GP-ALPS: Automatic Latent Process Selection for Multi-Output Gaussian Process Models

Pavel Berkovich[1,2], Eric Perim[1], and Wessel Bruinsma[1,3]
[1]Invenia Labs, [2]University College London, [3]University of Cambridge
<firstname>.<lastname>@invenialabs.co.uk

- The choice of the **latent space** for multi-output GPs forms a **model selection problem**, with the usual challenges.
- GP-ALPS: **binary variables** automatically **select latent processes** that meaningfully contribute to explaining the data.
- Exact inference is **intractable**. We propose a mean-field **approximate inference** scheme using continuous relaxations.

Motivation

- Gaussian processes commonly extended to multiple outputs by linearly combining latent GPs $(x_j)_{j=1}^m$ via a "mixing matrix" H:

$$x_j \sim \mathcal{GP}(0, k_j(t,t')), \quad f(t) = Hx(t), \quad y_i(t) \sim \mathcal{N}(f_i(t), \sigma_i^2).$$

- Choice of kernels and number of latent processes is important.
- Can this be done automatically?

GP-ALPS

- A binary, multiplicative "switch" b_j for each latent process.
- Generative model:

$$x_j \sim \mathcal{GP}(0, k_j(t,t')), \quad b_j \sim \mathrm{Bern}(\theta_j), \quad H_{ij} \sim \mathcal{N}(0, s_{ij}),$$
$$f(t) = H\,(x(t) \circ b), \quad y_i(t) \sim \mathcal{N}(f_i(t), \sigma_i^2).$$

- Each of the 2^m possible values of b corresponds to a model class.
- Bayesian Occam's razor encourages parsimonious posterior over b.

Approximate Inference Scheme

- Augment model with **inducing points** X^z [1] .
- Introduce **structured mean-field approximation** of posterior:

$$q(X, X^z, H, b) = p(X|X^z)q(X^z)q(H)q(b).$$

- Relax $p(b)$ and $q(b)$ using the **Concrete distribution** [2]:

$$p(b) \approx \prod_{j=1}^m \mathrm{Concrete}(b_j; \theta_j), \quad q(b) = \prod_{j=1}^m \mathrm{Concrete}(b_j; \rho_j).$$

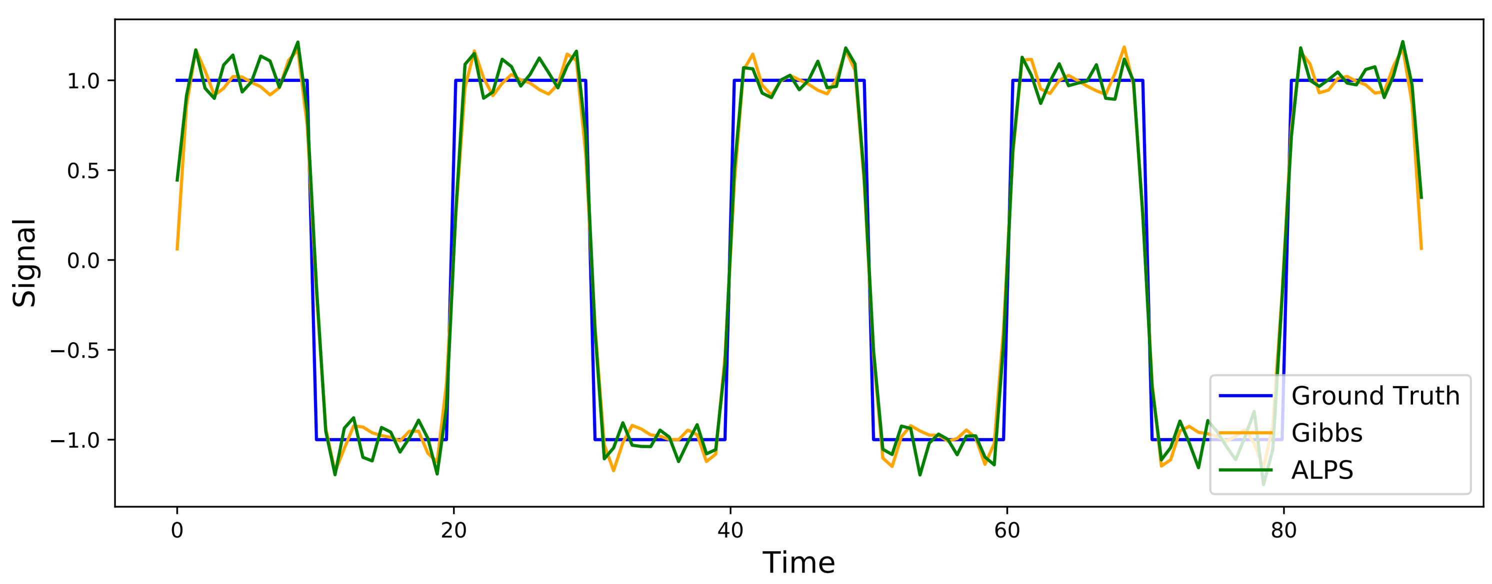
- Variational posterior q obtained by maximising the **evidence lower bound** (ELBO)

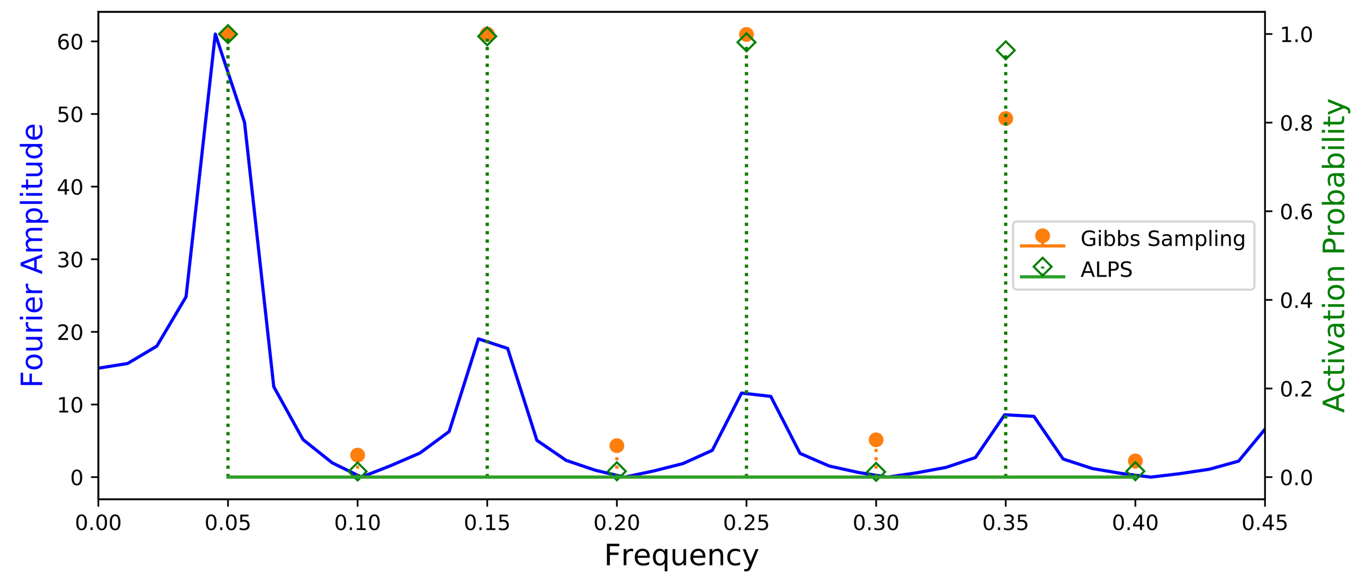
$$\log p(Y) \geq \mathcal{L} = \mathbb{E}_q \left[\log \frac{p(Y, X, X^z, H, b)}{q(X, X^z, H, b)} \right]$$

using stochastic, gradient-based optimisation in combination with the **reparametrisation trick** [3].

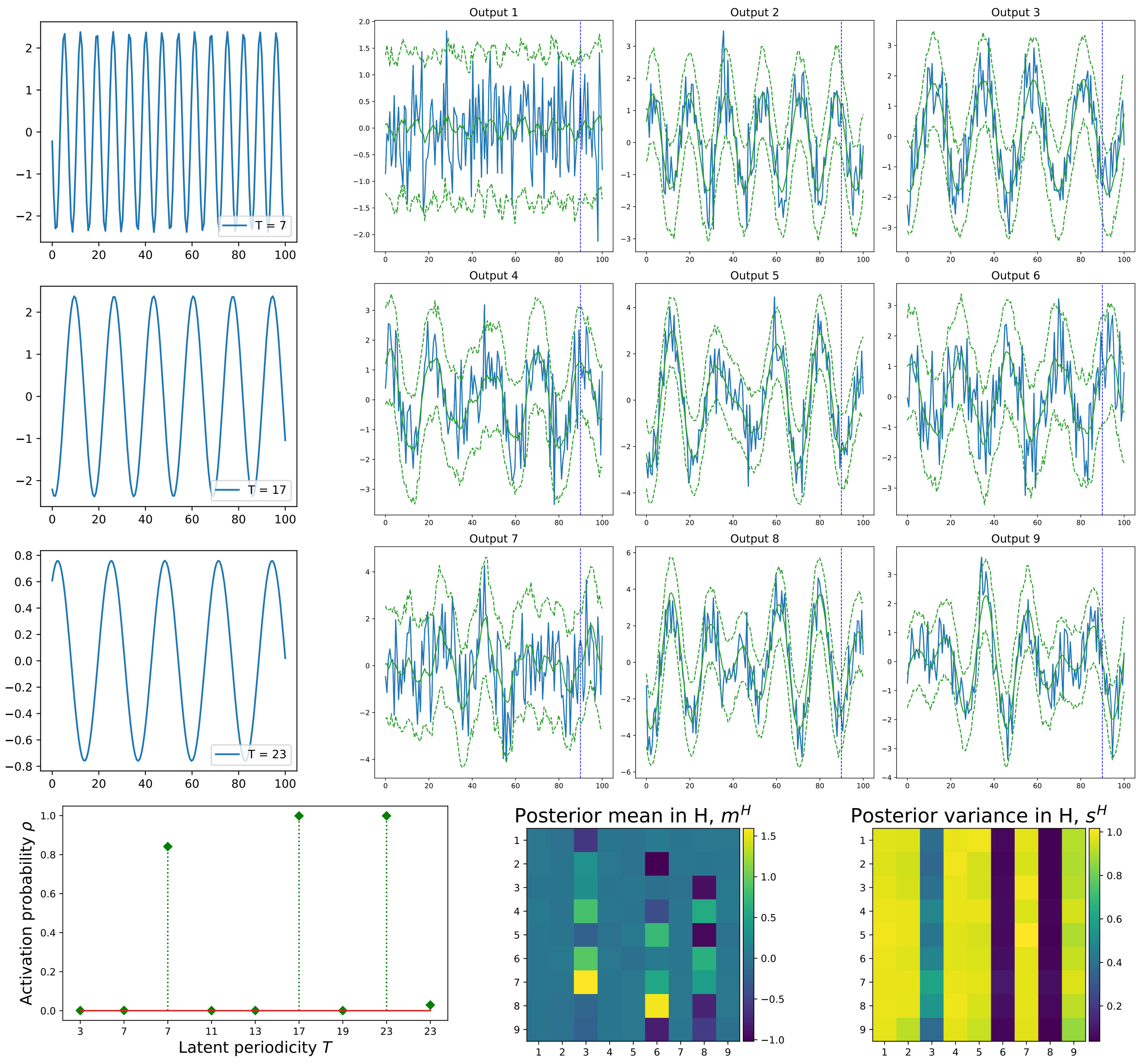
- ELBO is a sum over data points, hence **amenable to mini-batching** to scale to large datasets.

Square Wave Decomposition

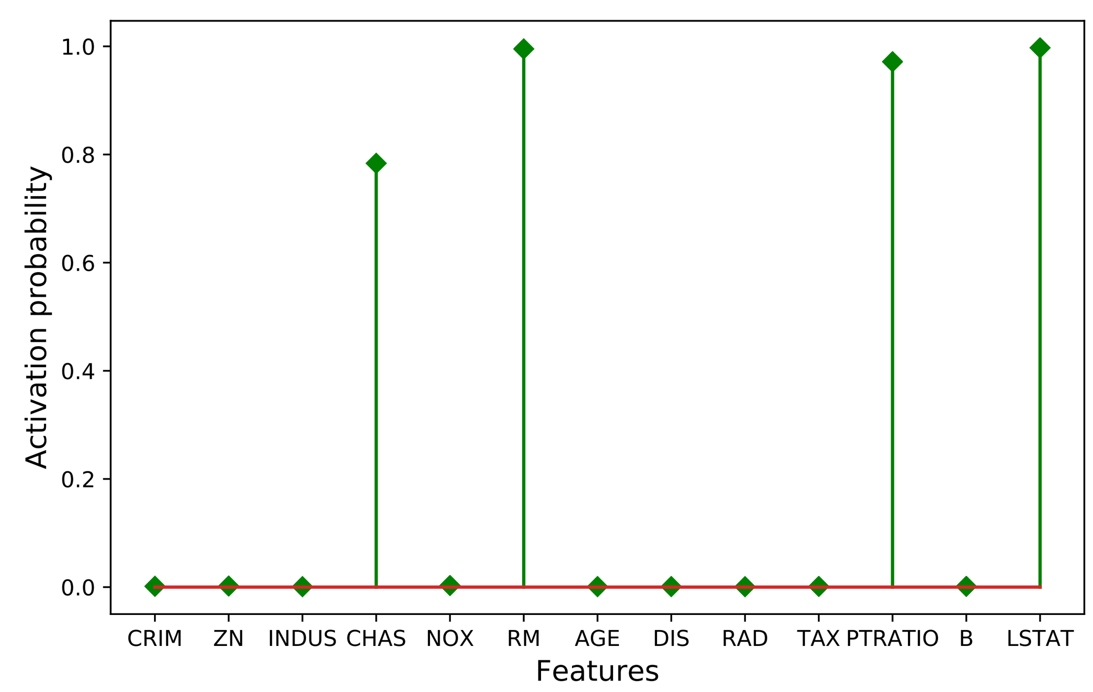


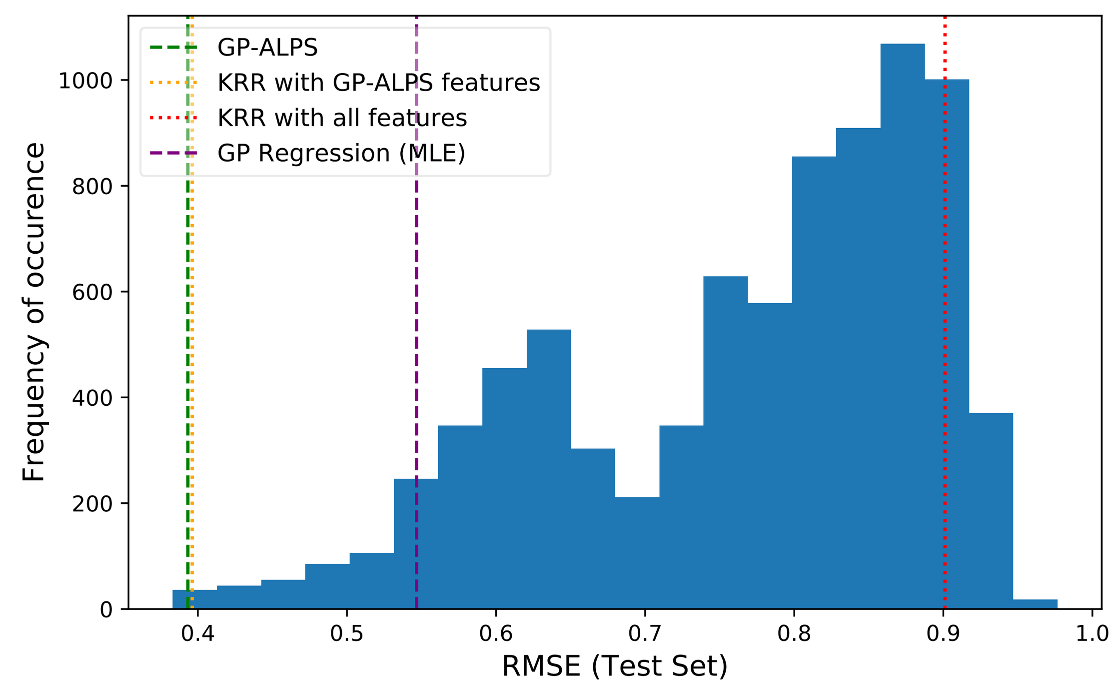


Noisy Mixture of Sinusoids



Boston Housing Dataset





[1] Dezfouli, Amir, and Edwin V. Bonilla. "Scalable inference for Gaussian process models with black-box likelihoods", 2015.
[2] Maddison, Chris J., Andriy Mnih, and Yee Whye Teh. "The concrete distribution: A continuous relaxation of discrete random variables', 2016.
[3] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes", 2013.