Thin and Deep Gaussian processes

Daniel Augusto de Souza, Alexander Nikitin, S. T. John, Magnus Ross, Mauricio A Álvarez, Marc Peter Deisenroth, João P. P. Gomes, Diego Mesquita, César L. C. Mattos

Scan me to see the poster, blogpost and more!

Why (deep) Gaussian processes?

- Gaussian processes (GPs) are widely used in machine learning for their simple uncertainty quantification;
- GP's kernel function → Modelling and uncertainty quality;
- Common stationary kernels, like square exponential and Matérn are unsuitable for non-stationary data;
  → Many geospatial processes, such as sea surface height, bathymetry.
- Popular research direction: how to make non-stationary kernels from common stationary ones.

From stationary to non-stationary

- Stationarity: $\mathrm{k}(\boldsymbol{a}, \boldsymbol{b}) = \mathrm{k}(\boldsymbol{a} - \boldsymbol{b}, 0)$;
  → Kernel is effectively a one-argument function, every slice looks the same.
- Isotropic stationary: $\mathrm{k}(\boldsymbol{a}, \boldsymbol{b}) = \pi_{\mathrm{k}}\left((\boldsymbol{a} - \boldsymbol{b})^{\mathrm{T}} \boldsymbol{\Delta}^{-1} (\boldsymbol{a} - \boldsymbol{b})\right)$;
  → Kernel is a function of the distance, weighted by the lengthscale matrix $\boldsymbol{\Delta}$.
- E.g., squared exponential kernel $\pi_{\mathrm{k}}(d^2) = \exp\left[-\frac{1}{2} d^2\right]$

Compositional kernels

- Non-linear warping function: $\boldsymbol{\tau}(\cdot) \to \mathrm{k}(\boldsymbol{\tau}(\boldsymbol{a}), \boldsymbol{\tau}(\boldsymbol{b}))$;
  → If $\tau(x) = \ell^{-1} \cdot x$, then ℓ gets absorbed in the kernel's lengthscales
- If $\boldsymbol{\tau}(\boldsymbol{x})$ is a neural network → Deep Kernel Learning;
- If $\boldsymbol{\tau}(\boldsymbol{x})$ is a GP → Compositional Deep Gaussian process;

Limitations

- The more complex $\boldsymbol{\tau}(\cdot)$ is, the harder the model is to interpret;
- For DKL, parameters are not Bayesian, thus the model overfits.
- For DGP, $\boldsymbol{\tau}(\cdot)$ cannot have zero prior mean → model collapse with increasing depth;

Lengthscale mixture kernels

- Positive semi-definite function field $\boldsymbol{\Delta}(\cdot)$

$$k_{\boldsymbol{\Delta}}(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\frac{\sqrt{|\boldsymbol{\Delta}(\boldsymbol{a})|}\sqrt{|\boldsymbol{\Delta}(\boldsymbol{b})|}}{|\boldsymbol{\Delta}(\boldsymbol{a}) + \boldsymbol{\Delta}(\boldsymbol{b})|}}\, \pi_{\mathrm{k}}\left((a - b)^{\mathrm{T}} \left[\frac{\boldsymbol{\Delta}(\boldsymbol{a}) + \boldsymbol{\Delta}(\boldsymbol{b})}{2}\right]^{-1} (a - b)\right)$$

- If $\sigma(\boldsymbol{\Delta}(\cdot))$ is a GP w/ linking function $\sigma(\cdot)$ → Deeply Non-stationary GP

Limitations

- Does not encode latent spaces; • Kernel scale affected by the pre-factor, giving rise to unwanted correlations;

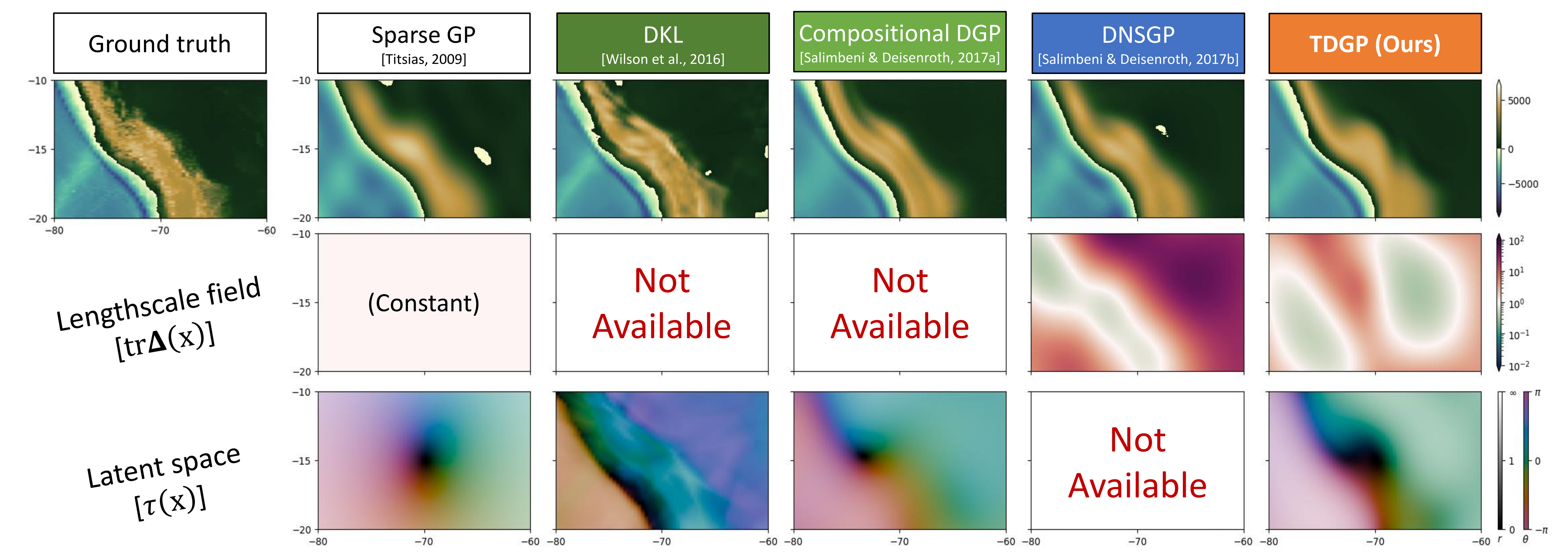
Our hybrid proposal

- We propose a compositional kernel: $\boldsymbol{\tau}(\boldsymbol{x}) = \mathbf{W}(\boldsymbol{x}) \cdot \boldsymbol{x}$.
  - For deeper layers, $\boldsymbol{\tau}^{(2)}\left(\boldsymbol{\tau}^{(1)}(\boldsymbol{x})\right) = \mathbf{W}^{(2)}(\boldsymbol{\tau}(\boldsymbol{x})) \cdot \mathbf{W}^{(1)}(\boldsymbol{x}) \cdot \boldsymbol{x}$;
- At the neighborhood of $\boldsymbol{x}$, we have $\boldsymbol{\Delta}(\boldsymbol{x}) = [\mathbf{W}^{\mathrm{T}}(\boldsymbol{x})\mathbf{W}(\boldsymbol{x})]^{-1}$;
- Therefore, we learn latent spaces and lengthscale fields;
- If $\mathbf{W}(\cdot)$ is a GP, then we obtain our deep model. $\mathbf{W}(\cdot)$ can be zero mean and with learned variances, we optimize for dim. reduction

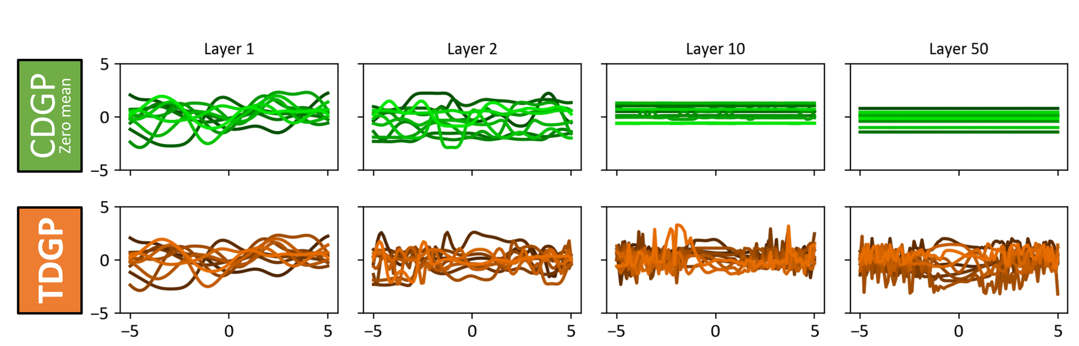
Limitations

- Adds $d \times q$ more GPs, making training harder; • Without adding a bias dimension to input, neighborhood of 0 is unaffected;

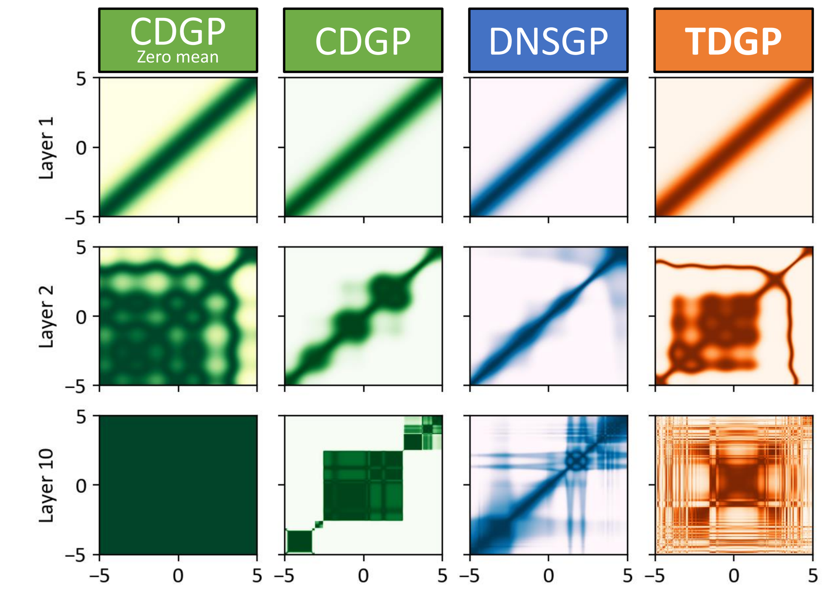
By **unifying** different approaches to **deep Gaussian processes**, we build **probabilistic models** that are more **interpretable** whilst learning lower-dimensional latent representations for **complex, non-stationary** data



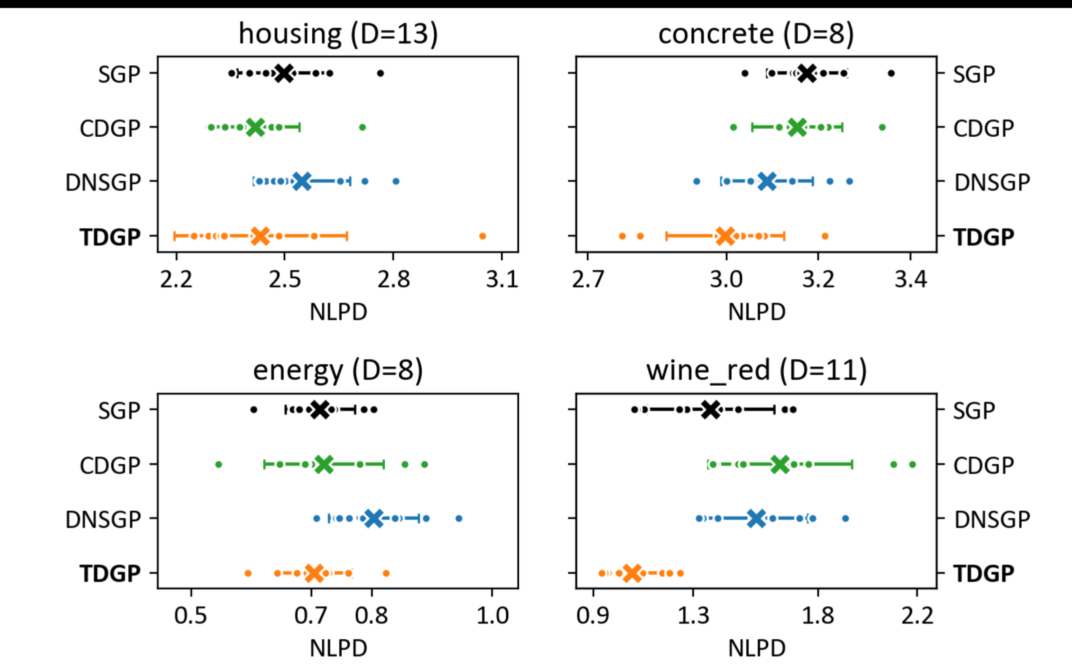
Very deep GP priors



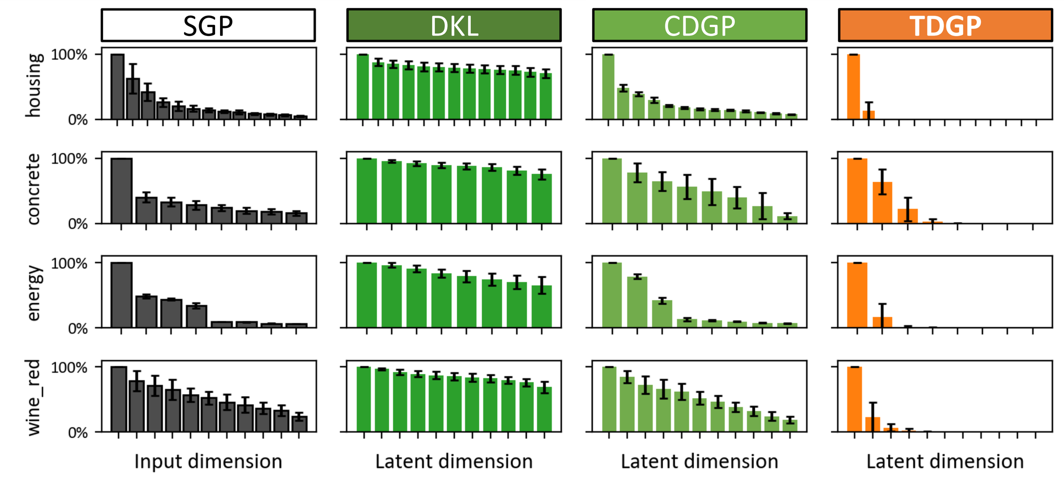
Prior kernel matrices



UCI datasets



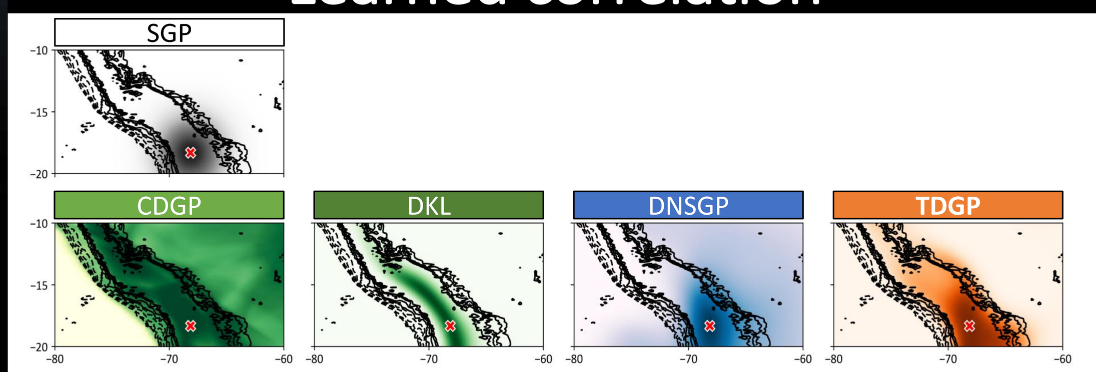
Dimensionality reduction



Bathymetry on Central Andes

	NLPD	MRAE
Sparse GP	-0.13 ± 0.09	1.19 ± 0.63
Deep Kernel Learning	3.85 ± 0.92	**0.59 ± 0.31**
Compositional DGP	-0.44 ± 0.12	0.83 ± 0.56
Deeply Nonstationary GP	-0.31 ± 0.12	1.12 ± 0.75
TDGP (Ours)	**-0.53 ± 0.10**	0.66 ± 0.43

Learned correlation



UCL | Aalto University | MANCHESTER 1824 The University of Manchester | UFC | FGV