Journal of Machine Learning Research 18 (2017) 1-58 Submitted 9/15; Revised 3/17; Published 12/17

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

Michael Riis Andersen* MICHAEL.ANDERSEN@AALTO.FI
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
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 problemblems 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 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 prob-

∗. Work done mainly while at Department of Applied Mathematics and Computer Science, Technical University of Denmark

lems 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 $\text{Ber}(z|p)$ 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

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

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

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

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

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

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 $\phi\left(\gamma_i\right)$ 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

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

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 $\boldsymbol{\Gamma} = \begin{bmatrix}\boldsymbol{\gamma}_1 & \boldsymbol{\gamma}_2 & \dots & \boldsymbol{\gamma}_T\end{bmatrix}$.

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

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

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

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

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 $\boldsymbol{\Sigma}_{\text{temporal}} \in \mathbb{R}^{T\times T}$ and $\boldsymbol{\Sigma}_{\text{spatial}} \in \mathbb{R}^{D\times D}$. 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 Spatio-temporal 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

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

We seek the posterior distribution of the parameters $\boldsymbol{X}, \boldsymbol{Z}$ and $\boldsymbol{\Gamma}$ conditioned on the observations $\boldsymbol{Y}$, which is obtained by applying Bayes's Theorem to the joint distribution in eq. (15)

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

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 $\mathrm{KL}\left(f_j(\boldsymbol{x}_j)\, Q^{\setminus j}(\boldsymbol{x}) \middle|\middle| \tilde{f}_j(\boldsymbol{x}_j)\, Q^{\setminus j}(\boldsymbol{x})\right)$. 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,

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

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

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

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

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

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 $f_1\left(\boldsymbol{X}\right)$ is decomposed into T terms of the form $f_{1,t}\left(\boldsymbol{x}_t\right) = p(\boldsymbol{y}_t \middle| \boldsymbol{x}_t)$, whereas the terms f_2 and f_3 are further decomposed as follows

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

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

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

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, $\tilde{f}_4(\boldsymbol{\Gamma}) = f_4(\boldsymbol{\Gamma}) = \mathcal{N}(\boldsymbol{\Gamma}|\mu_0, \boldsymbol{\Sigma}_0)$.

Define $\hat{\boldsymbol{m}}_{2,t} = \begin{bmatrix}\hat{m}_{2,t,1} & \hat{m}_{2,t,2} & \dots & \hat{m}_{2,t,D}\end{bmatrix}^T$, $\hat{\boldsymbol{V}}_{2,t} = \operatorname{diag}\begin{pmatrix}\hat{v}_{2,t,1} & \hat{v}_{2,t,2} & \dots & \hat{v}_{2,t,D}\end{pmatrix}^T$ 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

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

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

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

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

$$
\hat{\boldsymbol{\Sigma}} = \left[\boldsymbol{\Sigma}_0^{-1} + \hat{\boldsymbol{\Sigma}}_3^{-1}\right]^{-1}, \qquad (33)
$$

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

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

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 $\hat{\boldsymbol{m}}_{1,t}$, $\hat{\boldsymbol{V}}_{1,t}$, $\hat{\boldsymbol{m}}_{2,t}$, $\hat{\boldsymbol{V}}_{2,t}$, $\hat{\boldsymbol{\mu}}_{3,t}$, $\hat{\boldsymbol{\Sigma}}_{3,t}$, $\hat{\boldsymbol{\gamma}}_{2,t}$ and $\hat{\boldsymbol{\gamma}}_{3,t}$ 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.

- 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: $\mathrm{KL}\big(f_{1,n,t} Q^{\setminus 1,n,t} \big|\big| Q^{1,t,\mathrm{new}}\big)$ 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: $\mathrm{KL}\big(f_{2,i,t} Q^{\setminus 2,i,t} \big|\big| Q^{2,t,\mathrm{new}}\big)$ 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: $\mathrm{KL}\big(f_{3,i,t} Q^{\setminus 3,i,t} \big|\big| Q^{3,t,\mathrm{new}}\big)$ 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.

4.3 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

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

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.

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

The terms $\tilde{f}_{2,t} = \prod_{i=1}^{D} \tilde{f}_{2,i,t}(x_{i,t}, z_{i,t})$ 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 $\tilde{f}_{2,j,t}(x_{j,t}, z_{j,t})$. The first step is to compute the marginal cavity distributions by removing the contribution of $\tilde{f}_{2,j,t}(x_{j,t}, z_{j,t})$ from the marginal of the global approximation Q using eq. (20)

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

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

$$\hat{v}^{\backslash 2,j,t} = \left[\hat{V}_{t,jj}^{-1} - \hat{v}_{2,j,t}^{-1}\right]^{-1}, \tag{38}$$

$$\hat{m}^{\backslash 2,j,t} = \hat{v}^{\backslash 2,j,t}\left[\hat{V}_{t,jj}^{-1}\hat{m}_{j,t} - \hat{v}_{2,j,t}^{-1}\hat{m}_{2,j,t}\right], \tag{39}$$

$$\hat{\gamma}^{\backslash 2,j,t} = \hat{\gamma}_{3,j,t}. \tag{40}$$

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 $f_{2,j,t}Q^{\backslash 2,j,t}$ 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. $x_{j,t}$

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

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

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

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

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

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

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

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

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 $\mathcal{O}\left(N^2D\right)$. 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 $\mathcal{O}\left(TN^2D\right)$.

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 $\mathbf{\Gamma}$.

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 $\mathbf{\Sigma}_0$ with a low rank approximation plus a diagonal matrix $\mathbf{\Sigma}_0 \approx \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T + \mathbf{\Lambda}$, 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 $\mathbf{\Lambda}$ is chosen such that the diagonal in the exact prior covariance matrix $\mathbf{\Sigma}_0$ is preserved. This allows us to apply the matrix inversion lemma to compute the update of the posterior covariance matrix for $\mathbf{\Gamma}$ (see Section 4.6.1).

Computing the eigendecomposition of $\mathbf{\Sigma}_0 \in \mathbb{R}^{DT\times DT}$ scales in general as $\mathcal{O}\left(D^3T^3\right)$. However, when the prior covariance has Kronecker structure, the eigendecompositions of $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_t \otimes \mathbf{\Sigma}_s$ can be efficiently obtained from the eigendecompositions of $\mathbf{\Sigma}_t \in \mathbb{R}^{T\times T}$ and $\mathbf{\Sigma}_s \in \mathbb{R}^{D\times D}$. In this case, the eigendecomposition of $\mathbf{\Sigma}_0$ 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 $\mathbf{\Gamma}$ 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 $\mathbf{\Gamma}$ can be written in terms of Kronecker products, we can significantly speed up the computation of the posterior covariance matrix of $\mathbf{\Gamma}$ 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

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

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

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

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

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, $p(\boldsymbol{Y}\middle|\boldsymbol{\Omega}) \approx Q(\boldsymbol{Y}|\boldsymbol{\Omega})$. The approximation $Q(\boldsymbol{Y}|\boldsymbol{\Omega})$ 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

$$\ln Q(\boldsymbol{\Omega}|Y) \sim \mathcal{GP}\left(\mu\left(\boldsymbol{\Omega}\right), k\left(\boldsymbol{\Omega}, \boldsymbol{\Omega}'\right)\right), \tag{70}$$

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 $\boldsymbol{\Omega}$, 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 $\hat{\boldsymbol{X}} = \mathbb{E}_{Q(\boldsymbol{X}|\boldsymbol{Y})}\left[\boldsymbol{X}\right]$ 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 $\|\cdot\|_{\mathrm{F}}$ 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`.

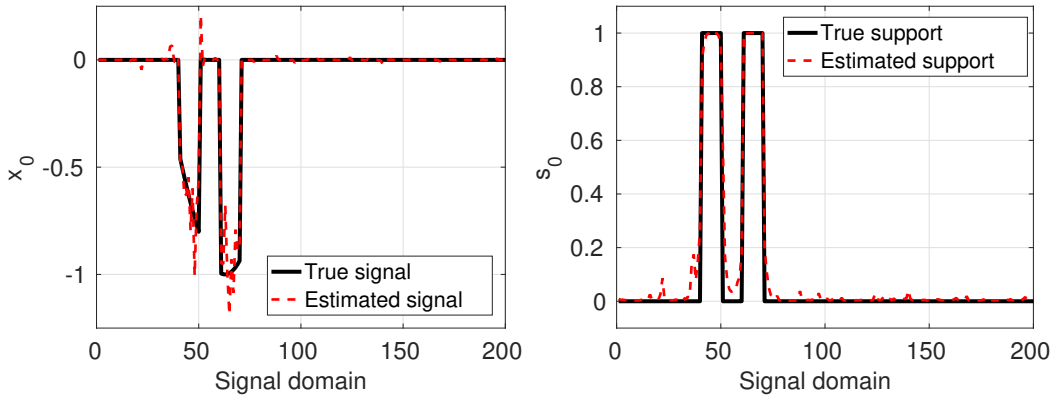
7.1 Experiment 1: The Effect of the Prior

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

$$\boldsymbol{\mu} = \nu \cdot \mathbf{1} \quad \text{and} \quad \boldsymbol{\Sigma}_{ij} = \kappa_1^2 \exp\left(-\frac{D_{ij}^2}{2\kappa_2^2}\right).$$

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 $\mathbb{E}_{Q(z_i|\boldsymbol{y})}[z_i]$.

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



(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.

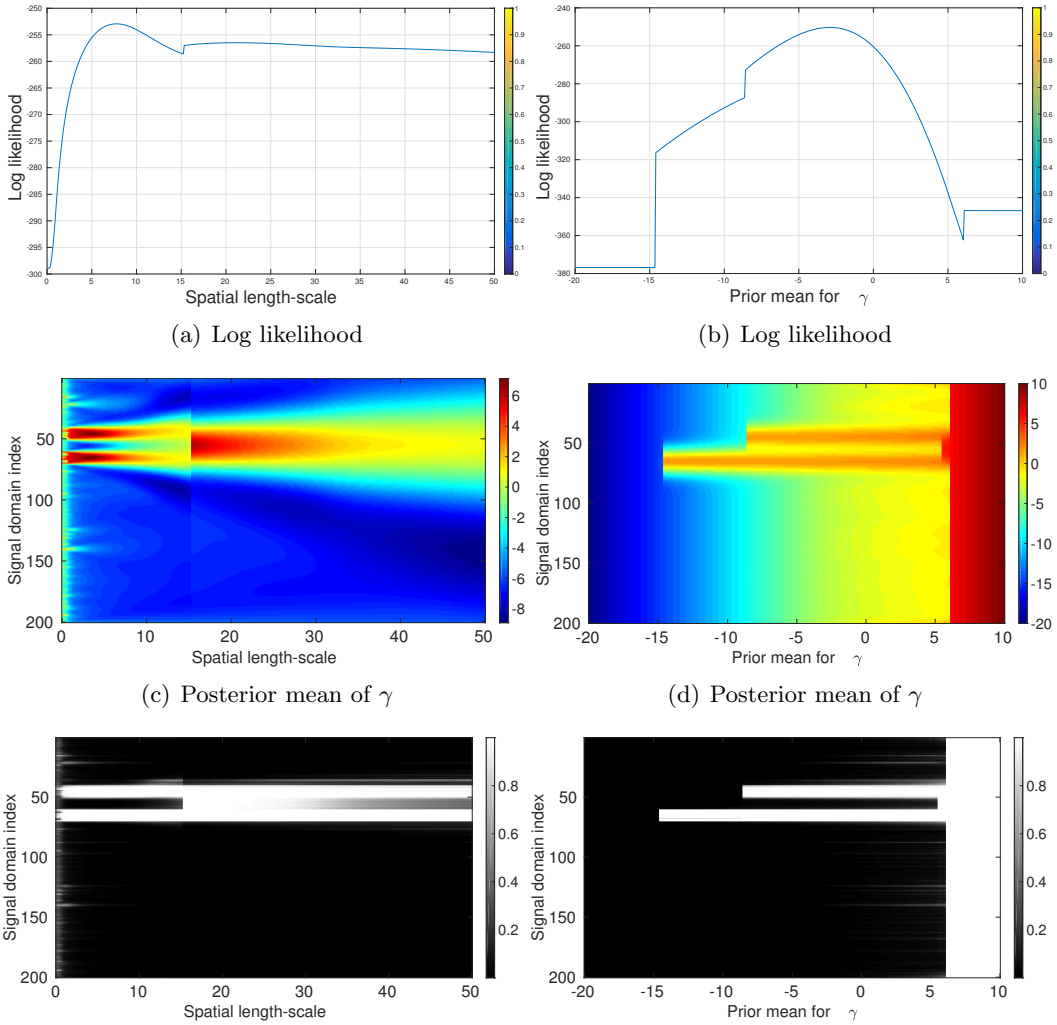
only present occasionally.

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.

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



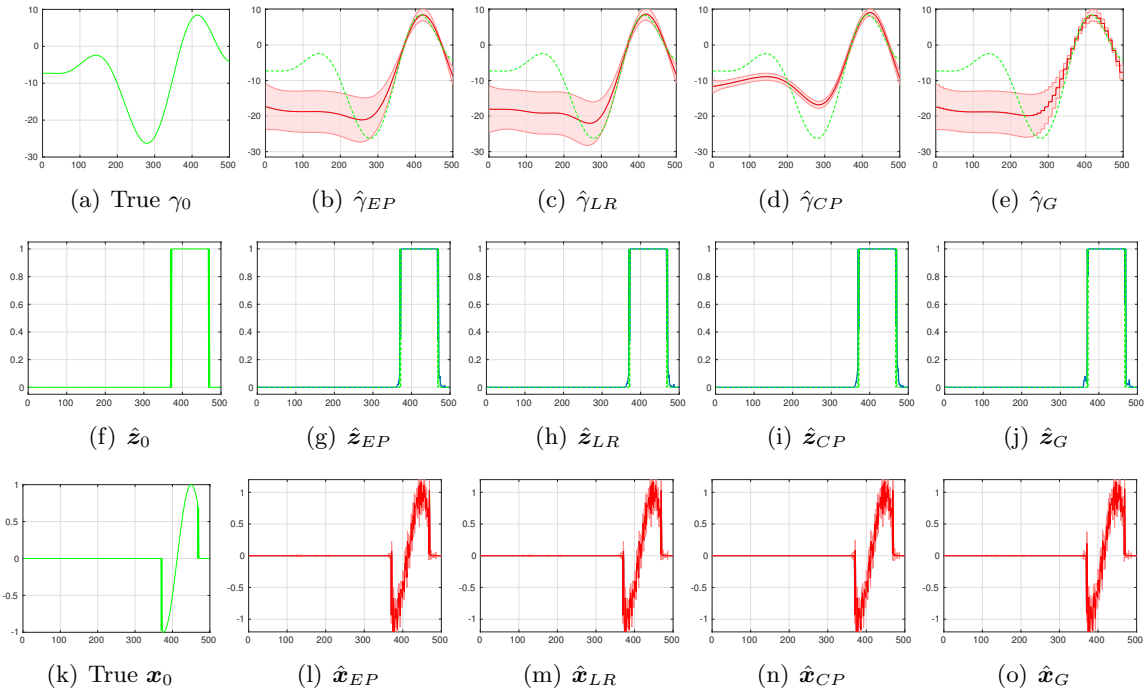
(a) Log likelihood (b) Log likelihood

(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 $\boldsymbol{\gamma}$.

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



(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.

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.

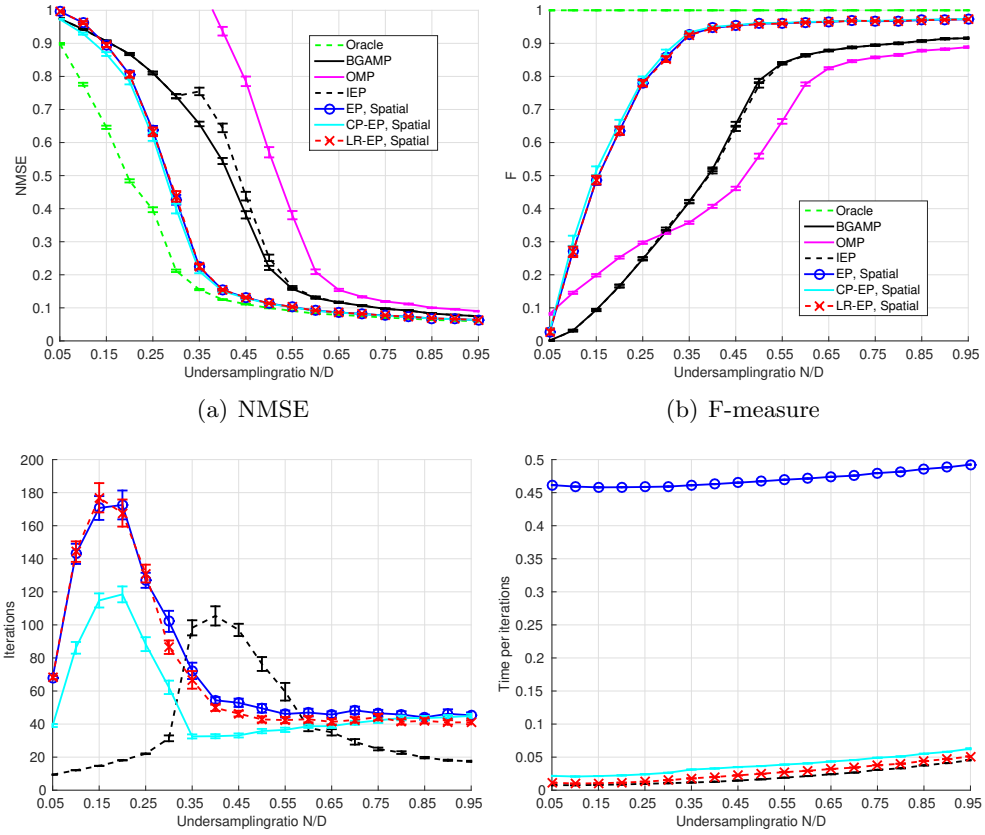
We can now compare the posterior distribution of $\boldsymbol{\gamma}_i$ for standard EP with the three approximations. Based on visual inspection one cannot tell the difference between the standard EP and EP with the low rank approximation, but the results for CP-EP and G-EP are quite different. For 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,



(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.

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 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.

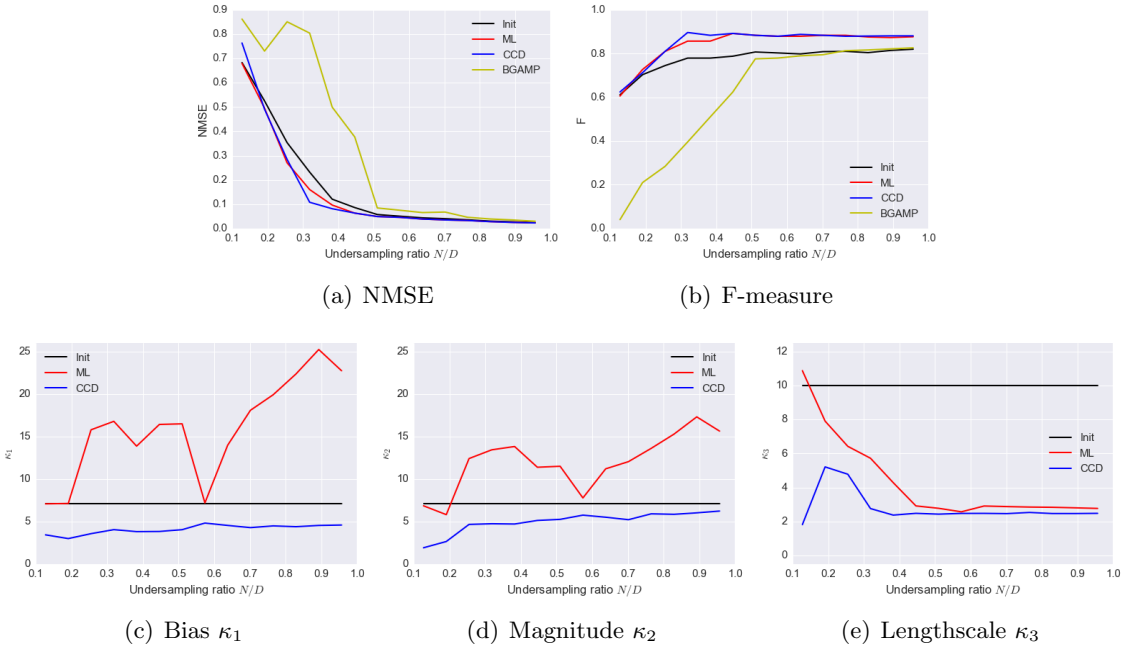
7.4 Experiment 4: Compressed Sensing of Optical Characters

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.



(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.

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 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).

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

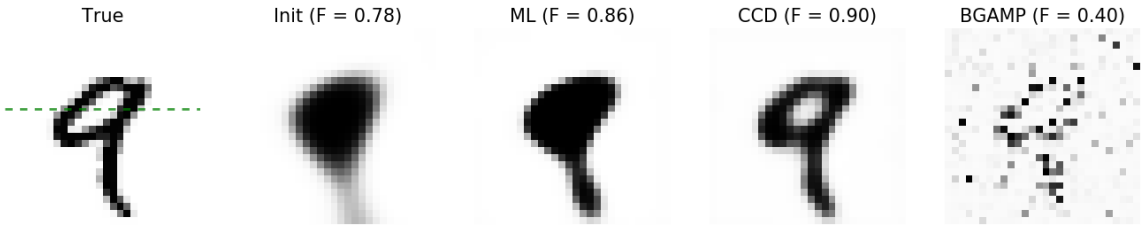
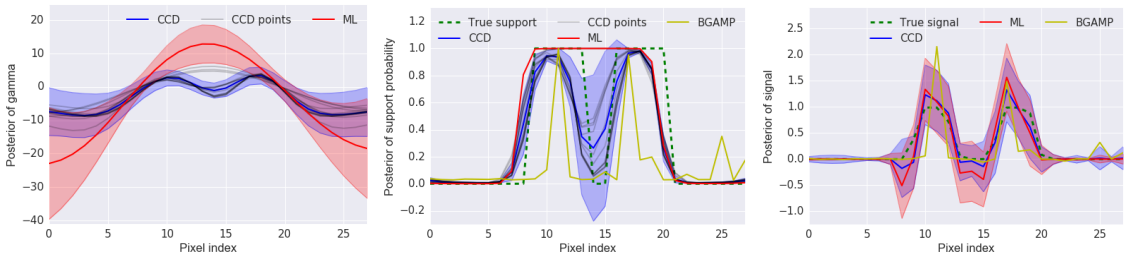
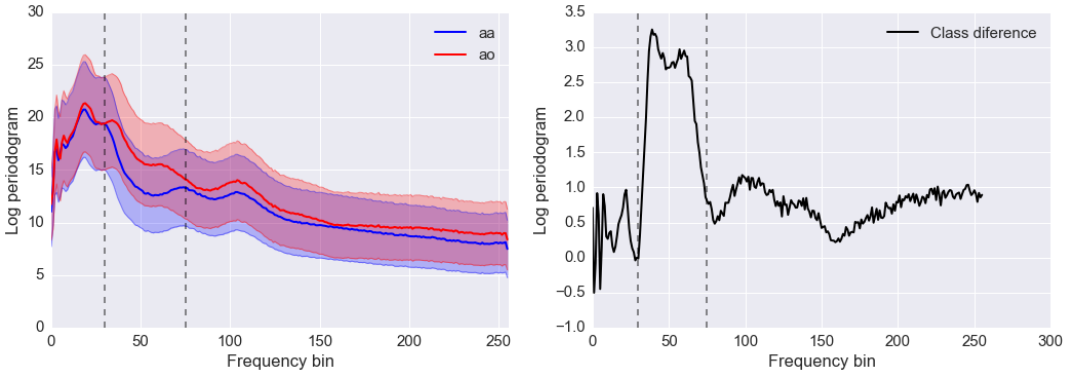


Figure 6: Posterior mean of $\boldsymbol{z}$ for compressed sensing of numerical digits, where $N/D \approx 0.3$. The panels in the top row show the posterior mean of the support, while the panels in the bottom row show the posterior mean of signal. Figure 7 shows the posterior distributions of the row indicated the dashed line.



(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 $N/D \approx 0.3$.



(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.

Method	Training error	Test error	Train LPPD	Test LPPD
NBSBC	9.7 (±0.3)	19.5 (±0.1)	−42.7 (±0.8)	−698.6 (±3.3)
LR-EP (ML)	13.3 (±0.3)	19.4 (±0.1)	−50.6 (±0.8)	−673.8 (±2.2)
LR-EP (CCD)	13.4 (±0.3)	**19.2** (±0.1)	−50.7 (±0.8)	**−665.5** (±1.5)

Table 2: Results for phoneme classification experiment

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

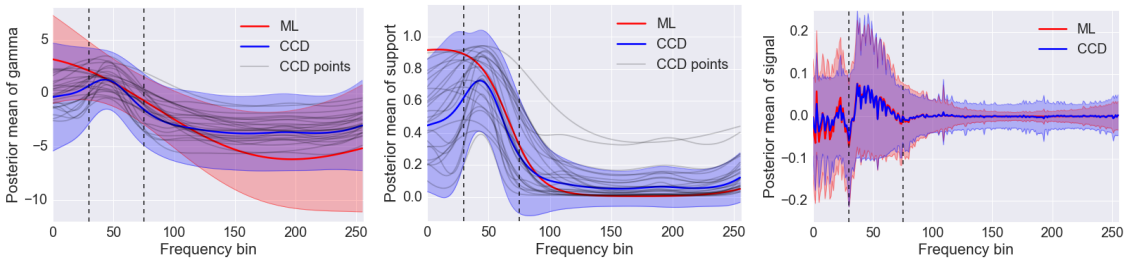
panel in Figure 2 shows the frequency-wise mean and standard deviation of the two classes and the right-most panel shows the difference of the two mean signals. We choose a squared exponential kernel for $\boldsymbol{\gamma}$ since it is assumed that frequency bands rather than single frequencies are 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.

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



(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 hyperparameter 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.

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

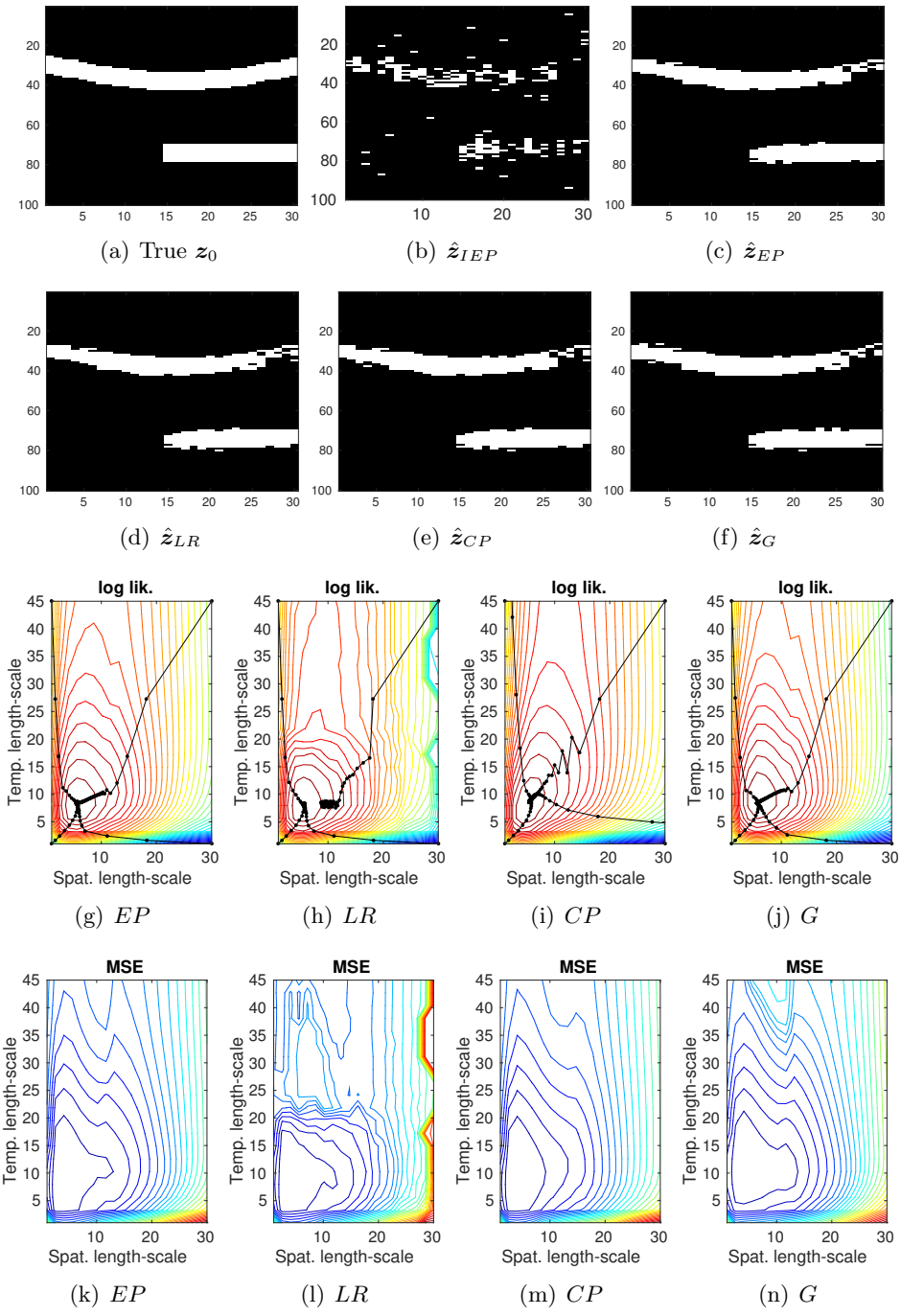
structure. Furthermore, the reconstructions using the approximation schemes do not differ significantly from the result using full EP.

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 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 $\boldsymbol{\Gamma}$ 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.



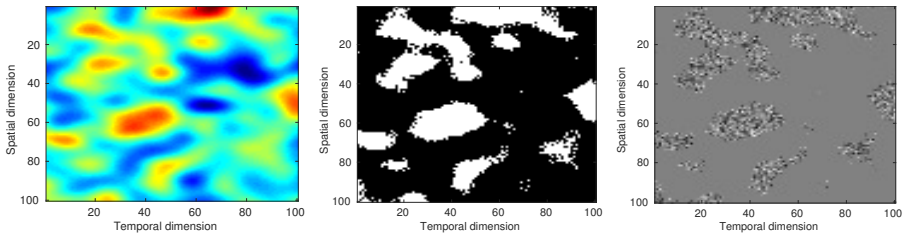
(a) True $\boldsymbol{z}_0$ (b) $\hat{\boldsymbol{z}}_{IEP}$ (c) $\hat{\boldsymbol{z}}_{EP}$

(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 $\boldsymbol{X}$

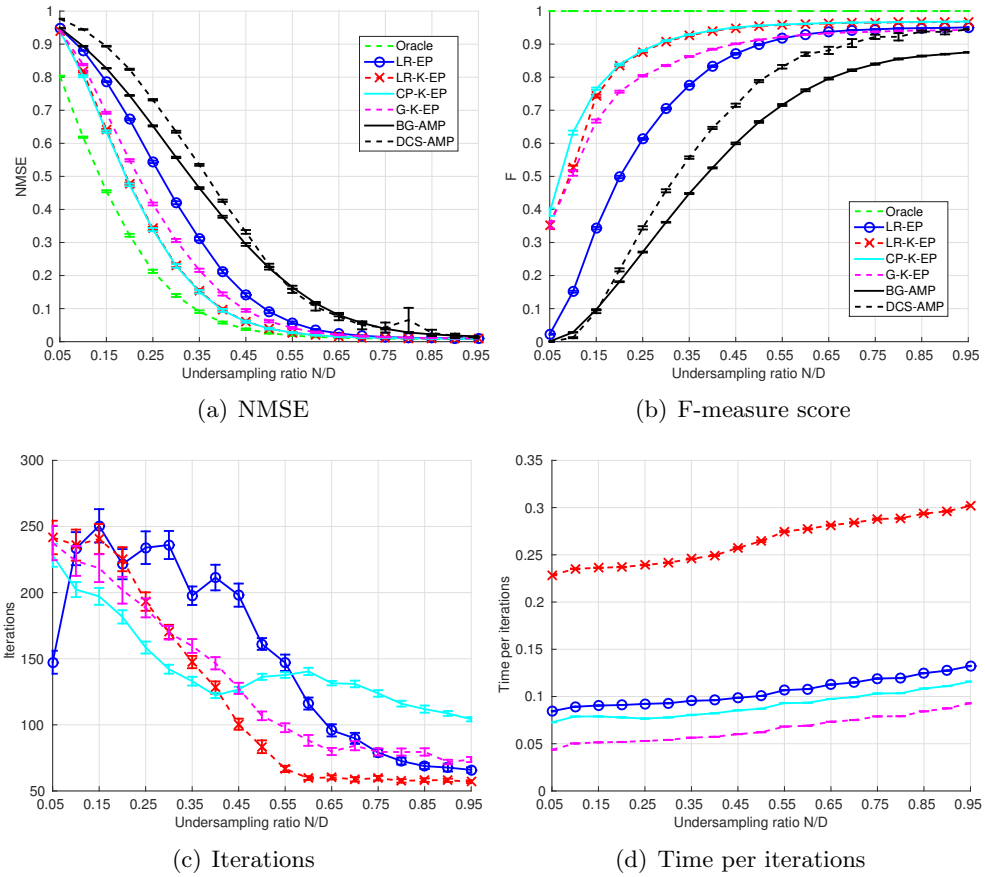
Figure 11: Example of a realization of the synthetic signals in Experiment 6.

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.

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 evenly distributed between the two conditions: face or scrambled face. Each epoch has a duration of $800ms$ corresponding to $T = 161$ samples in time and the stimuli are presented at $t = 0s$. We generated a forward model[1] with 5124 dipole sources, that is $\boldsymbol{A} \in \mathbb{R}^{128\times 5124}$. To encourage spatio-temporal coherence of the sources, we choose the covariance matrix for $\mathbf{\Gamma}$ to be of the form $\mathbf{\Sigma}_0 = \kappa^2 \cdot \mathbf{\Sigma}_{\text{temporal}} \otimes \mathbf{\Sigma}_{\text{spatial}}$, 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

1. We used the SPM8 software (Ashburner et al., 2010).



(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.

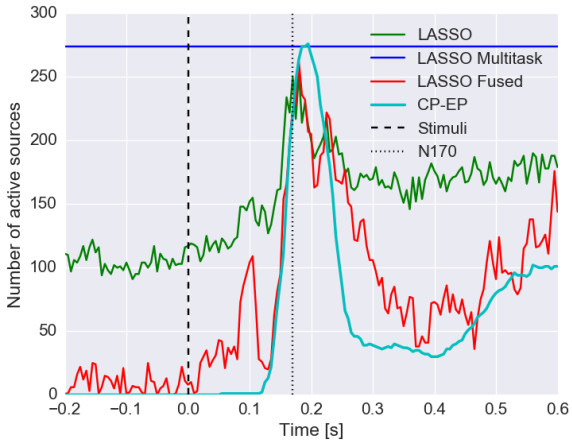


Figure 13: The number of active dipoles sources as a function of the time for the common precision approximation (CP-EP) and the LASSO, the Multi-task LASSO and the fused LASSO for a face perception experiment. The stimuli were presented at time $t = 0$.

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 of this problem we use the common precision approximation for this data set. In fact, a low rank approximation of the prior covariance matrix $\boldsymbol{\Sigma}_0$ 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 to 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

2. We used the implementation in scikit-learn toolbox (Pedregosa et al., 2011).

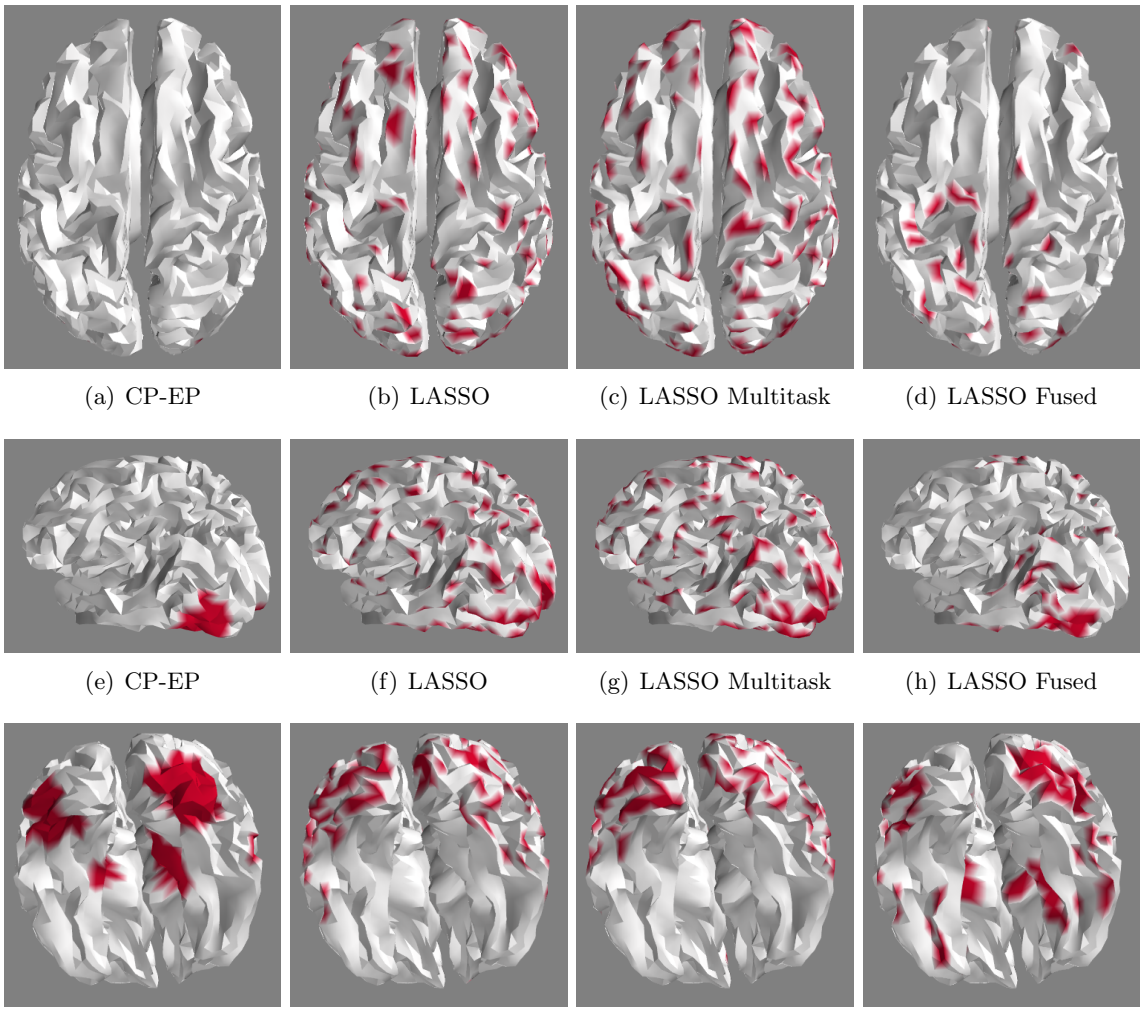
the fused LASSO[3] (Tibshirani et al., 2005). The LASSO, the multi-task LASSO and the fused LASSO all minimize a quadratic reconstruction error subject to an ℓ_1 constraint, but the multi-task LASSO also assumes that the sparsity pattern is constant in time (joint sparsity) and the fused LASSO has an additional constraint on the temporal first-order difference of the solution $\sum_{i,t} |x_{i,t} - x_{i,t-1}|$.

The CP-EP method has been informed with appropriate values of the kernel hyperparameters. This is possible because the structure of the sparsity pattern is encoded using generic covariance functions that are easily interpretable. However, this is not an option for the LASSSO methods as the regularization parameters of these methods are more difficult to interpret in terms of the geometry of the problem. To compensate, we used the estimated solution obtained using the CP-EP method to inform the LASSO methods as follows. For the regular LASSO and the multi-task LASSO, we chose the value of the regularization parameter such that the number of active sources matches the CP-EP solution at the time point with the largest number of active sources (see Figure 13). For the fused LASSO, we also matched the average autocorrelation across all sources.

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 consistent with the known time delay of approximately 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. We note that learning hyperparameters is often difficult in neuroimaging due to high-dimensional signals and poor signal to noise conditions (Varoquaux et al., 2017).

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 $Q\left(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}\right)$

$$\begin{aligned} Q^{\backslash 2,t,j}(x_{t,j}, z_{t,j}) &= \frac{Q\left(x_{t,j}, z_{t,j}\right)}{\tilde{f}_{2,t,j}\left(x_{t,j}, z_{t,j}\right)} = \frac{\mathcal{N}\left(x_{t,j} | \hat{m}_{t,j}, \hat{V}_{t,j}\right) \mathrm{Ber}\left(z_{t,j} | \phi\left(\hat{\gamma}_{t,j}\right)\right)}{\mathcal{N}\left(x_{t,j} | \hat{m}_{2,t,j}, \hat{V}_{2,t,j}\right) \mathrm{Ber}\left(z_{t,j} | \phi\left(\hat{\gamma}_{2,t,j}\right)\right)} \\ &= K^{\backslash 2,t,j} \cdot \mathcal{N}\left(x_{t,j} | \hat{m}^{\backslash 2,t,j}, \hat{V}^{\backslash 2,t,j}\right) \mathrm{Ber}\left(z_{t,j} | \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right), \end{aligned} \tag{74}$$

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. $z_{t,j}$

$$\begin{aligned} Z_1 &= \sum_{z_{t,j}} \int z_{t,j} f_{2,t,j}\left(x_{t,j}, z_{t,j}\right) Q^{\backslash 2,t,j}\left(x_{t,j}, z_{t,j}\right) \mathrm{d}x_{t,j} \\ &= \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right) \mathcal{N}\left(0 | \hat{m}^{\backslash 2,i} - \rho_0, \hat{V}^{\backslash 2,t,j} + \tau_0\right). \end{aligned} \tag{78}$$

Next, the zeroth moment w.r.t $x_{t,i}$ or the normalization constant of $f_{2,t,j}Q^{\backslash 2,t,j}$

$$\begin{aligned} X_0 &= \sum_{z_{t,j}} \int f_{2,t,j}\left(x_{t,j}, z_{t,j}\right) Q^{\backslash 2,t,j}\left(x_{t,j}, z_{t,j}\right) \mathrm{d}x_{t,j} \qquad (79) \\ &= \sum_{z_{t,j}} \int \left[\left(1 - z_{t,j}\right)\delta(x_{t,j}) + z_{t,j}\mathcal{N}\left(x_{t,j}\middle|\rho_0, \tau_0\right)\right] \\ &\quad \mathcal{N}\left(x_{t,j}\middle|\hat{m}^{\backslash 2,t,j}, \hat{V}^{\backslash 2,t,j}\right) \mathrm{Ber}\left(z_{t,j}\middle|\phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right) \mathrm{d}x_{t,j} \\ &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right)\mathcal{N}\left(0\middle|\hat{m}^{\backslash 2,i}, \hat{V}^{\backslash 2,i}\right) + \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\mathcal{N}\left(0\middle|\hat{m}^{\backslash 2,t,j} - \rho_0, \hat{V}^{\backslash 2,t,j} + \tau_0\right) \\ &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\right)\mathcal{N}\left(0\middle|\hat{m}^{\backslash 2,i}, \hat{V}^{\backslash 2,i}\right) + Z_1 \qquad (80) \end{aligned}$$

We now compute the (unnormalized) first moment w.r.t. $x_{t,j}$

$$\begin{aligned} X_1 &= \sum_{z_{t,j}} \int x_{t,j} f_{2,t,j}\left(x_{t,j}, z_{t,j}\right) Q^{\backslash 2,i}\left(x_{t,j}, z_{t,j}\right) \mathrm{d}x_{t,j} \\ &= \phi\left(\hat{\gamma}^{\backslash 2,t,j}\right)\mathcal{N}\left(0\middle|\hat{m}^{\backslash 2,t,j} - \rho_0, \hat{V}^{\backslash 2,t,j} + \tau_0\right) \frac{\frac{\hat{m}^{\backslash 2,t,j}}{\hat{V}^{\backslash 2,t,j}} + \frac{\rho_0}{\tau_0}}{\frac{1}{\tau_0} + \frac{1}{\hat{V}^{\backslash 2,t,j}}} \\ &= Z_1 \frac{\hat{m}^{\backslash 2,t,j}\tau_0 + \rho_0 \hat{V}^{\backslash 2,t,j}}{\tau_0 + \hat{V}^{\backslash 2,t,j}} \qquad (81) \end{aligned}$$

and the second (unnormalized) moment w.r.t. $x_{t,j}$

$$\begin{aligned} X_2 &= \sum_{z_{t,j}} \int x_i^2 f_{2,i}\left(x_i, z_i\right) Q^{\backslash 2,i}\left(x_i, z_i\right) \mathrm{d}x_i \\ &= \phi\left(\hat{\gamma}^{\backslash 2,i}\right)\mathcal{N}\left(0\middle|\hat{m}^{\backslash 2,i} - \rho_0, \hat{V}^{\backslash 2,i} + \tau_0\right)\left[\left(\frac{\frac{\hat{m}^{\backslash 2,i}}{\hat{V}^{\backslash 2,i}} + \frac{\rho_0}{\tau_0}}{\frac{1}{\tau_0} + \frac{1}{\hat{V}^{\backslash 2,i}}}\right)^2 + \frac{1}{\frac{1}{\tau_0} + \frac{1}{\hat{V}^{\backslash 2,i}}}\right] \\ &= Z_1\left[\left(\frac{\hat{m}^{\backslash 2,t,j}\tau_0 + \rho_0\hat{V}^{\backslash 2,t,j}}{\tau_0 + \hat{V}^{\backslash 2,t,j}}\right)^2 + \frac{\tau_0\hat{V}^{\backslash 2,i}}{\hat{V}^{\backslash 2,i} + \tau_0}\right] \qquad (82) \end{aligned}$$

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

$$E\left[x_{t,j}\right] = \frac{X_1}{X_0}, \qquad V\left[x_{t,j}\right] = \frac{X_2}{X_0} - \frac{X_1^2}{X_0^2}, \qquad E\left[z_{t,j}\right] = \frac{Z}{X_0}. \qquad (83)$$

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

$$Q^{\setminus 3,t,j}(z_{t,j}, \gamma_{t,j}) = \frac{Q\left(z_{t,j}, \gamma_{t,j}\right)}{\tilde{f}_{3,t,j}\left(z_{t,j}, \gamma_{t,j}\right)} = \frac{\text{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}_{t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j}, \hat{\mu}_{t,j}, \hat{\Sigma}_{t,jj}\right)}{\text{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}_{3,t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j}, \hat{\mu}_{3,t,j}, \hat{\Sigma}_{3,t,j}\right)}$$
$$= K^{\setminus 3,t,j} \cdot \text{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}^{\setminus 3,t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j} \middle| \hat{\mu}^{\setminus 3,t,j}, \hat{\Sigma}^{\setminus 3,t,j}\right), \tag{84}$$

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 $f_{3,t,j}Q^{\setminus 3,t,j}$

$$\begin{aligned}
G_0 &= \sum_{z_{t,j}} \int f_{3,t,j}\left(z_{t,j}, \gamma_{t,j}\right) Q^{\setminus 3,t,j}\left(z_{t,j}, \gamma_i\right) \mathrm{d}\gamma_{t,j} \\
&= \sum_{z_{t,j}} \int \text{Ber}\left(z_{t,j} \middle| \phi\left(\gamma_{t,j}\right)\right) \text{Ber}\left(z_{t,j} \middle| \phi\left(\hat{\gamma}^{\setminus 3,t,j}\right)\right) \mathcal{N}\left(\gamma_{t,j} \middle| \hat{\mu}^{\setminus 3,t,j}, \hat{\Sigma}^{\setminus 3,t,j}\right) \mathrm{d}\gamma_{t,j} \\
&= \sum_{z_i} \int \left[\left(1 - z_i\right)\left(1 - \phi\left(\gamma_i\right)\right)\left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) + z_i \phi\left(\gamma_i\right) \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right] \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\
&= \left(1 - \phi\left(\hat{\gamma}^{\setminus 3,i}\right)\right) \int \left(1 - \phi\left(\gamma_i\right)\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i \\
&\quad + \phi\left(\hat{\gamma}^{\setminus 3,i}\right) \int \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\setminus 3,i}, \hat{\Sigma}^{\setminus 3,i}\right) \mathrm{d}\gamma_i
\end{aligned}$$

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

$$\begin{aligned} Z_1 &= \sum_{z_i} \int z_i f_{3,i}\left(z_i, \gamma_i\right) Q^{\backslash 3,i}\left(z_i, \gamma_i\right) \mathrm{d}\gamma_i \\ &= \phi\left(\hat{\gamma}^{\backslash 3,i}\right) \phi\left(c_{3,i}\right), \end{aligned} \tag{92}$$

Then the first moment w.r.t. to $z_{i,t}$ is obtained as $E\left[z_{i,t}\right] = Z_1/G_0$.

For the moments w.r.t. γ_i, we get

$$\begin{aligned} G_1 &= \sum_{z_i} \int \gamma_i f_{3,i}\left(z_i, \gamma_i\right) Q^{\backslash 3,i}\left(z_i, \gamma_i\right) \mathrm{d}\gamma_i \\ &= \sum_{z_i} \int \gamma_i \left[\left(1 - z_i\right)\left(1 - \phi\left(\gamma_i\right)\right)\left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) + z_i \phi\left(\gamma_i\right) \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right] \\ &\quad \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i \\ &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \int \gamma_i \left(1 - \phi\left(\gamma_i\right)\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i \\ &\quad + \phi\left(\hat{\gamma}^{\backslash 3,i}\right) \int \gamma_i \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i \\ &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \left[\hat{\mu}^{\backslash 3,i} - \int \gamma_i \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i\right] \\ &\quad + \phi\left(\hat{\gamma}^{\backslash 3,i}\right) \int \gamma_i \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i \end{aligned} \tag{93}$$

Again we turn to (Rasmussen and Williams, 2006) for the analytical solution of the above integrals

$$\begin{aligned} \int \gamma_i \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i &= \phi\left(c_{3,i}\right) \hat{\mu}^{\backslash 3,i} + \phi\left(c_{3,i}\right) \frac{\hat{\Sigma}^{\backslash 3,i} \mathcal{N}\left(c_{3,i} \middle| 0, 1\right)}{\phi\left(c_{3,i}\right) \sqrt{1 + \hat{\Sigma}^{\backslash 3,i}}} \\ &= \phi\left(c_{3,i}\right) \hat{\mu}^{\backslash 3,i} + \phi\left(c_{3,i}\right) d_{3,i}, \end{aligned} \tag{94}$$

where we have defined

$$d_{3,i} \triangleq \frac{\hat{\Sigma}^{\backslash 3,i} \mathcal{N}\left(c_{3,i} \middle| 0, 1\right)}{\phi\left(c_{3,i}\right) \sqrt{1 + \hat{\Sigma}^{\backslash 3,i}}}. \tag{95}$$

Plugging eq. (94) back into eq. (93) and simplifying yields

$$\begin{aligned} G_1 &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \left[\left(1 - \phi\left(c_{3,i}\right)\right) \hat{\mu}^{\backslash 3,i} - \phi\left(c_{3,i}\right) d_{3,i}\right] + \phi\left(\hat{\gamma}^{\backslash 3,i}\right) \phi\left(c_{3,i}\right) \left[\hat{\mu}^{\backslash 3,i} + d_{3,i}\right] \\ &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \left(1 - \phi\left(c_{3,i}\right)\right) \hat{\mu}^{\backslash 3,i} - \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \phi\left(c_{3,i}\right) d_{3,i} + Z_1 \left[\hat{\mu}^{\backslash 3,i} + d_{3,i}\right] \\ &= \left(G_0 - Z_1\right) \hat{\mu}^{\backslash 3,i} - \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \phi\left(c_{3,i}\right) d_{3,i} + Z_1 \left[\hat{\mu}^{\backslash 3,i} + d_{3,i}\right] \\ &= G_0 \hat{\mu}^{\backslash 3,i} + \left(2 Z_1 - \phi\left(c_{3,i}\right)\right) d_{3,i} \end{aligned} \tag{96}$$

Thus, the first moment w.r.t. $\gamma_{i,t}$ is given by $\mathbb{E}\left[\gamma_{i,t}\right] = G1/G0$.

Similarly, we compute the second moment w.r.t. γ_i

$$\begin{aligned} G_2 &= \sum_{z_i} \int \gamma_i^2 f_{3,i}\left(z_i, \gamma_i\right) Q^{\backslash 3,i}\left(z_i, \gamma_i\right) \mathrm{d}\gamma_i \\ &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \int \gamma_i^2 \left(1 - \phi\left(\gamma_i\right)\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i \\ &\quad + \phi\left(\hat{\gamma}^{\backslash 3,i}\right) \int \gamma_i^2 \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i \end{aligned} \tag{97}$$

The solution to the above integrals are given by (Rasmussen and Williams, 2006)

$$\begin{aligned} &\int \gamma_i^2 \phi\left(\gamma_i\right) \mathcal{N}\left(\gamma_i \middle| \hat{\mu}^{\backslash 3,i}, \hat{\Sigma}^{\backslash 3,i}\right) \mathrm{d}\gamma_i \\ &= \phi\left(c_{3,i}\right) \left[2\hat{\mu}^{\backslash 3,i}\left(\hat{\mu}^{\backslash 3,i} + d_{3,i}\right) + \left(\hat{\Sigma}^{\backslash 3,i} - \left(\hat{\mu}^{\backslash 3,i}\right)^2\right) - b_{3,i}\right] \end{aligned} \tag{98}$$

where

$$b_{3,i} \triangleq \frac{\left(\hat{\Sigma}^{\backslash 3,i}\right)^2 c_{3,i} \mathcal{N}\left(c_{3,i} \middle| 0, 1\right)}{\phi\left(c_{3,i}\right)\left(1 + \hat{\Sigma}^{\backslash 3,i}\right)} \tag{99}$$

Furthermore, we define

$$w_{3,i} \triangleq 2\hat{\mu}^{\backslash 3,i}\left(\hat{\mu}^{\backslash 3,i} + d_{3,i}\right) + \left(\hat{\Sigma}^{\backslash 3,i} - \left(\hat{\mu}^{\backslash 3,i}\right)^2\right) - b_{3,i} \tag{100}$$

Substituting the above result back into eq. (97) and rearranging yields

$$\begin{aligned} G_2 &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \left[\left(\hat{\mu}^{\backslash 3,i}\right)^2 + \hat{\Sigma}^{\backslash 3,i} - \phi\left(c_{3,i}\right) w_{3,i}\right] + \phi\left(\hat{\gamma}^{\backslash 3,i}\right) \phi\left(c_{3,i}\right) w_{3,i} \\ &= \left(1 - \phi\left(\hat{\gamma}^{\backslash 3,i}\right)\right) \left[\left(\hat{\mu}^{\backslash 3,i}\right)^2 + \hat{\Sigma}^{\backslash 3,i} - \phi\left(c_{3,i}\right) w_{3,i}\right] + Z_1 w_{3,i} \end{aligned} \tag{101}$$

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

$$\mathbb{E}\left[\gamma_{j,t}\right] = \frac{G_1}{G_0}, \qquad \mathbb{V}\left[\gamma_{j,t}\right] = \frac{G_2}{G_0} - \frac{G_1^2}{G_0^2}, \qquad \mathbb{E}\left[z_{j,t}\right] = \frac{Z_1}{G_0}. \tag{102}$$

These moments completely determine the distribution $Q^{3,\text{new}}$ and thus, we compute the updates for $f_{3,i}$ as follows

$$\hat{\Sigma}_{3,i}^{\text{new}} = \left[\mathbb{V}\left[\gamma_i\right]^{-1} - \left(\hat{\Sigma}^{\backslash 3,i}\right)^{-1}\right]^{-1}, \tag{103}$$

$$\hat{\mu}_{3,i}^{\text{new}} = \hat{\Sigma}_{3,i}^{\text{new}} \left[\mathbb{V}\left[\gamma_i\right]^{-1} \mathbb{E}\left[\gamma_i\right] - \left(\hat{\Sigma}^{\backslash 3,i}\right)^{-1} \hat{\mu}^{\backslash 3,i}\right], \tag{104}$$

$$\hat{\gamma}_{3,i}^{\text{new}} = d\left(\phi\left(\mathbb{E}\left[z_i\right]\right), \hat{\gamma}^{\backslash 3,i}\right), \tag{105}$$

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

$$\begin{aligned}\mathbb{V}\left[\boldsymbol{x}_t\right] = &\operatorname{diag}\left(\boldsymbol{V}^{\setminus 1,t,n}\right) \\ &- \alpha_{1,t,n} \cdot \frac{\left(\tilde{\boldsymbol{A}}_{n,\cdot}\mathbb{E}\left[\boldsymbol{x}_t\right] + \alpha_{1,t,n}\right)}{1+b_{1,t,n}} \left(\tilde{\boldsymbol{A}}_{n,\cdot} \circ \operatorname{diag}\left(\boldsymbol{V}^{\setminus 1,t,n}\right)\right) \circ \left(\tilde{\boldsymbol{A}}_{n,\cdot} \circ \operatorname{diag}\left(\boldsymbol{V}^{\setminus 1,t,n}\right)\right).\end{aligned} \tag{118}$$

Using these moments, we compute the updates for $\tilde{f}_{1,t,n}$ as follows

$$\hat{\boldsymbol{V}}_{1,t,n}^{\text{new}} = \left[\boldsymbol{V}\left[\boldsymbol{x}_t\right]^{-1} - \left(\boldsymbol{V}^{\setminus 1,t,n}\right)^{-1}\right]^{-1}, \tag{119}$$

$$\hat{\boldsymbol{m}}_{1,t,n}^{\text{new}} = \hat{\boldsymbol{V}}_{1,t,n}^{\text{new}}\left[\boldsymbol{V}\left[\boldsymbol{x}_t\right]^{-1}\mathbb{E}\left[\boldsymbol{x}_t\right] - \left(\boldsymbol{V}^{\setminus 1,t,n}\right)^{-1}\boldsymbol{m}^{\setminus 1,t,n}\right]. \tag{120}$$

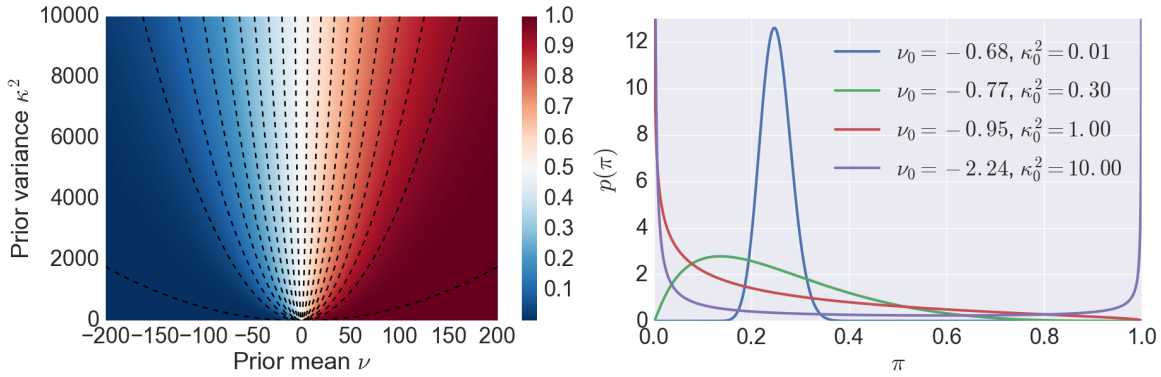
Appendix D. On the Prior Mean and Variance of Γ

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

$$p(\pi) = p_\gamma(\phi^{-1}(\pi))\left|\frac{\mathrm{d}\phi^{-1}(\pi)}{\mathrm{d}\pi}\right| = \mathcal{N}\left(\phi^{-1}(\pi)|\nu_0, \kappa_0^2\right)\left|\frac{\mathrm{d}\phi^{-1}(\pi)}{\mathrm{d}\pi}\right| = \frac{\mathcal{N}\left(\phi^{-1}(\pi)|\nu_0, \kappa_0^2\right)}{\mathcal{N}\left(\phi^{-1}(\pi)|0, 1\right)}. \tag{122}$$

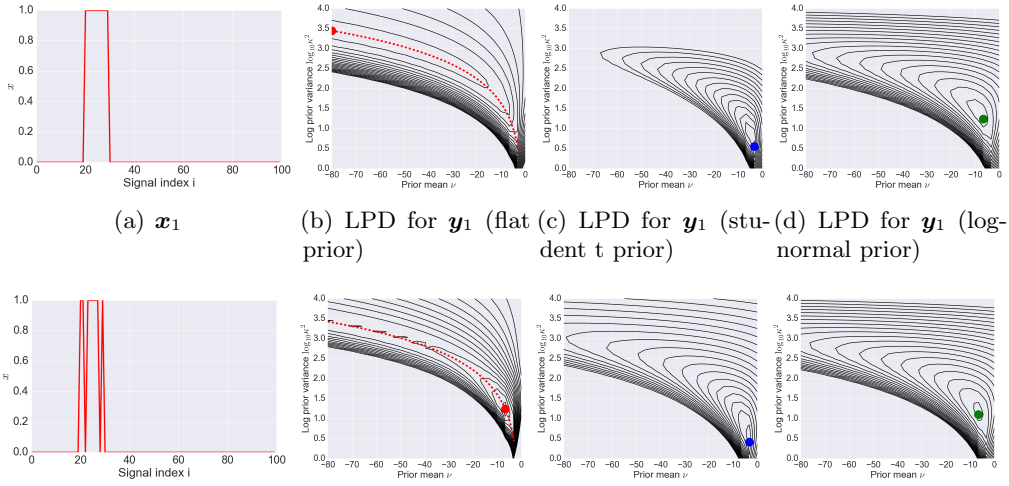


(a) Expected sparsity vs. prior mean and variance

(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)

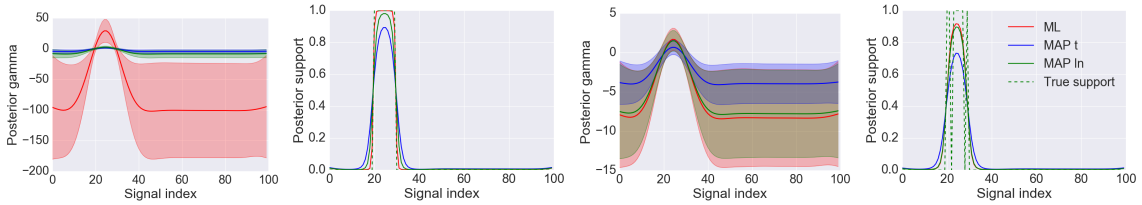
Figure 16: (a) Signal, where the support is contiguous. (b), (c), (d): Log posterior density for $\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{\epsilon}$ with a flat prior, half student t prior (df = 4) and a log normal prior (mean 6, std. dev 3) for κ_0, respectively. (e) Signal, 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.