\}$. The model given in eq. (2) is depicted as a graphical model in Figure 1(a).

The dictionary of covariance matrix components, $\mathcal{S}$, is shared across both time and subjects. Loosely speaking, $\mathcal{S}$ is a common basis of covariance matrices for all time points and all subjects, where $(\alpha_{1,t}^n, \ldots, \alpha_{K,t}^n)$ is the coordinates for the specific covariance matrix at time t for the n'th subject. This shared representation allows us to pool data across multiple subjects to estimate each $\boldsymbol{S}_k$. As we show, estimating each covariance matrix component, $\boldsymbol{S}_k$, using data from multiple subjects simultaneously rather than estimating them independently for each subject leads to more robust estimates.

We estimate the covariance matrix components, $\mathcal{S}$, and the dynamic mixing weights, $\mathcal{A} = \{\mathcal{A}^n\}_{n=1}^N$, simultaneously from the observed time series $\mathcal{D}$. We use the Bayesian paradigm for inference, i.e. we impose prior distributions on all random variables of interest to inject our prior assumptions and uncertainty into the model and then we seek to obtain the posterior distribution of the random variables given the data. To estimate the quantities of interest, we use the posterior expectation of the random variables conditioned on the data, e.g. $\hat{\boldsymbol{S}}_k = \mathbb{E}\left[\boldsymbol{S}_k | \mathcal{D}\right]$.

We impose sparsity promoting priors on each $\boldsymbol{S}_k$ for robustness and improved interpretation and for the mixing weights, $\mathcal{A}$, we impose prior distributions that promote both sparsity and smoothness in time. Here the sparsity assumption means that the model seeks to explain the instantaneous covariance matrix for a given subject using as few covariance matrix components as possible. The priors are defined in such a way that each $\boldsymbol{\alpha}_{k,\cdot}^n = (\alpha_{k,1}^n, \alpha_{k,2}^n, \ldots, \alpha_{k,T}^n) \in \mathbb{R}^T$ can have separate correlation length in time, i.e. the model can encode both slow and rapid fluctuations.

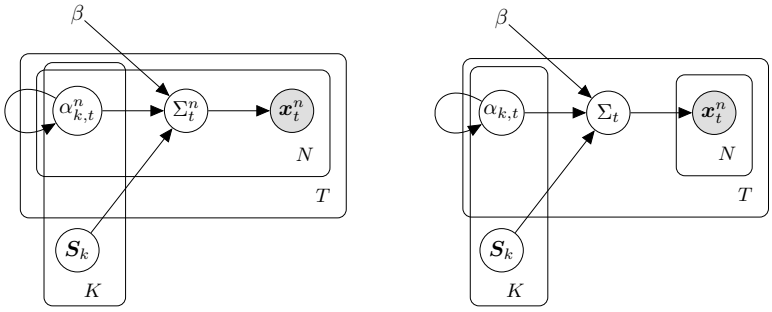
The set of mixing weights $\mathcal{A}$ controls the second order dynamics. From a Bayesian modeling perspective, there are many choices of prior distributions on $\mathcal{A}$. In this work, we choose the prior distribution for $\mathcal{A}$ based on the following three desired properties: sparsity, temporal smoothness and non-negativity. Sparsity ensures that only a small subset of elements from the dictionary $\mathcal{S}$ contribute to the instantaneous covariance $\boldsymbol{\Sigma}_t^n$ at any given time t. Temporal smoothness of $\boldsymbol{\alpha}_k^n$ implies that $\boldsymbol{\Sigma}_t$ will change slowly in time, i.e. two samples $\boldsymbol{x}_t^n$ and $\boldsymbol{x}_{t'}^n$ are more likely to have similar second order moments, if t and t' are close in time. The properties of temporal smoothness and sparsity can both be interpreted as a way to regularize the model. Finally, non-negativity is a technical requirement that ensures that the instantaneous covariance matrix $\boldsymbol{\Sigma}_t^n$ remains positive definite for all time points and for all subjects.

Inspired by the recent success of the use of the so-called rectified linear transformation in the field of deep learning (Nair & Hinton, 2010), we model the set of mixing coefficients as linear rectified Gaussian processes (Rasmussen & Williams, 2005)

$$\alpha_{k,t}^n = \max\left(0, a_{k,t}^n\right) \tag{3}$$

$$\boldsymbol{a}_k^n \sim \mathcal{GP}\left(\boldsymbol{m}_k^n, \boldsymbol{C}_k^n\right). \tag{4}$$

That is, we model each $\boldsymbol{\alpha}_k^n$ as a rectified linear transformation of a Gaussian process with prior mean $\boldsymbol{\mu}_k^n \in \mathbb{R}^T$



(a) Subject specific mixing weights (b) Shared mixing weights across subjects

Figure 1. Graphical representation of the proposed models. The left-most figure shows the model with subject specific mixing weights as in eq. (2) and the right-most figure shows the model with shared mixing weights across subjects as in eq. (9). The self-edges of $\alpha^n_{k,t}$ (left-panel) and $\alpha_{k,t}$ (right-panel) are indicating that the mixing weights are correlated in time (Hensman et al., 2013). The variables $a^n_{k,t}, v_{k,i}$, and $s_{k,i}$ have been left out of the figure for clarity.

and prior covariance $\boldsymbol{C}^n_k \in \mathbb{R}^{T\times T}$. Using this construction, we can explicitly control the smoothness properties of $\boldsymbol{\alpha}^n_k$ using the prior covariance matrix $\boldsymbol{C}_k$. Furthermore, the marginal probability of a given weight $\alpha^n_{k,t}$ being non-zero is given by

$$p(\alpha^n_{k,t} > 0) = p(a^n_{k,t} > 0) = \Phi\left(\frac{\mu^n_{k,t}}{\sqrt{C^n_{k,tt}}}\right), \quad (5)$$

where $\Phi : \mathbb{R} \to (0, 1)$ is the standardized normal cumulative distribution function.

In the proposed hierarchical model, the covariance matrix components $\boldsymbol{S}_k$ are shared across both subjects and time. For simplicity, each $\boldsymbol{S}_k$ is assumed to be sparse, symmetric and of rank one, i.e. $\boldsymbol{S}_k = \boldsymbol{v}_k\boldsymbol{v}_k^T$, where $\boldsymbol{v}_k$ is a sparse vector. To encourage sparsity of $\boldsymbol{v}_k$ we impose the so-called spike-and-slab prior (Mitchell & Beauchamp, 1988) on $\boldsymbol{v}_k$ as follows

$$\boldsymbol{v}_k = \boldsymbol{s}_k \circ \boldsymbol{u}_k \quad (6)$$

$$\boldsymbol{s}_k \sim \prod_{i=1}^{D} \text{Bernoulli}\,(p_k) \quad (7)$$

$$\boldsymbol{u}_k \sim \prod_{i=1}^{D} \mathcal{N}\,(0, \tau_k)\,, \quad (8)$$

where $\circ$ is the element-wise Hadamard product, $s_{k,i} \in \{0, 1\}$ is a binary support variable for $v_{k,i}$, and $p_k \in (0, 1)$ is a hyperparameter controlling the degree of sparsity, i.e. the expected fraction of non-zero entries in $\boldsymbol{v}_k$ is p_k.

There are two main kinds of experiments in neuroimaging, namely "resting state" data – collected while subjects are at rest and "task" data – collected while subjects are actively participating in an experimental task. For our purposes, the task data has the advantage of temporal synchronization i.e. it is reasonable to expect that all subjects will be in similar brain states at similar time-points during the experiment. Thus, we model each subject as i.i.d. observations of the same underlying spatio-temporal process. In particular, we assume that both the spatial maps and dynamic weights are shared, i.e. $\alpha_{k,t} = \alpha^n_{k,t}$ for all n. The simplified model becomes

$$\boldsymbol{\Sigma}_t = \beta^{-1}\boldsymbol{I} + \sum_{k=1}^{K} \alpha_{k,t}\boldsymbol{S}_k. \quad (9)$$

This modelling assumption is reflected in Figure 1(b).

2.2. Approximate Inference

Since $\boldsymbol{S}_k = \boldsymbol{v}_k\boldsymbol{v}_k^T$ is of rank one, we can rewrite eq. (9) as $\boldsymbol{\Sigma}_t = \beta^{-1}\boldsymbol{I} + \boldsymbol{V}\boldsymbol{A}_t\boldsymbol{V}^T$, where $\boldsymbol{V} = \begin{bmatrix}\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_K\end{bmatrix}$ and $\boldsymbol{A}_t = \text{diag}\,(\alpha_{1,t}\ \alpha_{2,t}\ \ldots\ \alpha_{K,t})$. This form is recognized as the marginalized covariance matrix of a factor model with Gaussian latent variables, i.e. $\boldsymbol{x}^n_t \sim \mathcal{N}\left(\boldsymbol{V}\boldsymbol{z}^n_t, \beta^{-1}\boldsymbol{I}\right)$ with $\boldsymbol{z}^n_t \sim \mathcal{N}\,(0, \boldsymbol{A}_t)$ and using the factor model form allows us to derive a more efficient inference scheme. However, the exact posterior distribution of the parameters of interest is intractable and hence, we resort to approximate inference using a mean-field approximation. We use the family of Gaussian distributions to approximate the posterior distributions over $a_{k,t}$ and $z^n_{k,t}$ and we use the family of approximate distributions described in (Lázaro-gredilla & Titsias, 2011) to approximate the posterior of the spike and slab

variables $\boldsymbol{V}$. That is,

$$Q(\boldsymbol{a}_k) = \prod_{t=1}^{T} \mathcal{N}\left(a_{k,t}|\hat{\gamma}_{k,t}, \hat{\lambda}_{k,t}\right) \tag{10}$$

$$Q(\boldsymbol{u}_k|\boldsymbol{s}_k) = \prod_{i=1}^{D} \mathcal{N}\left(u_{i,k}|s_{i,k}\hat{\mu}_{i,k}, s_{i,k}\hat{\tau}_{i,k} + (1-s_{i,k})\tau_k\right)$$

$$Q(\boldsymbol{s}_k) = \prod_{i=1}^{D} \text{Bernoulli}(s_{i,k}|\hat{\pi}_{i,k}), \tag{11}$$

$$Q(\boldsymbol{z}_k^n) = \prod_{t=1}^{T} \mathcal{N}\left(z_{k,t}^n|\hat{\eta}_{k,t}^n, \hat{\theta}_{k,t}^n\right), \tag{12}$$

where $\boldsymbol{z}_k^n = (z_{k,1}^n, \ldots, z_{k,T}^n) \in \mathbb{R}^T$. We minimize the KL-divergence, $\text{KL}\left[Q||P\right]$, between the approximation Q and the exact posterior distribution P by optimizing the Evidence Lower Bound (ELBO) (Blei et al., 2016) w.r.t. to the variational parameters.

3. Numerical Experiments

To study and quantify the performance of the model and the inference algorithm, we conducted a number of numerical experiments using both synthetic data and real fMRI data. There is no notion of ground truth in covariance estimation using real data and therefore, it is hard to objectively evaluate and compare models. However, experiments using simulated data allows us to evaluate model performance using the ground truth quantities.

The model requires a priori specification of the mean function and the functional form of the prior covariance matrix in of the Gaussian process prior distribution in eq. (4). In all experiments, we choose the mean function to be a constant, i.e. $\boldsymbol{m}_k = m_k\mathbf{1} \in \mathbb{R}^T$, and we choose the covariance function for $\boldsymbol{C}_k$ to be the Matérn covariance (Rasmussen & Williams, 2005) function plus a scaled identity matrix

$$\boldsymbol{C}_k(t,t') = \tag{13}$$
$$c_k\left[1+\frac{\sqrt{5}|t-t'|}{\ell_k}+\frac{5|t-t'|^2}{3\ell_k^2}\right]\exp\left[\frac{\sqrt{5}|t-t'|}{\ell_k}\right]+d_k\boldsymbol{I}.$$

The parameters m_k, c_k, d_k, & ℓ_k for each component are estimated from the training data by maximizing the ELBO.

3.1. Simulations

First we investigate the performance of the model using simulated data. To quantify the quality of the estimated covariance matrices, we use the Log-Euclidean Riemannian Metric (LERM), which defines a metric on the manifold of symmetric positive definite matrices (Vemulapalli & Jacobs, 2015; Huang et al., 2015) and is given by

$$\text{LERM}\left(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\right) = \|\log\left(\boldsymbol{\Sigma}_1\right) - \log\left(\boldsymbol{\Sigma}_2\right)\|_F^2, \tag{14}$$

where $\log(\cdot)$ is the matrix logarithm. More specifically, for a sequence of estimated covariance matrices $\left\{\hat{\boldsymbol{\Sigma}}_t\right\}_{t=1}^T$, we compute the time-averaged LERM-distance to the ground truth sequence $\{\boldsymbol{\Sigma}_t\}_{t=1}^T$ as follows

$$\text{Avg. LERM}\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) = \frac{1}{T}\sum_{t=1}^{T}\text{LERM}\left(\boldsymbol{\Sigma}_t, \hat{\boldsymbol{\Sigma}}_t\right). \tag{15}$$

To ease interpretation of this metric and to have a frame of reference, we also compute the Avg. LERM distance to the empirical covariance matrix for all time points, i.e. ignoring any second order dynamics.

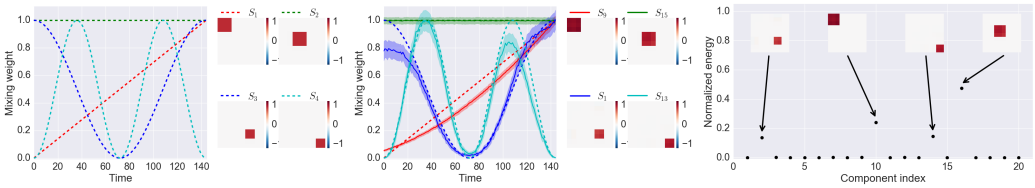
3.1.1. Continuous mixing data set

In the first experiment, we generated time series for a number of subjects using the model in eq. (9) assuming that all subjects share the same time-varying covariance structure. In particular, we generated a sequence of ground truth covariance matrices $\boldsymbol{\Sigma}_t$ using four components and 145 time points, i.e. $K = 4$ and $T = 145$. We ran the experiment with three different of number of dimensions, $D = 10, 30, 50$, respectively. The ground truth mixing weights for the four components are chosen to be a linear function, a constant function, and two sinusoidal functions with different frequency, respectively. The ground truth mixing weights are shown in Figure 2(a) along with the corresponding covariance matrix components $\boldsymbol{S}_k$.

For this experiment, we initialized the model using $K = 20$ random covariance components. The amount of energy that the k'th component contributes to the total energy depends on both $\boldsymbol{S}_k$ and $\boldsymbol{\alpha}_k$ and for each k, we computed the energy contribution for each component as the expectation of $E_k = \text{Trace}\left(\boldsymbol{S}_k\right)\sum_{t=1}^T \alpha_{k,t}$. Figure 2(c) shows the energy contribution of each component along with the corresponding estimated component after fitting the model to a data set with $N = 5$ subjects. It is seen that the model correctly identifies four true non-zero components. Figure 2(b) shows the mean of the posterior distribution for $\boldsymbol{\alpha}_k$ superimposed with 2 standard deviations (left panel) and the corresponding non-zero covariance components (right panel). There is a scaling ambiguity in the model, i.e. in eq. (9) we can scale $\boldsymbol{\alpha}_k$ with some non-zero constant and divide $\boldsymbol{S}_k$ with the same constant to obtain the same covariance matrix $\boldsymbol{\Sigma}_t$. To facilitate comparison with the ground truth values, we scale each estimate of $\boldsymbol{\alpha}_k$ such that the maximum value is 1.

We also compared the performance of the proposed model to competing methods from the literature[1]. Specifically, we

[1]Ideally, we would also compare with the method proposed by Kastner (2016) (see related work in Section 4), but as far as we can tell, the associated software package does not support the problem dimensions used in our experiments.



(a) Ground truth weights & components (b) Estimated weights & components (c) Component Energy

Figure 2. Ground truth and model estimates for simulated experiment using continuous mixing dynamics for $D = 10$ and for $N = 5$ subjects. The leftmost panel in figure (a) shows the true mixing weights, while the rightmost panel in figure (a) shows the corresponding ground truth components. The left and right panel in figure (b) show the estimated mixing weights $\boldsymbol{\alpha}_k$ and the estimated covariance components $\boldsymbol{S}_k$ from the model. (c) Energy contribution for each component normalized wrt. total energy for continuous data set.

consider the sample covariance estimator (completely ignoring any dynamics), the sliding window estimator and the hidden Markov model using multivariate Gaussian emission distributions (HMM) (Rabiner, 1989). We considered three different window sizes for the sliding window estimator, i.e. $L = 10, 20, 30$, where $L \approx 20$ is optimal w.r.t. Avg. LERM metric. For training of the HMMs, we choose the number of states based on log likelihood of data from one hold out subject and hence, here we need at least $N \geq 2$ (one for fitting and one for model selection).

The three rows in Figure 3(a) show the performance of each method as a function of number of training subjects for three different values of number of regions, i.e. $D = 10, 30, 50$, averaged over $R = 20$ realizations of the data. First, it is seen that the proposed method performs uniformly better than the reference methods. Furthermore, it is also seen that as the dimension of data D increases, the performance of the reference methods drop significantly. In particular, when the number of training subjects $N < 10$ and $D > 10$, the sliding window estimators and the HMMs performs equal or worse than the sample covariance estimator.

3.1.2. Discrete switching data set

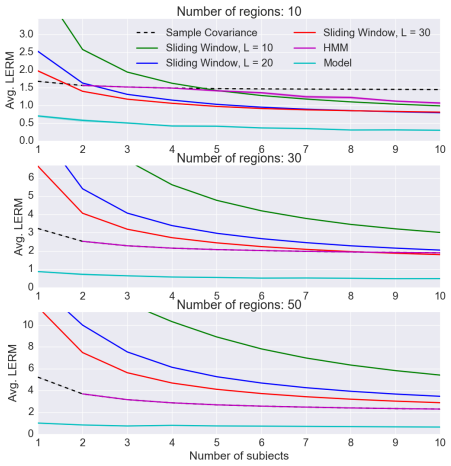
Inspired by the simulated experiments by Calhoun et al. (2014), we also performed an experiment with simulated data where the ground truth covariance switches instantaneously between a set of discrete states rather than continuous mixing as in the previous experiment. As in Calhoun et al. (2014), we considered four different states with four different covariance matrices as shown in Figure 4. Using the same fixed state sequence of $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 2$ for each subject, we generated time series for each subject such that the samples within each state are drawn i.i.d. from a multivariate Gaussian distribution with a state-specific covariance matrix, i.e. the emission model of the Gaussian HMM.

Figure 5 shows the posterior mean for each $\boldsymbol{\alpha}_k$ for the non-zero components and the corresponding covariance components after training the model on $N = 5$ subjects. It is seen that the model captures the dynamics and the covariance matrices for all states correctly, even though the estimated mixing weights are temporally smoothed. This is due to the fact that step functions are not well-modelled by the Gaussian process priors on $\boldsymbol{a}_k$ using generic stationary kernels. The bottom-most panels in Figure 5 show how the model decomposes the four unique covariance matrices into 8 rank one components, where some of the components are used in multiple states. For example, state 1 is decomposed into components $\boldsymbol{S}_{16}$ and $\boldsymbol{S}_{13}$, while state 3 is decomposed into components $\boldsymbol{S}_{13}, \boldsymbol{S}_{10}$, and $\boldsymbol{S}_1$. Thus, the samples from states 1 and 3 both contribute to the estimation of $\boldsymbol{S}_{13}$ even though the complete covariance matrices for the two states are different. The decomposition of distinct states into a set of shared components aligns well with the hypothesis in neuroscience, which states that brain function is decomposable into a set of elementary cognitive processes (Posner et al., 1988). The model also produces accurate estimates of the instantaneous covariance matrices as weighted sums of the estimated covariance components $\hat{\mathcal{S}}$ as evidenced in Figure 3b. It is seen that for $D = 10$, the proposed model outperforms the reference methods for $N < 4$, while it achieves the same level of performance as the HMM for $N \geq 4$. For $D = 30, 50$, the proposed model performs uniformly better than the reference methods.

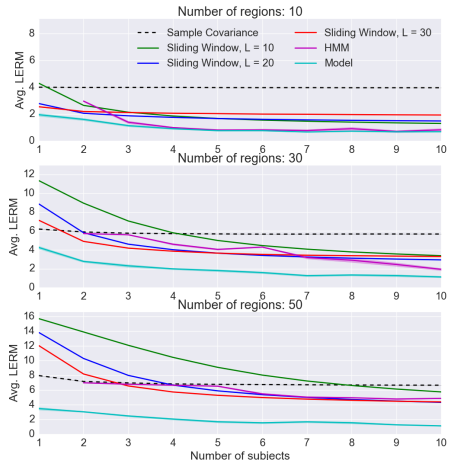
3.2. Analysis of fMRI motor task data

In this experiment, we applied the proposed model to an fMRI data set from the Human Connectome Project (HCP) (Van Essen et al., 2013). Specifically, we analyzed data from 20 subjects from a motor task experiment, where the subjects were asked to perform 5 different motor tasks: left hand tapping, right hand tapping, tongue wagging, left foot tapping and right foot tapping at different time points.

We cannot evaluate our method using the LERM metric since there is no notion of ground truth for this data set. But



(a) Continuous mixing dynamics



(b) Discrete mixing dynamics

Figure 3. Estimator performance as a function of the number of training subjects for two different simulated data sets. The left-most panel shows the results for the data sets with continuous mixing, while the right-most panel shows the results for the data sets with discrete switching. Furthermore, the three rows show the results using $D = 10, 20, 30$ number of regions, respectively.

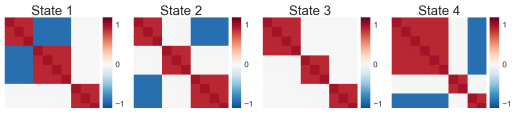


Figure 4. True states for simulated experiment with instantaneous switching dynamics

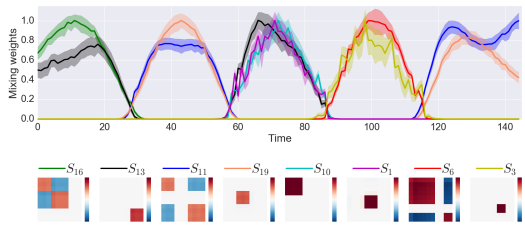


Figure 5. The topmost panel shows estimated mixing weights $\boldsymbol{\alpha}_k$ for each of the non-zero components and the bottommost panel shows the corresponding estimated covariance components $\boldsymbol{S}_k$.

to validate the results and show that the sequences of estimate covariance matrices carry relevant information, we demonstrate that we can distinguish between the different motor tasks using the sequences estimated of covariance matrices from the model.

The multivariate time series for each subject was preprocessed (bandpass filtered and de-trended) and parcellated into 333 regions using the Gordon Atlas (Gordon et al., 2016). Each subject time series has length $T = 284$. Our main interest is to analyze the connectivity structure of the data and hence, the task onset sequences and the motion parameters have been regressed out the time series for each voxel, i.e. we apply the proposed model to the residuals after fitting a linear model to each voxel time series with motion parameters and the task paradigm as explanatory variables.

We divided the data set into a training set and a test set with $N_{\text{training}} = N_{\text{test}} = 10$ subjects. We fitted the model to the training set and for each time point, we used the posterior expectation, $\hat{\boldsymbol{\Sigma}}_t = \mathbb{E}_Q\left[\boldsymbol{\Sigma}_t | \mathcal{D}\right]$, as an estimate of the instantaneous group covariance matrix at time t. Next, we computed the time-averaged group covariance matrix for each task

$$\hat{\boldsymbol{\Sigma}}_{\text{task } i} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \hat{\boldsymbol{\Sigma}}_t, \tag{16}$$

where $\mathcal{T}_i$ is the set of time points for the i'th task and $|\mathcal{T}_i|$ is the number of volumes within task i. That is, we obtain a covariance estimate, $\hat{\Sigma}_{\text{left hand tapping}}, \ldots, \hat{\Sigma}_{\text{right foot tapping}}$, for each task and one for the resting periods between the onsets of the task blocks. Using a flat prior for the task label, $p(\text{task } i)$, we classified the label of each task block of each

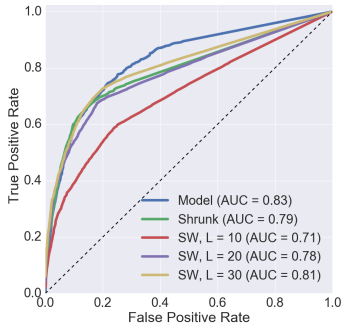


Figure 6. Macro (across all tasks) ROC curves for motor task classification for $N_{\text{train}} = N_{\text{test}} = 10$ subjects.

held-out test subject using Bayes' rule

$$p(\text{task } i|\boldsymbol{X}^*) \propto p(\boldsymbol{X}^*|\text{task } i)p(\text{task } i) \tag{17}$$

$$\propto \prod_{t\in\mathcal{T}_i} \mathcal{N}\left(\boldsymbol{x}_t^*|\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\text{task } i}\right), \tag{18}$$

where $\boldsymbol{X}^*$ is the block of data from the test subject to be classified and $p(\boldsymbol{X}^*|\text{task } i)$ is the likelihood of task i. We compared the proposed method with three reference methods: a regularized covariance matrix estimator, the sliding window approach and random guessing (uniformly). First, the regularized covariance matrix estimator denoted the *shrunk covariance estimator* given by

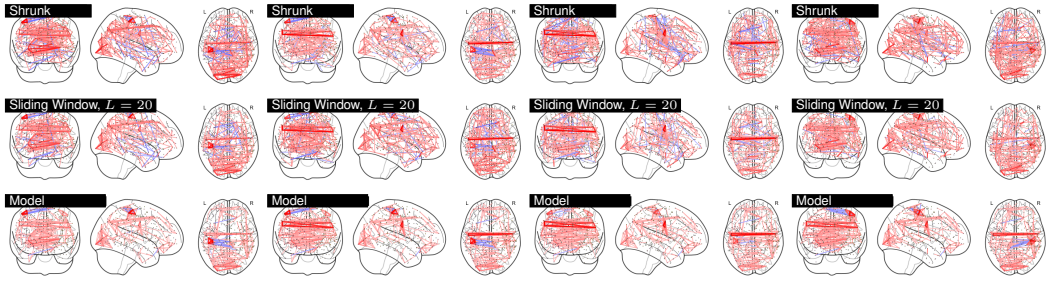
$$\boldsymbol{C}_{\text{task}_i}^{\text{shrunk}} = \gamma\boldsymbol{I} + (1-\gamma)\hat{\boldsymbol{S}}_{\text{task}_i} \tag{19}$$

where $\hat{\boldsymbol{S}}_{\text{task}_i}$ is the sample covariance matrix of the samples within task i across all training subjects and $\gamma \in [0, 1]$ is a shrinkage parameter. This estimator is basically a (regularized) sample covariance matrix of all data point belonging to a given task across all training subjects. In order to be able evaluate the classification likelihoods in eq. (18), it is necessary to use a regularized estimator since the number of regions is larger than the number of time points within each task. Furthermore, we also consider the sliding window estimator (also based on the above shrunk estimator rather than the sample estimator) with window sizes of $10, 20,$ & 30. We use a fixed value of the shrinkage parameter $\gamma = 0.85$ for both $C_{\text{task}_i}^{\text{shrunk}}$ and $C_{\text{task}_i}^{\text{sliding}}$ (The value is chosen based on test classification accuracy using three subjects). For the proposed model, we fixed the initial number of components to $K = 25$ (also based on classification accuracy using three subjects). Figure 6 shows the macro ROC curves, which is a multi-class generalization of the classic ROC curves (Sokolova & Lapalme, 2009), for the multi-class classification problem with $N = 10$, and it is seen that all methods perform better than random. This suggests that covariance matrix estimates produced by the model is indeed time-varying and they contain meaningful information. Figure 7 visualizes the inverses of the estimated task covariance matrices as brain networks using $N = 5$ training subjects for 4 different tasks. It is seen that the general structure of the networks is similar, although the networks from the proposed model are sparser due to the sparsity assumptions of the model. By inspecting the lower panels in Figure 7(a) and (d), it is seen that the 'left finger tapping' task induces a localized network component in the right hand side and vice versa, which is consistent with the expectations from earlier studies (Saladin, 2010).

Figure 8 shows the classification accuracy for the individual tasks as a function of the number of training subjects averaged over $R = 20$ random splits of the data. It is seen that for 4 out of 5 tasks, the proposed method performs as good or better than the reference methods and that the performance in general is increasing as a function of number of training subjects as expected. Furthermore, the classification accuracies for the two finger tapping tasks are in general higher than accuracies of the remaining three tasks. However, the performance for the proposed model for the tongue wagging task is significant worse than the reference methods and the fact the all three sliding window methods have increasing accuracy as a function of number of training subjects suggests that it is possible to do better than random guessing, but for some reason it is not being captured by the model. From the confusion matrix shown in Figure 9(a), it is seen that a large proportion of the blocks belonging to the 'tongue wagging'-class is being classified as either 'left foot tapping' (28%) or 'right foot tapping' (21%). In the experimental paradigm, these two classes appear right before and right after the onset of the 'tongue wagging' task. This might suggest that the model cannot distinguish between these segments. This is also consistent with the fact that estimated networks for the 'left foot tapping'-task and the 'tongue wagging'-task are very similar as evidenced in Figure 7(b)-(c). Finally, we also compared the set of estimated dynamic mixing weights to the task onset sequences of the experimental paradigm. The top-most panels in Figure 10 show the task activation sequences for three different tasks superimposed with the posterior mean of the $\boldsymbol{\alpha}_k$ that matches the support of the specific task activation. The bottom-most panels visualizes the corresponding covariance matrix components. Interestingly, the results show that the networks expected to be associated with left and right hand movements (right and left motor cortex, respectively) have indeed non-zero mixing weights near the task-onsets of the relevant tasks.

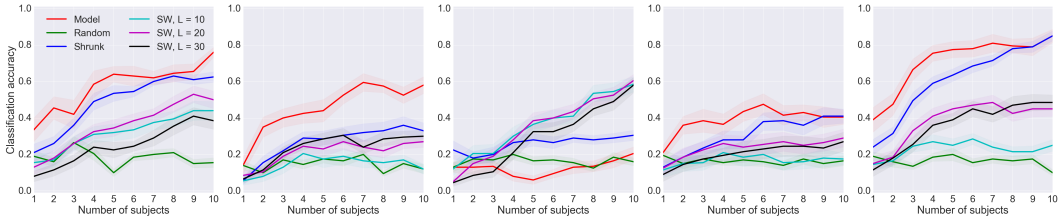
4. Related work

Related work includes sliding window methods (Hutchison et al., 2013; Damaraju et al., 2014; Calhoun et al.,



(a) Right hand tapping (b) Left foot tapping (c) Tongue wagging (d) Left hand tapping

Figure 7. Visualization of estimated precision matrices for four different tasks.



(a) Right hand tapping (b) Left foot tapping (c) Tongue wagging (d) Right foot tapping (e) Left hand tapping

Figure 8. Classification accuracy for each task averaged over 20 random splits.

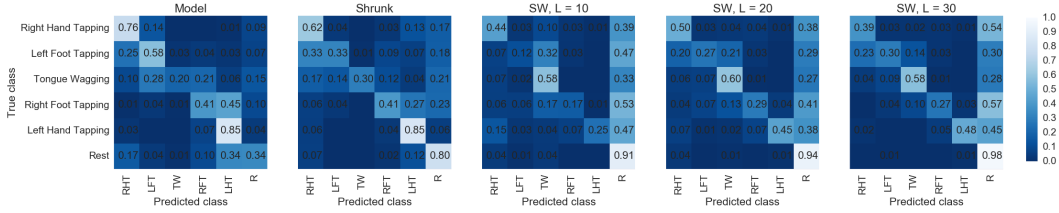
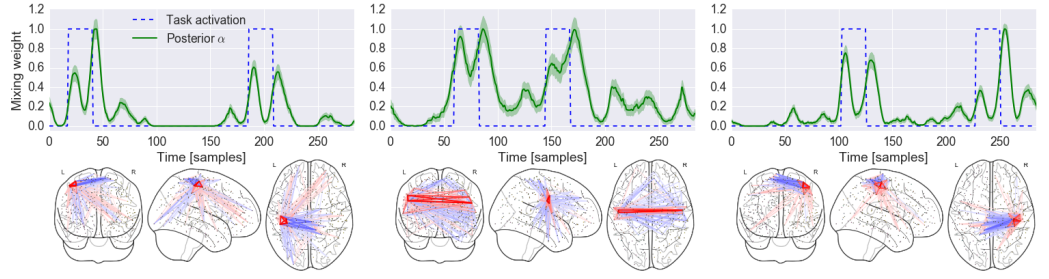


Figure 9. Confusion matrices of classification accuracies across all tasks for each method.



(a) Right hand tapping (b) Tongue wagging (c) Left hand tapping

Figure 10. Top panels: Dynamic mixing weights superimposed with task activation pattern. Bottom panels: Visualization of corresponding covariance matrix component. All plots are extracted from a random split with 5 training subjects.

2014; Gonzalez-Castillo et al., 2015) as mentioned in the introduction. Furthermore, hidden Markov models (Rabiner, 1989), state-space models (Yang et al., 2016; Olsson & Hansen, 2006) and independent component analysis models (Dyrholm et al., 2007; Hyvarinen & Oja, 2000) are also relevant. Finally, the factor model representation we propose here is similar to the model described by Kastner (2016) with the important differences that we use a linear rectified Gaussian processes to encode smoothness in the temporal evolution of the prior variances of the factor rather than first order Markovian dynamics and we use spike-and-slab priors to sparsify the loading matrix rather than conjugate scale-mixture priors. Finally, Kastner (2016) uses MCMC for inference, while we use a variational approximation to improve the scaling properties of the algorithm.

5. Conclusion and Future Work

We described a probabilistic model for time-varying covariance estimation, where the dynamics of the model is captured by a set of Gaussian processes. We proposed a mean-field inference algorithm for the model and evaluated it using synthetic data. We also applied the algorithm to an fMRI motor task data set, where we demonstrated that the estimated covariance matrices were predictive of the task label for hold out subjects in the experiment.interesting Future work includes extending the model and inference algorithm to analyze resting state and task data simultaneously. Furthermore, the prior in eq. (6)-(8) assumes that the each entry in $\boldsymbol{v}_k$ is independent, but the model can be extended to include a priori knowledge of spatial dependencies using structured spike and slab priors (Andersen et al., 2014). Including such information could improve robustness of the model.

References

Allen, Elena A, Damaraju, Eswar, Plis, Sergey M, Erhardt, Erik B, Eichele, Tom, and Calhoun, Vince D. Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex*, 24(3):663–676, March 2014.

Andersen, Michael R, Winther, Ole, and Hansen, Lars K. Bayesian inference for structured spike and slab priors. In Ghahramani, Z, Welling, M, Cortes, C, Lawrence, N D, and Weinberger, K Q (eds.), *Advances in Neural Information Processing Systems 27*, pp. 1745–1753. Curran Associates, Inc., 2014.

Bastos, André M and Schoffelen, Jan-Mathijs. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.*, 9: 175, 2015.

Blei, David M, Kucukelbir, Alp, and McAuliffe, Jon D. Variational inference: A review for statisticians. 4 January 2016.

Calhoun, Vince D, Miller, Robyn, Pearlson, Godfrey, and Adalı, Tulay. The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron*, 84(2):262–274, 22 October 2014.

Damaraju, E, Allen, E A, Belger, A, Ford, J M, McEwen, S, Mathalon, D H, Mueller, B A, Pearlson, G D, Potkin, S G, Preda, A, Turner, J A, Vaidya, J G, van Erp, T G, and Calhoun, V D. Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *Neuroimage Clin*, 5:298–308, 24 July 2014.

Dimitriadis, Stavros I, Laskaris, Nikolaos A, Del Rio-Portilla, Yolanda, and Koudounis, George Ch. Characterizing dynamic functional connectivity across sleep stages from EEG. *Brain Topogr.*, 22(2):119–133, September 2009.

Dyrholm, Mads, Makeig, Scott, and Hansen, Lars Kai. Convolutive ica for spatio-temporal analysis of eeg. *Neural Computation*, 19:934–955, 2007. ISSN 1530888x, 08997667.

Fingelkurts, Andrew A, Fingelkurts, Alexander A, and Kähkönen, Seppo. Functional connectivity in the brain – is it an elusive concept?

Friston, K J, Frith, C D, Liddle, P F, and Frackowiak, R S. Functional connectivity: the principal-component analysis of large (PET) data sets. *J. Cereb. Blood Flow Metab.*, 13(1):5–14, January 1993.

Friston, Karl J. Functional and effective connectivity: a review. *Brain Connect.*, 1(1):13–36, 2011.

Gonzalez-Castillo, Javier, Hoy, Colin W, Handwerker, Daniel A, Robinson, Meghan E, Buchanan, Laura C, Saad, Ziad S, and Bandettini, Peter A. Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proc. Natl. Acad. Sci. U. S. A.*, 112(28):8762–8767, 14 July 2015.

Gordon, Evan M, Laumann, Timothy O, Adeyemo, Babatunde, Huckins, Jeremy F, Kelley, William M, and Petersen, Steven E. Generation and evaluation of a cortical area parcellation from Resting-State correlations. 26(1): 288–303, January 2016.

Hensman, James, Fusi, Nicolo, and Lawrence, Neil D. Gaussian processes for big data. 26 September 2013.

Hindriks, R, Adhikari, M H, Murayama, Y, Ganzetti, M, Mantini, D, Logothetis, N K, and Deco, G. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *Neuroimage*, 26 November 2015.

Huang, Zhiwu, Wang, Ruiping, Shan, Shiguang, Li, Xianqiu, and Chen, Xilin. Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 720–729, 2015.

Hutchison, R Matthew, Womelsdorf, Thilo, Allen, Elena A, Bandettini, Peter A, Calhoun, Vince D, Corbetta, Maurizio, Della Penna, Stefania, Duyn, Jeff H, Glover, Gary H, Gonzalez-Castillo, Javier, Handwerker, Daniel A, Keilholz, Shella, Kiviniemi, Vesa, Leopold, David A, de Pasquale, Francesco, Sporns, Olaf, Walter, Martin, and Chang, Catie. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378, 15 October 2013.

Hyvarinen, A and Oja, E. Independent component analysis: algorithms and applications. *Neural Networks*, 13 (4-5):411–430, 2000. ISSN 18792782, 08936080. doi: 10.1016/S0893-6080(00)00026-5.

Kastner, Gregor. Sparse bayesian time-varying covariance estimation in many dimensions. 19 July 2016.

Lázaro-gredilla, Miguel and Titsias, Michalis K. Spike and slab variational inference for Multi-Task and multiple kernel learning. In Shawe-Taylor, J, Zemel, R S, Bartlett, P L, Pereira, F, and Weinberger, K Q (eds.), *Advances in Neural Information Processing Systems 24*, pp. 2339–2347. Curran Associates, Inc., 2011.

Li, Kaiming, Guo, Lei, Nie, Jingxin, Li, Gang, and Liu, Tianming. Review of methods for functional brain connectivity detection using fMRI. *Comput. Med. Imaging Graph.*, 33(2):131–139, March 2009.

Mitchell, T J and Beauchamp, J J. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, 83(404): 1023–1032, 1988.

Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

Olsson, Rasmus Kongsgaard and Hansen, Lars Kai. Linear state-space models for blind source separation. *Journal of Machine Learning Research*, 7:2585–2602, 2006. ISSN 15337928, 15324435.

Posner, Michael I., Petersen, Steven E., Fox, Peter T., and Raichle, Marcus E. Localization of cognitive operations in the human brain. *Science*, 240(4859): 1627–1631, 1988. ISSN 10959203, 00368075. doi: 10.2307/1701013, 10.2307/1701013.

Rabiner, L R. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, February 1989.

Rasmussen, Carl Edward and Williams, Christopher K I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Saladin, Kenneth S. *Anatomy & physiology : the unity of form and function*. 2010. ISBN 9780071283410. URL `http://www.worldcat.org/isbn/9780071283410`.

Shakil, Sadia, Lee, Chin-Hui, and Keilholz, Shella Dawn. Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *Neuroimage*, 133:111–128, 4 March 2016.

Smith, Stephen M, Miller, Karla L, Moeller, Steen, Xu, Junqian, Auerbach, Edward J, Woolrich, Mark W, Beckmann, Christian F, Jenkinson, Mark, Andersson, Jesper, Glasser, Matthew F, Van Essen, David C, Feinberg, David A, Yacoub, Essa S, and Ugurbil, Kamil. Temporally-independent functional modes of spontaneous brain activity. *Proc. Natl. Acad. Sci. U. S. A.*, 109 (8):3131–3136, 21 February 2012.

Sokolova, Marina and Lapalme, Guy. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.*, 45(4):427–437, 2009.

Tagliazucchi, Enzo and Laufs, Helmut. Multimodal imaging of dynamic functional connectivity. *Front. Neurol.*, 6:10, 16 February 2015.

van den Heuvel, Martijn P and Hulshoff Pol, Hilleke E. Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.*, 20(8):519–534, August 2010.

Van Essen, David C, Smith, Stephen M, Barch, Deanna M, Behrens, Timothy E J, Yacoub, Essa, Ugurbil, Kamil, and WU-Minn HCP Consortium. The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79, 15 October 2013.

Vemulapalli, Raviteja and Jacobs, David W. Riemannian metric learning for symmetric positive definite matrices. 10 January 2015.

Yang, Ying, Aminoff, Elissa, Tarr, Michael, and Robert, Kass E. A state-space model of cross-region dynamic

connectivity in meg/eeg. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1234–1242. Curran Associates, Inc., 2016.

Bibliography

E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex*, 24(3):663–676, Mar. 2014.

M. R. Andersen. Sparse inference using approximate message passing. Master's thesis, M. Sc Thesis, Technical University of Denmark, 2014.

C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Mach. Learn.*, 50(1-2):5–43, 1 Jan. 2003.

S. Baillet, S. Baillet, J. C. Mosher, J. C. Mosher, R. M. Leahy, and R. M. Leahy. Electromagnetic brain mapping. *IEEE Signal Process. Mag.*, 18(6): 14–30, 2001.

L. Baldassarre, J. Mourao-Miranda, and M. Pontil. Structured sparsity models for brain decoding from fMRI data. In *2012 Second International Workshop on Pattern Recognition in NeuroImaging*, pages 5–8, July 2012.

S. Barthelmé and N. Chopin. Expectation-Propagation for Likelihood-Free inference. 29 July 2011.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. 4 Jan. 2016.

V. D. Calhoun, R. Miller, G. Pearlson, and T. Adalı. The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron*, 84(2):262–274, 22 Oct. 2014.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

F. Caron and A. Doucet. Sparse bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 88–95, New York, NY, USA, 2008. ACM.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *International Conference on Artificial Intelligence and Statistics*, pages 73–80, 2009.

G. Casella and R. L. Berger. *Statistical Inference*. Thomson Learning, 2002.

V. Cevher, M. F. Duarte, C. Hegde, and R. Baraniuk. Sparse signal recovery using Markov random fields. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 257–264. Curran Associates, Inc., 2009.

C. Chen and J. Huang. Compressive sensing MRI with wavelet tree sparsity. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1115–1123. Curran Associates, Inc., 2012.

S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *Signal Processing, IEEE Transactions on*, 53(7):2477–2488, 2005.

J. P. Cunningham, P. Hennig, and S. Lacoste-Julien. Gaussian probabilities and expectation propagation. 2013.

E. Damaraju, E. A. Allen, A. Belger, J. M. Ford, S. McEwen, D. H. Mathalon, B. A. Mueller, G. D. Pearlson, S. G. Potkin, A. Preda, J. A. Turner, J. G. Vaidya, T. G. van Erp, and V. D. Calhoun. Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *Neuroimage Clin*, 5:298–308, 24 July 2014.

G. Dehaene and S. Barthelmé. Expectation propagation in the large-data limit. 27 Mar. 2015.

A. Delorme, J. Palmer, J. Onton, R. Oostenveld, and S. Makeig. Independent EEG sources are dipolar. *PLoS One*, 7(2):e30135, 15 Feb. 2012.

D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, Apr. 2006.

D. L. Donoho and J. Tanner. Precise undersampling theorems. *Proc. IEEE*, 98 (6):913–924, 2010.

D. L. Donoho, A. Maleki, and A. Montanari. The Noise-Sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory*, 57(10):6920–6941, Oct. 2011.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 15(12): 3736–3745, Dec. 2006.

Y. C. Eldar. *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, 9 Apr. 2015.

D. Engemann, D. Strohmeier, E. Larson, and A. Gramfort. Mind the noise covariance when localizing brain sources with M/EEG. In *2015 International Workshop on Pattern Recognition in NeuroImaging*, pages 9–12, June 2015.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Stat. Sin.*, 20(1):101–148, Jan. 2010.

S. Flaxman, A. Wilson, D. Neill, H. Nickisch, and A. Smola. Fast kronecker inference in gaussian processes with non-gaussian likelihoods. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 607–616, 2015.

K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattout. Multiple sparse priors for the M/EEG inverse problem. *Neuroimage*, 39(3):1104–1120, 1 Feb. 2008.

K. J. Friston. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier/Academic Press, 2007.

K. J. Friston, C. D. Frith, P. F. Liddle, and R. S. Frackowiak. Functional connectivity: the principal-component analysis of large (PET) data sets. *J. Cereb. Blood Flow Metab.*, 13(1):5–14, Jan. 1993.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press, 1 Nov. 2013.

A. Gelman, A. Vehtari, P. Jylänki, T. Sivula, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. Robert. Expectation propagation as a way of life: A framework for bayesian inference on partitioned data. 2017.

M. V. Gerven, B. Cseke, R. Oostenveld, and T. Heskes. Bayesian source localization with the multivariate Laplace prior. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1901–1909. Curran Associates, Inc., 2009.

J. Gonzalez-Castillo, C. W. Hoy, D. A. Handwerker, M. E. Robinson, L. C. Buchanan, Z. S. Saad, and P. A. Bandettini. Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proc. Natl. Acad. Sci. U. S. A.*, 112(28):8762–8767, 14 July 2015.

I. F. Gorodnitsky, J. S. George, and B. D. Rao. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. *Electroencephalogr. Clin. Neurophysiol.*, 95(4):231–251, Oct. 1995.

S. T. Hansen and L. K. Hansen. Spatio-temporal reconstruction of brain dynamics from EEG with a markov prior. *Neuroimage*, 148:274–283, 1 Mar. 2017.

J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 28 Sept. 2001.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. 26 Sept. 2013.

R. Henson, Y. Goshen-Gottstein, T. Ganel, L. Otten, A. Quayle, and M. Rugg. Electrophysiological and haemodynamic correlates of face perception, recognition and priming. *Cerebral Cortex*, 13(7):793–805, 2003. ISSN 14602199, 10473211. doi: 10.1093/cercor/13.7.793.

R. N. A. Henson, E. Mouchlianitis, and K. J. Friston. MEG and EEG data fusion: Simultaneous localisation of face-evoked responses. *Neuroimage*, 47 (2):581–589, 2009.

D. Hernandez-Lobato, J. M. Hernandez-Lobato, and A. Suarez. Expectation propagation for microarray data classification. *Pattern Recognit. Lett.*, 31 (12):1618–1626, 2010.

D. Hernandez-Lobato, J. M. Hernandez-Lobato, and A. Suarez. Network-based sparse Bayesian classification. *Pattern Recognit.*, 44(4):886–900, 2011.

D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Generalized Spike-and-Slab priors for Bayesian group feature selection using expectation propagation. *J. Mach. Learn. Res.*, 14:1891–1945, 2013.

J. M. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez. Expectation propagation in linear regression models with spike-and-slab priors. *Mach. Learn.*, 99(3):437–487, 1 June 2015.

T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 216–223. Morgan Kaufmann Publishers Inc., 1 Aug. 2002.

T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter. Approximate inference techniques with expectation constraints. *J. Stat. Mech.*, 2005(11): P11015, 30 Nov. 2005.

R. Hindriks, M. H. Adhikari, Y. Murayama, M. Ganzetti, D. Mantini, N. K. Logothetis, and G. Deco. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *Neuroimage*, 127:242–256, 15 Feb. 2016.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 417–424, New York, NY, USA, 2009. ACM.

Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 720–729, 2015.

R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. Della Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo, D. A. Handwerker, S. Keilholz, V. Kiviniemi, D. A. Leopold, F. de Pasquale, O. Sporns, M. Walter, and C. Chang. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378, 15 Oct. 2013.

L. Jacob, G. Obozinski, and J. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, NY, USA, 2009. ACM.

M. A. Jatoi, N. Kamel, A. S. Malik, I. Faye, and T. Begum. A survey of methods used for source localization using EEG signals. *Biomed. Signal Process. Control*, 11:42–52, 2014.

B. D. Jeffs. Sparse inverse solution methods for signal and image processing applications. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 3, pages 1885–1888 vol.3, May 1998.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1 Nov. 1999.

S. C. Jun, S. M. Plis, D. M. Ranken, and D. M. Schmidt. Spatiotemporal noise covariance estimation from limited empirical magnetoencephalographic data. *Phys. Med. Biol.*, 51(21):5549–5564, 7 Nov. 2006.

P. Jylänki, J. Vanhatalo, and A. Vehtari. Robust Gaussian process regression with a Student-t likelihood. *J. Mach. Learn. Res.*, 12:3227–3257, 2011.

M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary gaussian process classification. *J. Mach. Learn. Res.*, 6(Oct):1679–1704, 2005.

B. Lautrup, L. K. Hansen, I. Law, C. Svarer, and S. C. Strother. Massive weight sharing: A cure for extremely Ill-Posed problems. In *Supercomputing in Brain Research: From Tomography to Neural Networks. World Scientific Pub. Corp*, pages 137–148. World Scientific, 1994.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, 2007.

Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2323–2331. Curran Associates, Inc., 2015.

D. J. C. MacKay. Hyperparameters: Optimize, or integrate out? In G. R. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, Fundamental Theories of Physics, pages 43–59. Springer Netherlands, 1996.

D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 25 Sept. 2003.

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. R. Stat. Soc. Series B Stat. Methodol.*, 70(1):53–71, 2008.

V. Michel, A. Gramfort, G. Varoquaux, and B. Thirion. Total variation regularization enhances Regression-Based brain activity prediction. In *2010 First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging*, pages 9–12, Aug. 2010.

T. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 362–369, San Francisco, CA, 2001. Morgan Kaufmann.

T. Minka. Power ep. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

T. Minka. Divergence measures and message passing. Technical report, 2005.

T. Minka. The EP energy function and minimization schemes. Aug. 2007.

T. J. Mitchell and J. J. Beauchamp. BAYESIAN VARIABLE SELECTION IN LINEAR-REGRESSION. *J. Am. Stat. Assoc.*, 83(404):1023–1032, 1988.

S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian and L1 approaches to sparse unsupervised learning. 6 June 2011.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

F. Natterer and G. Wang. The mathematics of computerized tomography. *Med. Phys.*, 29(1):107–108, 1 Jan. 2002.

H. Nickisch and C. E. Rasmussen. Approximations for binary gaussian process classification. *J. Mach. Learn. Res.*, 9(Oct):2035–2078, 2008.

S. F. V. Nielsen, K. H. Madsen, R. Røge, M. N. Schmidt, and M. Mørup. Nonparametric modeling of dynamic functional connectivity in fMRI data. 4 Jan. 2016.

P. L. Nunez and R. Srinivasan. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, Jan. 2006.

J. Onton and S. Makeig. Information-based modeling of event-related brain dynamics. *Prog. Brain Res.*, 159:99–120, 2006.

M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Comput.*, 12(11):2655–2684, 2000.

M. Opper and O. Winther. Expectation consistent approximate inference. *J. Mach. Learn. Res.*, 6(Dec):2177–2204, 2005.

W. Ou, P. Golland, and M. Hämäläinen. A distributed spatio-temporal EEG/MEG inverse solver. *Med. Image Comput. Comput. Assist. Interv.*, 11(Pt 1):26–34, 2008.

G. Parisi. *Statistical Field Theory*. Addison-Wesley, 1988.

T. Peltola, P. Jylänki, and A. Vehtari. {Expectation Propagation for Likelihoods Depending on an Inner Product of Two Multivariate Random Variables}. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 769–777, 2014.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6(Dec):1939–1959, 2005.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, Feb. 1989.

B. D. Rao. Signal processing with the sparseness constraint. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 3, pages 1861–1864 vol.3, May 1998.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, and S. C. Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognit.*, 45(6):2085–2100, 2012.

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1530–1538, 2015.

I. Rish. Functional MRI analysis with sparse models. In *Machine Learning and Knowledge Discovery in Databases*, pages 632–636. Springer, Berlin, Heidelberg, 23 Sept. 2013.

I. Rish and G. Grabarnik. *Sparse Modeling: Theory, Algorithms, and Applications*. CRC Press, 1 Dec. 2014.

V. Ročková and E. I. George. The Spike-and-Slab LASSO. *J. Am. Stat. Assoc.*, 0(ja):0–0, 2016.

H. Rue and S. Martino. Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations.

P. Sallee and B. A. Olshausen. Learning sparse multiscale image representations. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1351–1358. MIT Press, 2003.

M. Seeger. Expectation propagation for exponential families. Technical report, 2005.

M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf. Optimization of k-space trajectories for compressed sensing by bayesian experimental design. *Magn. Reson. Med.*, 63(1):116–126, Jan. 2010.

S. Shakil, C.-H. Lee, and S. D. Keilholz. Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *Neuroimage*, 133:111–128, 4 Mar. 2016.

S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 22 Nov. 2003.

S. M. Smith, K. L. Miller, S. Moeller, J. Xu, E. J. Auerbach, M. W. Woolrich, C. F. Beckmann, M. Jenkinson, J. Andersson, M. F. Glasser, D. C. Van Essen, D. A. Feinberg, E. S. Yacoub, and K. Ugurbil. Temporally-independent functional modes of spontaneous brain activity. *Proc. Natl. Acad. Sci. U. S. A.*, 109(8):3131–3136, 21 Feb. 2012.

E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In Y. Weiss, P. B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press, 2006.

J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 12 Mar. 2015.

M. Sourty, L. Thoraval, D. Roquet, J.-P. Armspach, J. Foucher, and F. Blanc. Identifying dynamic functional connectivity changes in dementia with lewy bodies based on product hidden markov models. *Front. Comput. Neurosci.*, 10:60, 23 June 2016.

C. Stahlhut, H. T. Attias, D. Wipf, L. K. Hansen, and S. S. Nagarajan. Probabilistic M/EEG source imaging from sparse spatio-temporal event structure. In *2nd NIPS Workshop on Machine Learning and Interpretation in NeuroImaging (MLINI 2012)*, 2012.

O. Stegle, C. Lippert, J. M. Mooij, N. D. Lawrence, and K. M. Borgwardt. Efficient inference in matrix-variate gaussian models with iid observation noise. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 630–638. Curran Associates, Inc., 2011.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58:267–288, 1994.

M. K. Titsias. Variational learning of inducing variables in sparse gaussian processes. *AISTATS*, 2009.

M. K. Titsias and M. Lazaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, Nips 2011, Adv. Neural Inf. Process. Syst. : Annu. Conf. Neural Inf. Process. Syst. , Nips*, 2011.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, K. Ugurbil, and WU-Minn HCP Consortium. The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79, 15 Oct. 2013.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1 Jan. 1979.

J. Vanhatalo, V. Pietiläinen, and A. Vehtari. Approximate inference for disease mapping with sparse gaussian processes. *Stat. Med.*, 29(15):1580–1607, 10 July 2010.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, Jan. 1995.

G. Varoquaux, M. Kowalski, and B. Thirion. Social-sparsity brain decoders: faster spatial sparsity. In *2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4, June 2016.

G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*, 145(Pt B):166–179, 15 Jan. 2017.

R. Vemulapalli and D. W. Jacobs. Riemannian metric learning for symmetric positive definite matrices. 10 Jan. 2015.

D. Vidaurre, A. J. Quinn, A. P. Baker, D. Dupret, A. Tejero-Cantero, and M. W. Woolrich. Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage*, 126:81–95, 1 Feb. 2016.

J. P. P. Vila and P. Schniter. Expectation-maximization Gaussian-mixture approximate message passing. *Signal Processing, IEEE Transactions on*, 61(19): 4658–4672, 2013.

E. M. Wainwright. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, 2008.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1 (1–2):1–305, 2008.

C. K. I. Williams and D. Barber. Bayesian classification with gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1342–1351, Dec. 1998.

A. Wilson, E. Gilboa, J. P. Cunningham, and A. Nehorai. Fast kernel learning for multidimensional pattern extrapolation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3626–3634. Curran Associates, Inc., 2014.

J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proc. IEEE*, 98 (6):1031–1044, June 2010.

Y. Yang, E. Aminoff, M. Tarr, and K. E. Robert. A state-space model of cross-region dynamic connectivity in MEG/EEG. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1234–1242. Curran Associates, Inc., 2016.

L. Yu, H. Sun, J. P. Barbot, and G. Zheng. Bayesian compressive sensing for cluster structured sparse signals. *Signal Processing*, 92(1):259–269, 2012.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.*, 68(1):49–67, 1 Feb. 2006.

R. Zhang, C. Czado, and K. Sigloch. A bayesian linear model for the high-dimensional inverse problem of seismic tomography. *Ann. Appl. Stat.*, 7(2): 1111–1138, June 2013.

J. Ziniel and P. Schniter. Dynamic compressive sensing of Time-Varying signals via approximate message passing. *IEEE Trans. Signal Process.*, 2013a.

J. Ziniel and P. Schniter. Efficient High-Dimensional inference in the multiple measurement vector problem. *IEEE Trans. Signal Process.*, 61(2):340–354, 2013b.

J. Ziniel, L. C. Potter, and P. Schniter. Tracking and smoothing of time-varying sparse signals via approximate belief propagation. In *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, pages 808–812, Nov. 2010.