Deep Mahalanobis Gaussian Process

Daniel Augusto de Souza, Diego Mesquita, César Lincoln Mattos, João Paulo Gomes

TL;DR

- In this work, we propose a new **non-stationary** GP model, focusing on greater interpretability;
- Our model models the lengthscales of the non-stationary kernel as a field of locally linear transformations;
- We compare this model with the traditional compositional deep GP in synthetic and popular regression datasets and obtain promising results.

From stationary kernels to non-stationarity

A kernel is stationary if:

$$k(a,b) = \phi\big((a-b)^T\Delta^{-1}(a-b)\big)$$

For a certain $\phi(d)\colon \mathbb{R} \to \mathbb{R}$ and lengthscale matrix Δ. For example, the RBF kernel is the one with $\phi(d) = \exp\left[-\frac{1}{2}d\right]$. As shown by Gibbs (1997), one can build a non-stationary kernel from these building blocks as follows:

$$\sqrt{2^d \frac{\sqrt{\Delta(a)}\sqrt{\Delta(b)}}{|\Delta(a)+\Delta(b)|}}\,\phi\big(2(a-b)^T[\Delta(a)+\Delta(b)]^{-1}(a-b)\big)$$

Monsters around the corner

1. The interpretation of $\Delta(x)$ becomes unclear due to the presence of the orange pre-factor (Gibbs, 1997);
2. The quadratic term inside ϕ usually defines the distance function of a semi-metric space, however, in the non-stationary case, the triangle inequality can be violated (Paciorek, 2003).

An alternative construction for RBF kernels

We propose using a different construction, first by rewriting the stationary kernel definition:

$$\begin{aligned} k(a,b) &= \phi\big((a-b)^T\Delta^{-1}(a-b)\big) \\ &= \phi\big((a-b)^T(W^TW)(a-b)\big) \\ &= \phi\big((Wa-Wb)^T(Wa-Wb)\big) \end{aligned}$$

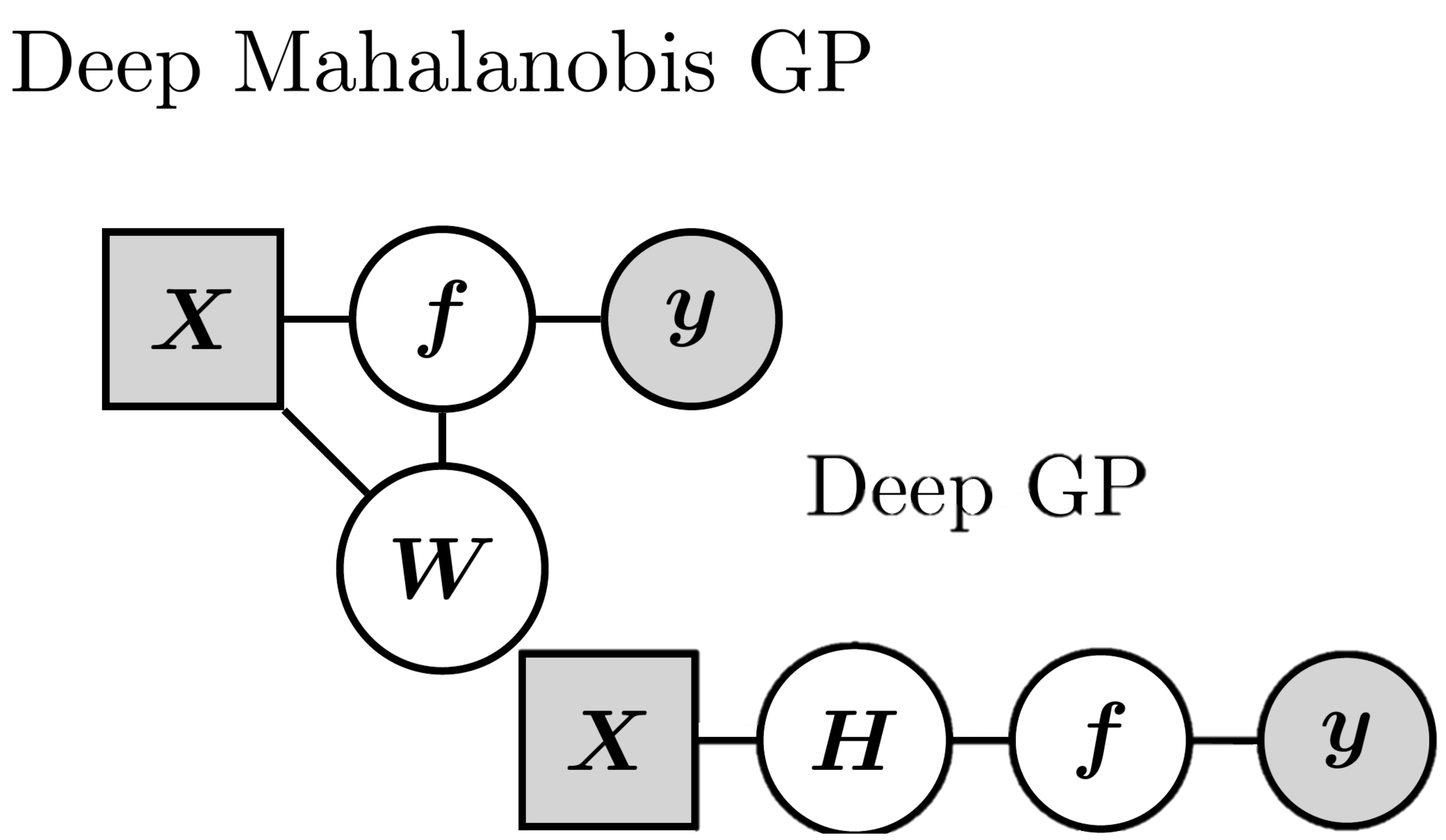
Where we split the positive-definite inverse lengthscale matrix into a product of ordinary matrices. From this point, now we can make W a function of the inputs:

$$\phi\big((W(a)\cdot a - W(b)\cdot b)^T(W(a)\cdot a - W(b)\cdot b)\big)$$

This new kernel is still a positive-definite function, doesn't contain the pre-factor from Gibbs and respects the triangle inequality.

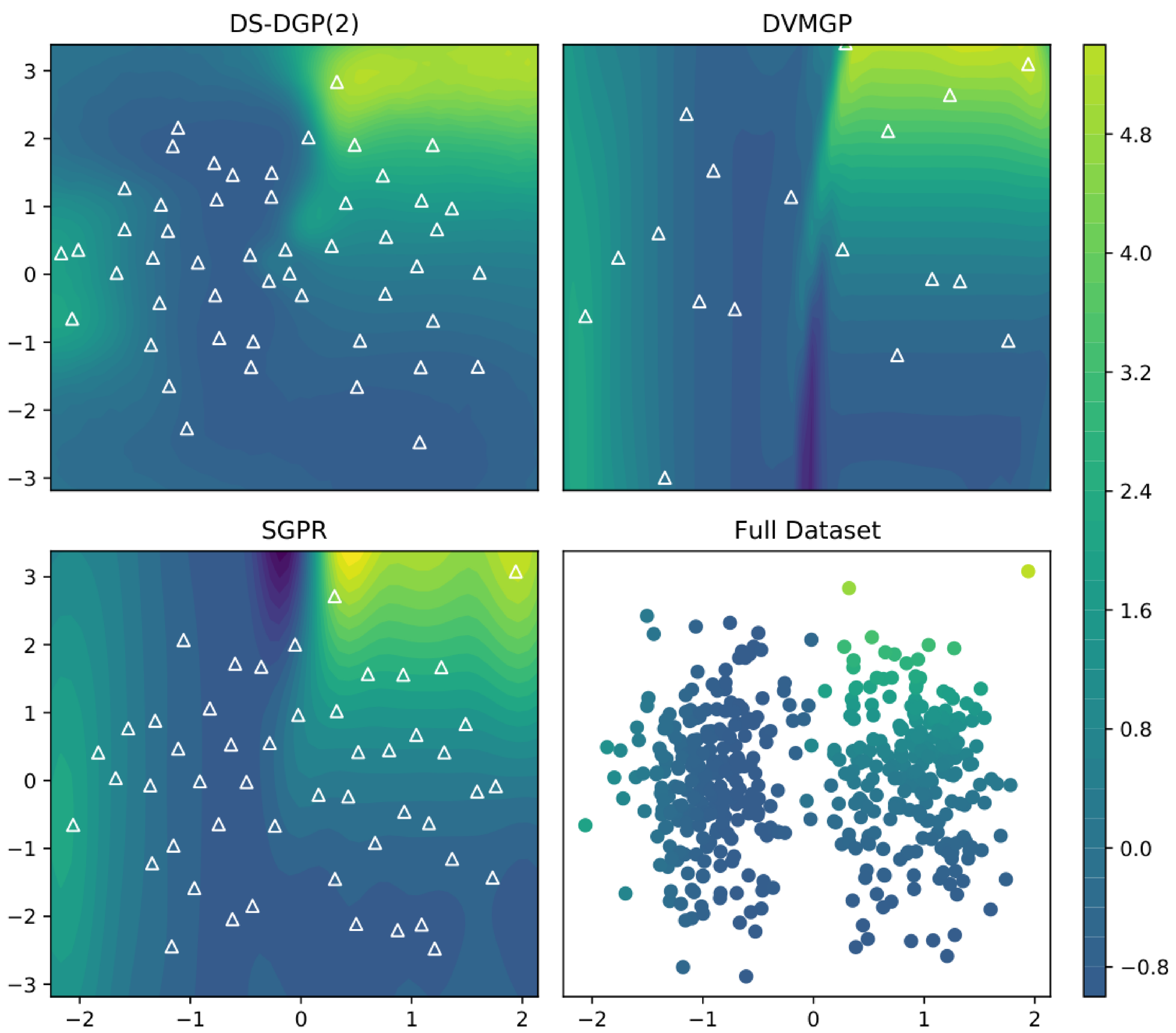
Deep Mahalanobis GP

By placing a Gaussian process prior on the entries of W, we get a deep model that maintains each GP layer still grounded on the input data.

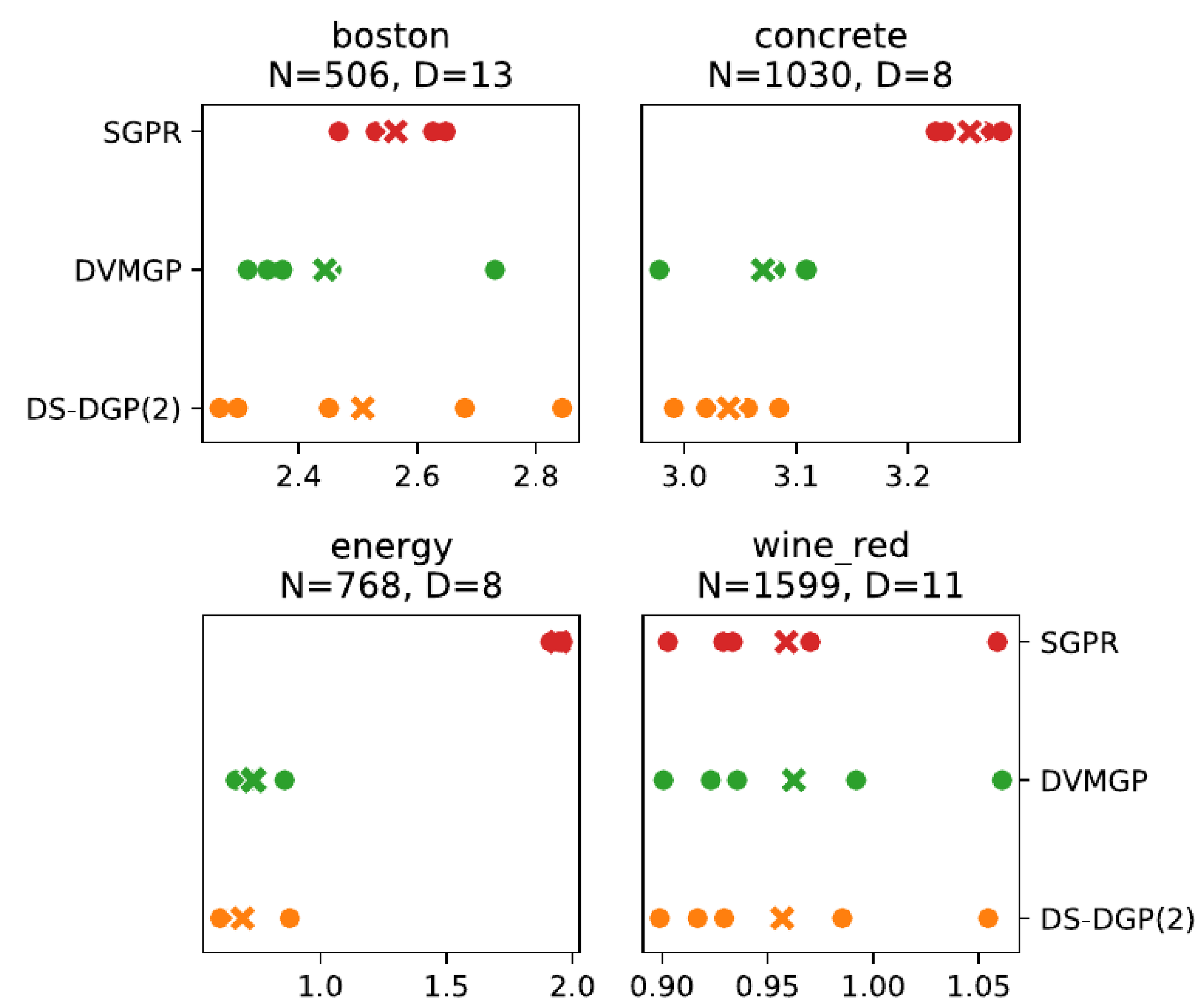


Results

Synthetic non-stationary



UCI Regression



Deep Mahalanobis Gaussian Process

Daniel Augusto de Souza, Diego Mesquita, César Lincoln, João Paulo Pordeus

TL;DR

- In this work, we propose a new **non-stationary** GP model, focusing on greater interpretability;
- Our model models the lengthscales of the non-stationary kernel as a field of locally linear transformations;
- We compare this model with the traditional compositional deep GP in synthetic and popular regression datasets and obtain promising results.

From stationary kernels to non-stationarity

A kernel is stationary if:

$$k(a,b) = \phi\big((a-b)^T \Delta^{-1} (a-b)\big)$$

For a certain $\phi(d)\colon \mathbb{R} \to \mathbb{R}$ and lengthscale matrix Δ. For example, the RBF kernel is the one with $\phi(d) = \exp\left[-\frac{1}{2}d\right]$. As shown by Gibbs (1997), one can build a non-stationary kernel from these building blocks as follows:

$$\sqrt{2^d \frac{\sqrt{\Delta(a)}\sqrt{\Delta(b)}}{|\Delta(a)+\Delta(b)|}}\,\phi\big(2(a-b)^T[\Delta(a)+\Delta(b)]^{-1}(a-b)\big)$$

Monsters around the corner

1. The interpretation of $\Delta(x)$ becomes unclear due to the presence of the orange pre-factor (Gibbs, 1997);
2. The quadratic term inside ϕ usually defines the distance function of a semi-metric space, however, in the non-stationary case, the triangle inequality can be violated (Paciorek, 2003).

An alternative construction for RBF kernels

We propose using a different construction, first by rewriting the stationary kernel definition:

$$\begin{aligned} k(a,b) &= \phi\big((a-b)^T \Delta^{-1} (a-b)\big) \\ &= \phi\big((a-b)^T (W^T W)(a-b)\big) \\ &= \phi\big((Wa-Wb)^T (Wa-Wb)\big) \end{aligned}$$

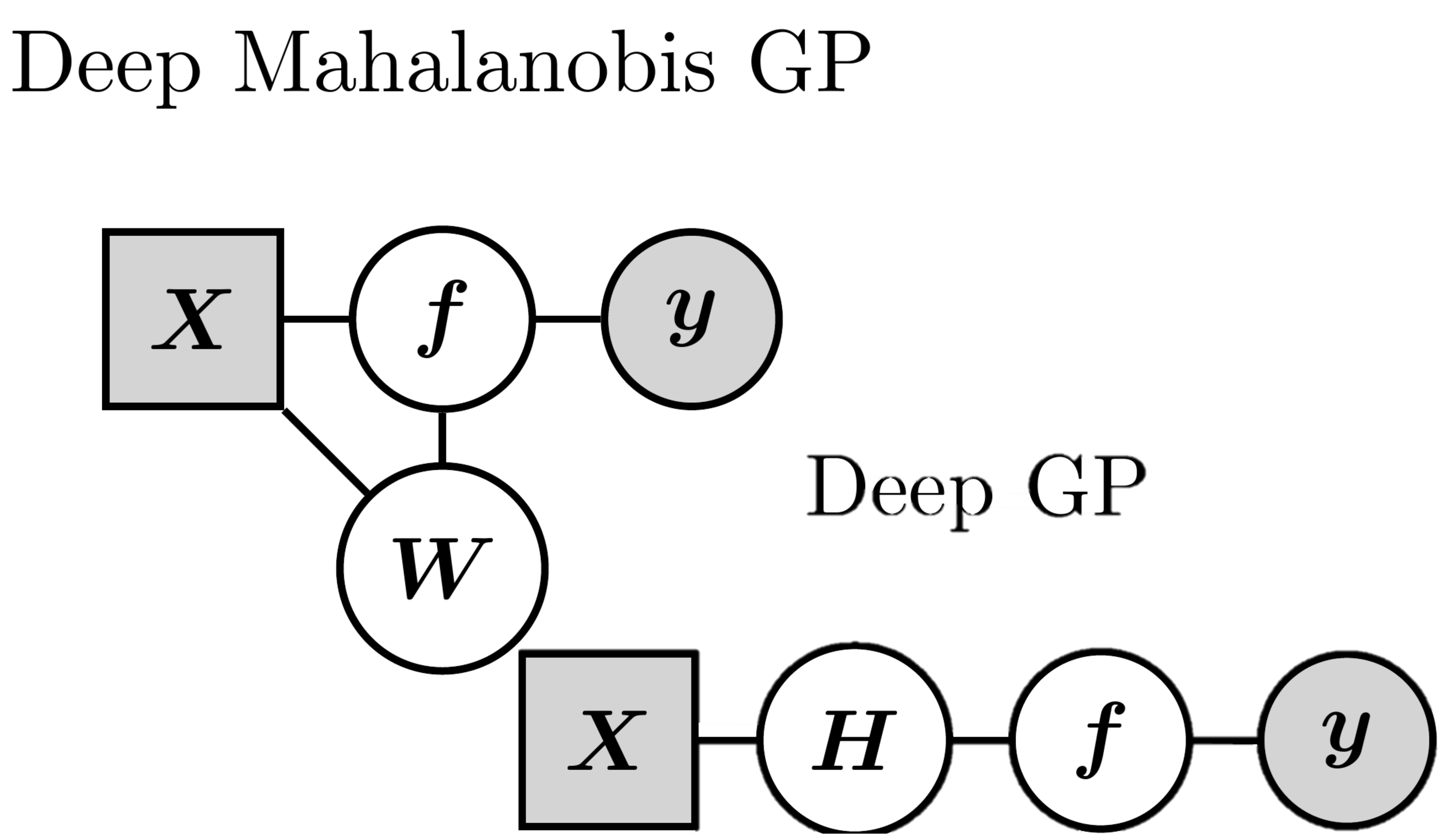
Where we split the positive-definite inverse lengthscale matrix into a product of ordinary matrices. From this point, now we can make W a function of the inputs:

$$\phi\big((W(a)\cdot a - W(b)\cdot b)^T (W(a)\cdot a - W(b)\cdot b)\big)$$

This new kernel is still a positive-definite function, doesn't contain the pre-factor from Gibbs and respects the triangle inequality.

Deep Mahalanobis GP

By placing a Gaussian process prior on the entries of W, we get a deep model that maintains each GP layer still grounded on the input data.



Results