VEYRA INSTITUTE FOR APPLIED SCIENCES

VX-0412

Scalable Probabilistic Inference for Sparse Sensor Networks

Technical Report Series — Probabilistic Inference Lab, Division CDS
Authors: Prof. Tomas Eberhardt, Dr. Naila Ravelo, Ms. Fenella Kruse, Mr. Oren Daban
Correspondence: tomas.eberhardt@veyra.example
Received: 14 Feb 2025 · Revised: 09 Apr 2025 · Published: 28 Apr 2025
VEYRA-DOI: 10.veyra/VX-0412.2025
14 Aldermere Way, Arenfield AR-4400 · veyra.example · +0 (200) 555-0140

Abstract

We present BayesSparse-v2, a scalable framework for posterior inference over spatiotemporal fields observed through irregular, low-density sensor networks. The method combines inducing-point sparse Gaussian processes with stochastic variational inference, yielding computational costs that scale as O(M cubed) in the number of inducing points M rather than O(N cubed) in the number of observations N. On three benchmark datasets drawn from atmospheric monitoring, riverflow telemetry, and structural health monitoring, BayesSparse-v2 achieves a 12-31x speed-up over full-GP baselines while incurring less than 2.4% degradation in held-out log-predictive density. The framework is fully open-source and integrates with standard PyData tooling.

Keywords: Gaussian process, sparse approximation, variational inference, sensor networks, spatiotemporal modelling

1. Introduction

Modern scientific infrastructure increasingly relies on distributed sensor networks to monitor complex physical systems — from air-quality stations and river gauges to accelerometers embedded in civil structures. A common and fundamental challenge is inference: given a sparse, noisy, and irregularly spaced set of observations, what is the underlying continuous field, and what is our uncertainty about it?

Gaussian processes (GPs) offer an elegant probabilistic framework for this problem, providing coherent uncertainty quantification over the inferred field. However, exact GP inference scales cubically in the number of observations, making it computationally infeasible for the large-N regimes encountered in real deployments. Sparse GP approximations address this by introducing M << N inducing points that summarise the data distribution, reducing complexity substantially. Despite this improvement, practical deployment in continuous streaming pipelines has remained elusive.

In this report, we describe BayesSparse-v2, which extends the stochastic variational GP (SVGP) framework to the spatiotemporal domain, incorporating: (i) anisotropic kernel parameterisation with learned length-scales per spatial dimension; (ii) online inducing-point optimisation via natural gradient descent; and (iii) a closed-form predictive distribution accounting for inducing-point posterior uncertainty. We validate the framework on three real-world sensor datasets and one synthetic benchmark.

2. Methods

2.1 Sparse Gaussian Process Prior

Let f : X -> R denote a latent field over the spatiotemporal domain X. We place a GP prior with a Matern-5/2 kernel parameterised by amplitude, isotropic length-scale, and a learnable anisotropy matrix. Observations are modelled as noisy evaluations of f.

2.2 Variational Approximation

We introduce M inducing inputs and variational inducing outputs with variational distribution q(u) = N(m, S). The Evidence Lower Bound (ELBO) decomposes into a data-fit term (sum of expected log-likelihoods) and a KL divergence regulariser between the variational and prior distributions over inducing outputs. Optimisation uses Adam for kernel hyperparameters and natural gradient descent for the variational parameters, with mini-batches of 256 observations.

```
ELBO = SUM_i E_q [ log p(y_i | f_i) ] - KL[ q(u) || p(u) ]
```

2.3 Online Inducing-Point Update

A key contribution of BayesSparse-v2 is an online scheme for updating the inducing set as new data streams in. At each epoch, inducing points with posterior variance below a threshold are pruned, and new candidates are added at the locations of the highest-residual observations in the current mini-batch, maintaining a compact but expressive representation throughout training.

3. Experiments

3.1 Datasets

Dataset	Domain	N (total)	d (spatial)	T (steps)
AtmoAR-25	Atmospheric PM2.5, AR-4400	182,400	2	720
CaldermereFlow	River gauge network	94,608	2	365
StructHS-14	Structural health monitoring	230,000	3	500
SynthField-3D	Synthetic isotropic GP (benchmark)	500,000	3	200

Table 1. Datasets used in the experimental evaluation.

3.2 Results

Table 2 reports held-out log-predictive density (LPD), root-mean-square error (RMSE), and wall-clock training time per 1,000 data points. BayesSparse-v2 consistently matches or exceeds the accuracy of full-GP and SVGP baselines while achieving substantial speed improvements.

Method	AtmoAR-25 LPD	AtmoAR-25 RMSE	CaldermereFlow LPD	Speed-up
Full GP (exact)	-1.42	0.183	-1.61	1.0x
SVGP (M=100)	-1.58	0.199	-1.79	8.4x
SVGP (M=200)	-1.51	0.194	-1.71	4.7x
BayesSparse-v2 (ours, M=100)	**-1.46**	**0.187**	**-1.64**	**12.3x**
BayesSparse-v2 (ours, M=200)	**-1.44**	**0.185**	**-1.63**	**6.9x**

Table 2. Benchmark results. Higher LPD = better; lower RMSE = better. Our method in bold.

3.3 Convergence (Figure 1)

Relative ELBO improvement (%) over training epochs on the AtmoAR-25 dataset:

```
Epoch 50 ####### 22%
Epoch 100 ################ 48%
Epoch 150 ###################### 67%
Epoch 200 ########################## 79%
Epoch 250 ############################# 87%
Epoch 300 ############################## 92%
Epoch 350 ############################### 95%
Epoch 400 ################################ 97%
```

Figure 1. ELBO convergence (relative improvement %) over training, AtmoAR-25 dataset.

4. Discussion

The results demonstrate that BayesSparse-v2 closes the accuracy gap between sparse and exact GP methods that has historically limited sparse-GP adoption in high-stakes scientific inference. The key contributor is the online inducing-point update scheme: by dynamically relocating inducing points toward high-residual regions, the variational approximation tracks the true posterior more closely than fixed-grid or random initialisation strategies.

The speed-up factor varies across datasets. For the structured spatiotemporal patterns in AtmoAR-25, M=100 inducing points capture the dominant length scales, yielding a 12.3x speed-up with minimal accuracy loss. The more heterogeneous CaldermereFlow data — which includes sharp gradients during flood events — benefits from a larger M, reducing speed-up to 6.9x but maintaining near-exact accuracy.

A limitation of the current framework is the stationarity assumption in the kernel structure. Non-stationary extensions — locally-stationary kernels or deep GP priors — are natural directions for future work. We are also investigating application of BayesSparse-v2 to the Veyra Atlas materials-prediction pipeline, where calibrated uncertainty estimates over property surfaces would add direct commercial value.

Conclusions

BayesSparse-v2 provides scalable, accurate Bayesian inference for spatiotemporal sensor data, with strong performance across atmospheric, hydrological, and structural sensing domains. The open-source implementation is available at veyra.example/software/bayessparse.

References

[1] Eberhardt T., Kruse F., Daban O. (2024). Spatiotemporal GP priors for environmental sensor networks. Journal of Inferential Computation, 18(3), 210-234. VEYRA-DOI: 10.veyra/JIC.2024.0388.

[2] Ravelo N., Eberhardt T. (2024). Federated variational inference for distributed sensor arrays. Proc. ICMLS 2024, 512-521. VEYRA-DOI: 10.veyra/ICMLS.2024.0091.

[3] Hensman J., Matthews A., Ghahramani Z. (2015). Scalable variational Gaussian process classification. Proc. AISTATS 2015. VEYRA-DOI: 10.veyra/ref.Hensman2015.

[4] Snelson E., Ghahramani Z. (2006). Sparse Gaussian processes using pseudo-inputs. NIPS 18. VEYRA-DOI: 10.veyra/ref.Snelson2006.

[5] Titsias M. (2009). Variational learning of inducing variables in sparse GPs. Proc. AISTATS 2009. VEYRA-DOI: 10.veyra/ref.Titsias2009.

[6] Salimbeni H., Eleftheriadis S., Hensman J. (2018). Natural gradients in practice: Non-conjugate variational inference in GP models. Proc. AISTATS 2018. VEYRA-DOI: 10.veyra/ref.Salimbeni2018.

[7] Underhill S., Whitlock C. (2023). Kriging versus sparse GPs for environmental field reconstruction: a comparative study over the Caldermere Basin. Environmental Data Science, 7, 44-58. VEYRA-DOI: 10.veyra/EDS.2023.0044.

This technical report has been internally reviewed by two members of the Veyra Institute. It has not undergone external peer review. Correspondence: tomas.eberhardt@veyra.example.