Multi-resolution Multi-task Gaussian Processes

Oliver Hamelijnck[1,2] ohamelijnck@turing.ac.uk
Theodoros Damoulas[1,2,3] tdamoulas@turing.ac.uk
Kangrui Wang[1,2] kwang@turing.ac.uk
Mark Girolami[1,4] mgirolami@turing.ac.uk

[1]The Alan Turing Institute
[2]University of Warwick, Department of Computer Science
[3]University of Warwick, Department of Statistics
[4]University of Cambridge, Department of Engineering

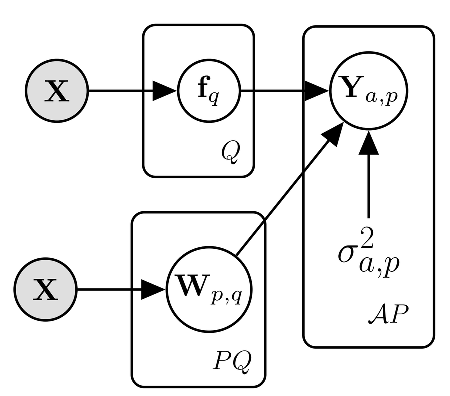
Introduction

- Consider integrating observations at varying spatio-temporal sampling resolutions (*multi-resolution*, MR), noise levels (*multi-fidelity*) and tasks
- Develop MR-GPRN that extends the Gaussian Process Regression Network (GPRN) of [4] to handle multi-resolution observations, additionally we utilise a composite likelihood to adjust posterior uncertainty under model misspecification
- Dervive MR-DGP that extends the Deep GP of [3] to handle multi-resolution data and any biases between the observation processes

Modelling Dependent Observations

- Construct $\mathcal{A}$ datasets $\{(\mathbf{X}_a, \mathbf{Y}_a)\}_{a=1}^{\mathcal{A}}$ where $\mathbf{Y}_a \in \mathbb{R}^{N_a \times P}$ for P tasks and N_a observations and $\mathbf{X}_a \in \mathbb{R}^{N_a \times |\mathcal{S}_a| \times D_a}$ over a (discretised) sampling area $\mathcal{S}_a$
- Introduce Q latent GPs $\mathbf{f}_q \sim \mathcal{GP}(0, \mathbf{K}_q)$ and PQ task-specific GPs $\mathbf{W}_{p,q} \sim \mathcal{GP}(0, \mathbf{K}_{p,q})$. Link these to the different resolutions through the likelihood:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{f}) = \prod_{a=1}^{\mathcal{A}} \prod_{p=1}^{P} \prod_{n=1}^{N_a} \mathcal{N}(\mathbf{Y}_{a,p,n} | \frac{1}{|\mathcal{S}_a|} \int_{\mathcal{S}_{a,n}} \sum_{q=1}^{Q} \mathbf{W}_{p,q}(\mathbf{x}) \odot \mathbf{f}_q(\mathbf{x})\, d\mathbf{x}, \sigma^2_{a,p}\mathbf{I})^{\phi}$$



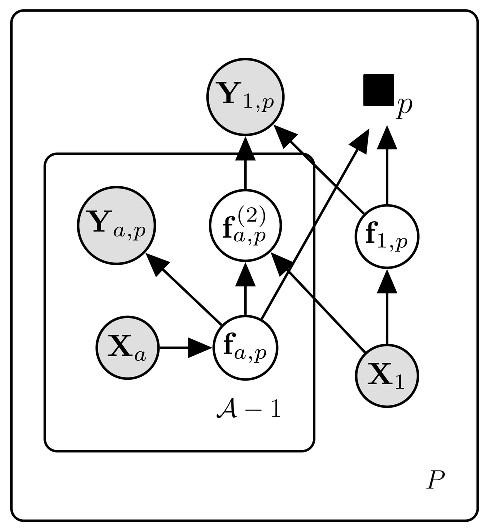
Algorithm 1: Inference of MR-GPRN

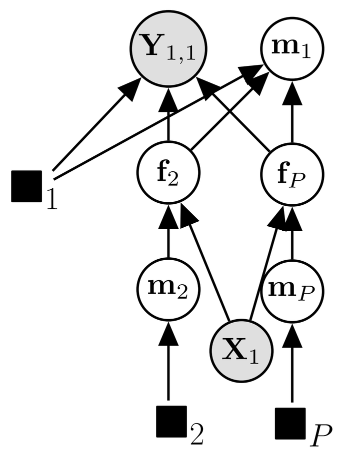
Input: $\{(\mathbf{X}_a, \mathbf{Y}_a)\}_{a=1}^{\mathcal{A}}$, initial θ,
$\hat{\theta} \leftarrow \arg\max_{\theta} \sum_{a=1}^{\mathcal{A}} \ell(\mathbf{Y}_a|\theta)$
$\mathbf{H} \leftarrow \sum_{a=1}^{\mathcal{A}} (\nabla \ell(\mathbf{Y}_a|\hat{\theta})(\nabla \ell(\mathbf{Y}_a|\hat{\theta}))^T$
$\mathbf{J} \leftarrow \nabla^2 \ell(\mathbf{Y}|\hat{\theta})$
$\phi \leftarrow \begin{cases} \frac{|\hat{\theta}|}{\mathrm{Tr}[\mathbf{H}(\hat{\theta})^{-1}\mathbf{J}(\hat{\theta})]} \\ \frac{\mathrm{Tr}[\mathbf{H}(\hat{\theta})\mathbf{J}(\hat{\theta})^{-1}\mathbf{H}(\hat{\theta})]}{\mathrm{Tr}[\mathbf{H}(\hat{\theta})]} \end{cases}$
$\theta_1 \leftarrow \arg\min_{\theta} \left(\sum_{a=1}^{\mathcal{A}} \phi \mathbb{E}_q \left[\ell(\mathbf{Y}_a|\theta)\right] + \mathcal{KL}\right)$

Modelling Biased Observations

- We assume that the highest resolution is the observation of interest and learn the mapping and calibration from the lower resolution observations
- A mixture of DGP experts allows for non-overlapping datasets

- Each $\mathbf{f}^{(\cdot)}{}_{a,p} \sim \mathcal{GP}(0, \mathbf{K}^{(\cdot)}{}_{a,p})$
- $\mathbf{m}_p = \sum_{a=1}^{\mathcal{A}} \beta_a \odot \mathbf{f}^{(\cdot)}_{a,p}$
- $\beta_a = (1 - \mathbf{V}_a) \sum_i^a \mathbf{V}_i$

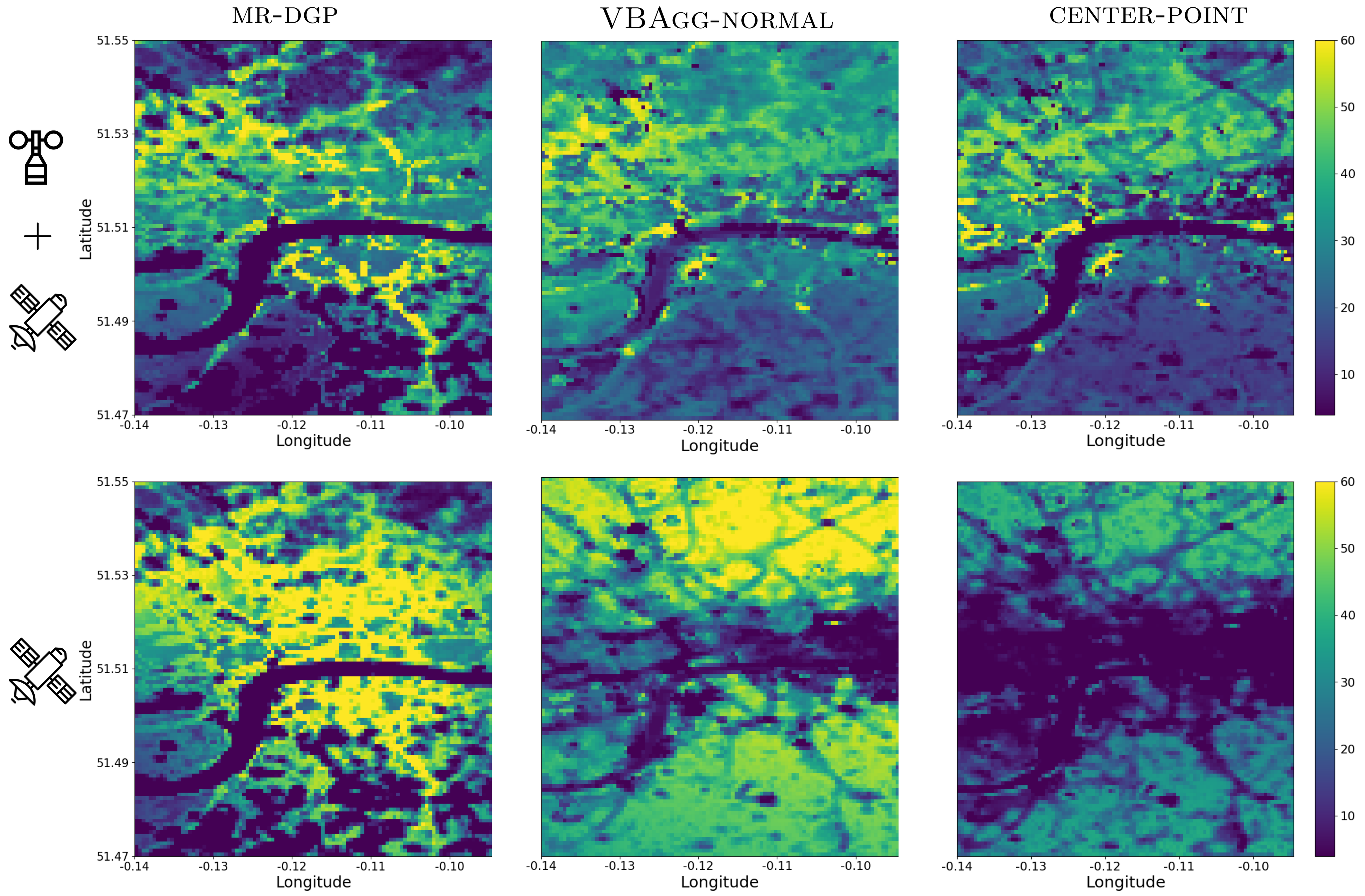




- Each likelihood has its own noise term allowing for multi-fidelity learning
- Propagating samples allows predictions at all resolutions and tasks

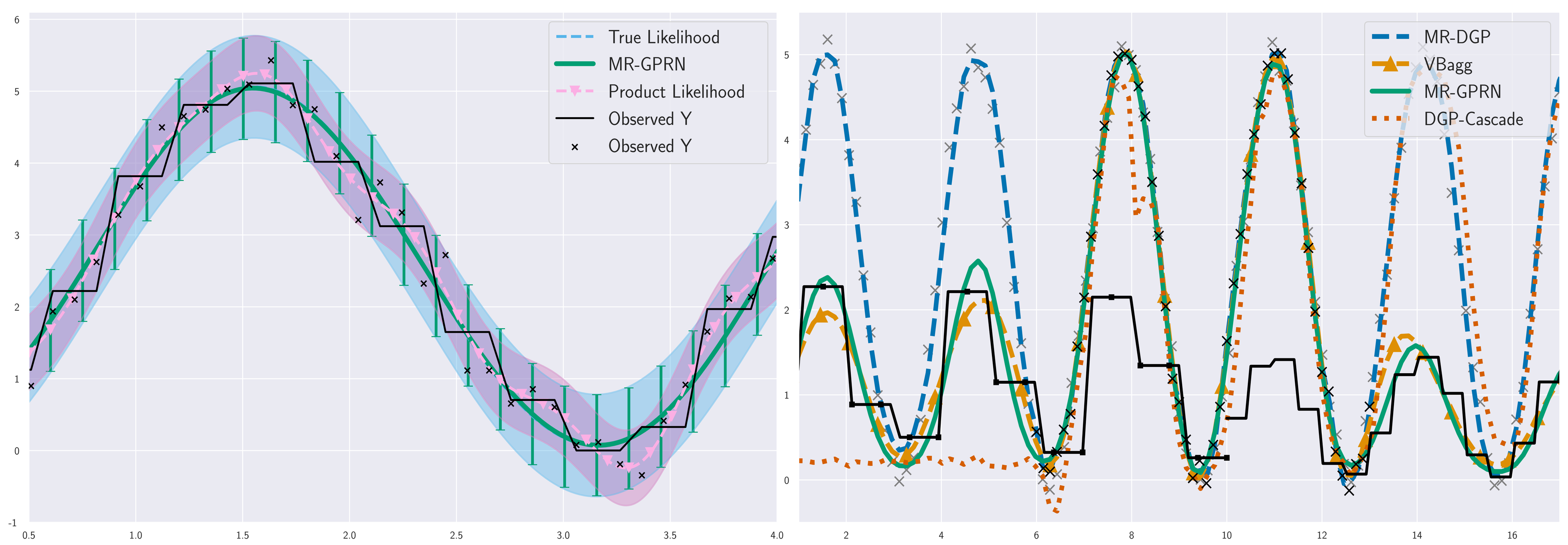
Forecasting NO_2 Across London

- Spatio-temporal estimation and forecasting of NO_2 levels in London



- **Top row**: Spatial slices with observations from both LAQN and the satellite model (low spatial resolution) are present. All models are able to capture the high resolution structure
- **Bottom row**: Spatial slices from the same models where *only* observations from the satellite model are present. Only MR-DGP retains the high resolution structure

Biased and Dependent Observations



- **Left**: MR-GPRN corrects for model misspecification from a product likelihood through the use of a composite likelihood
- **Right**: MR-DGP learns a scaling bias between multi-resolution datasets allowing the true predictive mean to be recovered instead of resorting to the uncalibrated observations. Whereas DGP-Cascade is unable to handle the non-overlapping multi-resolution datasets

Variational Lower Bounds

- For MR-GPRN we derive efficient closed form variational lower bounds. We augment all latent GPs with inducing points and derive the ELL:

$$\mathrm{ELL}_{a,p,n,k} = \pi_k \log \mathcal{N}\left(Y_{a,p,n} \mid \frac{1}{|\mathcal{S}_{a,n}|} \sum_{\mathbf{x} \in \mathcal{S}_{a,n}} \sum_{q=1}^{Q} \boldsymbol{\mu}^{(w)}_{k,p,q}(\mathbf{x}) \boldsymbol{\mu}^{(f)}_{k,q}(\mathbf{x}), \sigma^2_{a,p}\right)$$
$$- \frac{\pi_k}{2\sigma^2_{a,p}} \frac{1}{|S_{a,n}|^2} \sum_{q=1}^{Q} \sum_{\mathbf{x}_1,\mathbf{x}_2} \boldsymbol{\Sigma}^{(w)}_{k,p,q} \boldsymbol{\Sigma}^{(f)}_{k,q} + \boldsymbol{\mu}^{(f)}_{k,q}(\mathbf{x}_1) \boldsymbol{\Sigma}^{(w)}_{k,p,q} \boldsymbol{\mu}^{(f)}_{k,q}(\mathbf{x}_2) \boldsymbol{\mu}^{(w)}_{k,p,q}(\mathbf{x}_1) \boldsymbol{\Sigma}^{(f)}_{k,q} \boldsymbol{\mu}^{(w)}_{k,p,q}(\mathbf{x}_2)$$

- For MR-DGP we sample from the base GPs and propagate the samples up:

$$q(\mathbf{m}^*_1) = \int q(\mathbf{m}^*_1 | \mathrm{Pa}(\mathbf{m}^*_1)) \prod_{\mathbf{f} \in \mathrm{Pa}(\mathbf{m}^*_1)} q(\mathbf{f})\, d\mathrm{Pa}(\mathbf{m}^*_1) \approx \frac{1}{S} \sum_{s=1}^{S} q(\mathbf{m}^*_1 | \{\mathbf{f}^{(s)}\}_{\mathbf{f} \in \mathrm{Pa}(\mathbf{m}^*_1)})$$

Results

Biased Mean

Model	RMSE	MAPE
MR-CASCADE	2.12	0.16
VBAgg-normal	1.68	0.14
MR-GPRN	1.6	0.14
MR-DGP	**0.19**	**0.02**

NO2 Across London

Model	RMSE	MAPE
Single GP	20.55 ± 9.44	0.8 ± 0.16
center-point	18.74 ± 12.65	0.65 ± 0.21
VBAgg-normal	16.16 ± 9.44	0.69 ± 0.37
MR-GPRN w/o CL	12.97 ± 9.22	0.56 ± 0.32
MR-GPRN w CL	11.92 ± 6.8	0.45 ± 0.17
MR-DGP	$\mathbf{6.27 \pm 2.77}$	$\mathbf{0.38 \pm 0.32}$

- MR-DGP is able to substantially outperform both VBAgg-normal, MR-GPRN
- MR-DGP can handle biases between observation processes

Future Work

- Incorporate physical constraints in latent space through physics-informed machine learning
- Reduce computational complexity through state-space GP formulations
- Explore further model robustness through recent advances in Generalised Variance Inference [5]
- Explore further MR constructions e.g. the concurrent submissions [6, 7]

Key References

[1] Varin, C., Reid, N., and Firth, D. An overview of composite likelihood methods. *Statist.Sinica*, 2011.
[2] Law, H. C. L., Sejdinovic, D., Cameron, E., Lucas, T. C., Flaxman, S., Battle, K., and Fukumizu,K. Variational learning on aggregate outputs with Gaussian processes. NeurIPS, 2018.
[3] Salimbeni, H. and Deisenroth, M. Doubly stochastic variational inference for Deep Gaussian processes. *Advances in Neural Information Processing Systems 30*, 2017.
[4] Wilson, A. G., Knowles, D. A., and Ghahramani, Z. Gaussian process Regression Networks, ICML, 2012.
[5] Knoblauch, J. Jewson, J. Damoulas, T. Generalized Variational Inference, arXiv, 2019.
[6] Yousefi, F. Smith, M. T. and Alvarez, Mauricio. A. Multi-task learning for aggregated data using Gaussian processes. NeurIPS 2019.
[7] Tanaka, Y. Tanaka T. Iwata T. Kurashima T. Okawa M. Akagi Y. and Toda H. Spatially aggregated Gaussian processes with multivariate areal outputs. NeurIPS 2019.

Acknowledgments

This work is funded by the Lloyd's Register Foundation programme and supported by The Alan Turing Institute under EPSRC grant EP/N510129/1 in collaboration with the Greater London Authority.

CODE

PAPER

Vancouver 2019