Structured Sparse Modelling with Hierarchical GP

Danil Kuzin, Olga Isupova, Lyudmila Mihaylova

The University of Sheffield, Sheffield, UK

Email: dkuzin1@sheffield.ac.uk, oisupova1@sheffield.ac.uk, l.s.mihaylova@sheffield.ac.uk

Abstract—In this paper a new Bayesian model for sparse linear regression with a spatio-temporal structure is proposed. It incorporates the structural assumptions based on a hierarchical Gaussian process prior for spike and slab coefficients. We design an inference algorithm based on Expectation Propagation and evaluate the model over the real data.

I. INTRODUCTION

Sparse regression problems arise often in various applications, e.g., model selection, compressive sensing, EEG source localisation and gene modelling [1], [2]. One of the Bayesian approaches to force the coefficients being zeros is the spike and slab prior [3]: each component is modelled as a mixture of spike, that is the delta-function in zero, and slab, that is some vague distribution. Following the Bayesian approach, latent variables that are indicators of spikes are added to the model [4] and the relevant distribution is placed over them [5].

In this model each component is modelled to be spike or slab independently. However, in many applications non-zero elements tend to appear in groups forming an unknown structure: wavelet coefficients of images are usually organised in trees [6], chromosomes have a spatial structure along the genome [2].

We propose an extension of the spike and slab model by imposing a hierarchical Gaussian process (GP) prior on the latent variables. Such hierarchical prior allows to model spatial structural dependencies for coefficients that can evolve in time. The new model is flexible as spatial and temporal dependencies are decoupled by different levels of the hierarchical GP prior.

II. PROPOSED MODEL

The observations $\mathbf{y}_t \in \mathbb{R}^K$ are collected with the design matrix $\mathbf{X} \in \mathbb{R}^{K \times N}$ from the unknown coefficients $\boldsymbol{\beta}_t \in \mathbb{R}^N$ at every time moment $t \in [1, \dots, T]$, with independent noise:

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{y}_t | \mathbf{X}\boldsymbol{\beta}_t, \sigma^2 \mathbf{I}). \quad (1)$$

We consider the case when $K < N$, therefore the problem of recovery of $\boldsymbol{\beta}_t$ from $\mathbf{y}_t$ is ill-posed and regularisation is required.

a) Sparsity: The vectors $\boldsymbol{\beta}_t$ are assumed to be sparse, that is implemented in the model using the spike and slab approach:

$$\beta_{it} \sim \omega_{it}\delta_0(\beta_{it}) + (1 - \omega_{it})\mathcal{N}(\beta_{it}|0, \sigma_\beta^2), \quad (2)$$

where ω_{it} are the latent indicators of spike and slab.

b) Spatial clustering: Non-zero elements in $\boldsymbol{\beta}_t$ are assumed to be clustered in groups at every timestamp. Therefore spatial dependencies for the positions of spikes in β_{it} are modelled with the GP:

$$\omega_{it} \sim \text{Ber}(\omega_{it}|\Phi(\gamma_{it})), \quad \Phi(\cdot) \text{ is the standard Gaussian cdf} \quad (3)$$

$$\boldsymbol{\gamma}_t \sim \mathcal{N}(\boldsymbol{\gamma}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\Sigma}_0(i,j) = \alpha_\Sigma \exp\left(-\frac{(i-j)^2}{2l_\Sigma^2}\right). \quad (4)$$

GPs specify prior over an unknown structure. This is particularly useful as it allows to avoid a specification of any structural patterns — structural modelling is governed only by the GP covariance function.

c) Temporal evolution: Clusters of spikes in $\boldsymbol{\beta}_t$ are assumed to evolve in time. This evolution is addressed with the hierarchical GP dynamic system model [7]. The mean for the spatial GP changes over time according to the top-level temporal GP:

$$\boldsymbol{\mu}_t \sim \mathcal{N}(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t-1}, \mathbf{W}), \quad \mathbf{W}(i,j) = \alpha_W \exp\left(-\frac{(i-j)^2}{2l_W^2}\right). \quad (5)$$

This allows to implicitly specify the prior over the transition function of the structure. The rate of the evolution is controlled with the top-level GP covariance function.

The exact posterior of the parameters is intractable, therefore approximate inference methods are required. Inference is based on Expectation Propagation [8] in this paper.

The structural assumptions in sparse models are studied in the literature. The group lasso [9] provides sparse solutions for predefined groups of coefficients. Group constraints for sparse models include smooth relevance vector machines [10], Boltzmann machine prior [11]; spatio-temporal coupling of the parameters [12], [13]. In [14] a spatio-temporal structure is modelled with a one-level GP prior. In contrast to that model the new one introduces an additional level of a GP prior for temporal dependencies, therefore the temporal and spatial structures are decoupled, adding flexibility to the model. The high-level GP controls the change of spike groups in time while the low-level GP allows the local changes within each group.

III. NUMERICAL EXPERIMENTS

The performance of the proposed hierarchical GP algorithm is compared with the one-level GP prior introduced in [14]. We apply both algorithms to the problem of object detection in video sequences. The Convoy dataset [15] is used where the frame difference is applied for moving object detection. The sparse observations are obtained as $\mathbf{y}_t = \mathbf{X}\boldsymbol{\beta}_t$, where $\mathbf{X}$ is the matrix with i.i.d. Gaussian elements, the number of observations is $\approx 40\%$ of the dimension of the hidden signal $\boldsymbol{\beta}_t$. This procedure corresponds to compressive sensing observations [16].

The reconstruction results on the sample frames are presented in Figure 1. The obtained performance measures values can be found in Table I. The F-measure compares binary masks computed by thresholding the true vectors $\boldsymbol{\beta}$ and the posterior estimates $\widehat{\boldsymbol{\beta}}$. NMSE represents normalised mean squared error between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$. The proposed algorithm shows better results in terms of both measures.

IV. CONCLUSION

In this work we propose a new approach for spatial structure modelling in sparse models that allows to capture complex temporal evolution of data patterns. We also develop an efficient inference method based on EP. The numerical experiments show superiority of the proposed model over the current state-of-the-art.

(a) Original frame   (b) Frame with subtracted background   (c) Hierarchical GP reconstruction   (d) One-level GP reconstruction
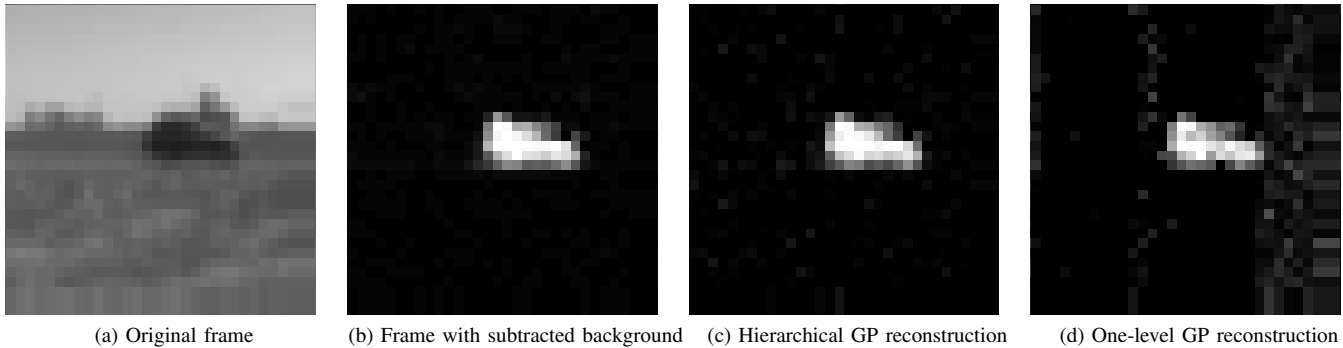
Fig. 1. Sample frame with reconstruction results from sparse measurements. (a): the original non-compressed frame; (b): object detection results based on non-compressed frame difference (static background frame is subtracted from the original frame); (c): reconstruction of compressed object detection results based on the proposed hierarchical GPs method; (d): reconstruction of the compressed object detection results based on the one-level GP method. (b) represents the true $\boldsymbol{\beta}_t$, while (c) and (d) are posterior estimates $\widehat{\boldsymbol{\beta}}_t$ obtained by the reconstruction algorithms.

TABLE I
PERFORMANCE OF SIGNAL RECONSTRUCTION FOR BACKGROUND
SUBTRACTION IN VIDEO

Measure	Hierarchical GP	One-level GP
NMSE	**0.0153**	0.0392
F-measure	**0.9870**	0.9403

REFERENCES

[1] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.

[2] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

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

[4] N. G. Polson and J. G. Scott, "Shrink globally, act locally: Sparse Bayesian regularization and prediction," *Bayesian Statistics*, vol. 9, pp. 501–538, 2010.

[5] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[6] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd ed. Academic Press, 2008.

[7] M. Deisenroth and S. Mohamed, "Expectation propagation in Gaussian process dynamical systems," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2609–2617.

[8] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 2001, pp. 362–369.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[10] A. Schmolck, "Smooth relevance vector machines," Ph.D. dissertation, University of Exeter, 2008.

[11] A. Drémeau, C. Herzet, and L. Daudet, "Boltzmann machine and mean-field approximation for structured sparse decompositions," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3425–3438, 2012.

[12] M. A. Van Gerven, B. Cseke, F. P. De Lange, and T. Heskes, "Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior," *NeuroImage*, vol. 50, no. 1, pp. 150–161, 2010.

[13] 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 27*, 2014, pp. 1628–1636.

[14] M. R. Andersen, A. Vehtari, O. Winther, and L. K. Hansen, "Bayesian inference for spatio-temporal spike and slab priors," *arXiv preprint arXiv:1509.04752*, 2015.

[15] G. Warnell, S. Bhattacharya, R. Chellappa, and T. Basar, "Adaptive-rate compressive sensing using side information," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3846–3857, 2015.

[16] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, and R. G. Baraniuk, "Compressive sensing for background subtraction," in *Proceedings of 10th European Conference on Computer Vision*, 2008, pp. 155–168.