Circular Pseudo-Point Approximations for Scaling Gaussian Processes

UNIVERSITY OF CAMBRIDGE

INVENIA LABS

Will Tebbutt, Thang Bui, Richard E. Turner

will.tebbutt@invenialabs.co.uk, {tdb40, ret26}@cam.ac.uk

Problem Definition

- Gaussian Processes (GPs) are useful regression models with an infinite number of parameters.
- Zero-mean GP marginal likelihood is

$$\mathcal{N}\left(\mathbf{y} \,|\, 0, K_{D,D} + \beta^{-1}\mathcal{I}\right)$$

where $(K_{D,D})_{i,j} = k(x_i, x_j)$, β^{-1} = variance of observation noise.
- Computing $K_{D,D}^{-1}$ is $O(N^3)$ operation $\rightarrow$ infeasible for large N.
- Circulant approximation and others exploit special structure in **data input locations / covariance function** to accelerate inference.
- Pseudo-point approximations accelerate inference if **data over-sampled**.
- Possible to get the best of both worlds?

Circulant Approximations

- Requires data on **regular grid** and k **stationary** (translation invariant).
- Learning + inference is $\mathcal{O}(N \log N)$ as $K_{D,D} \approx U \Gamma_D U^{\dagger}$, where U = the (unitary) DFT matrix, Γ_D = diagonal matrix of eigenvalues (see [Gray, 2006]).
- Wide input domain relative to kernel length-scale $\rightarrow$ highly accurate.
- Narrow input domain relative to kernel length-scale $\rightarrow$ circular 'wrap-around' is problematic.

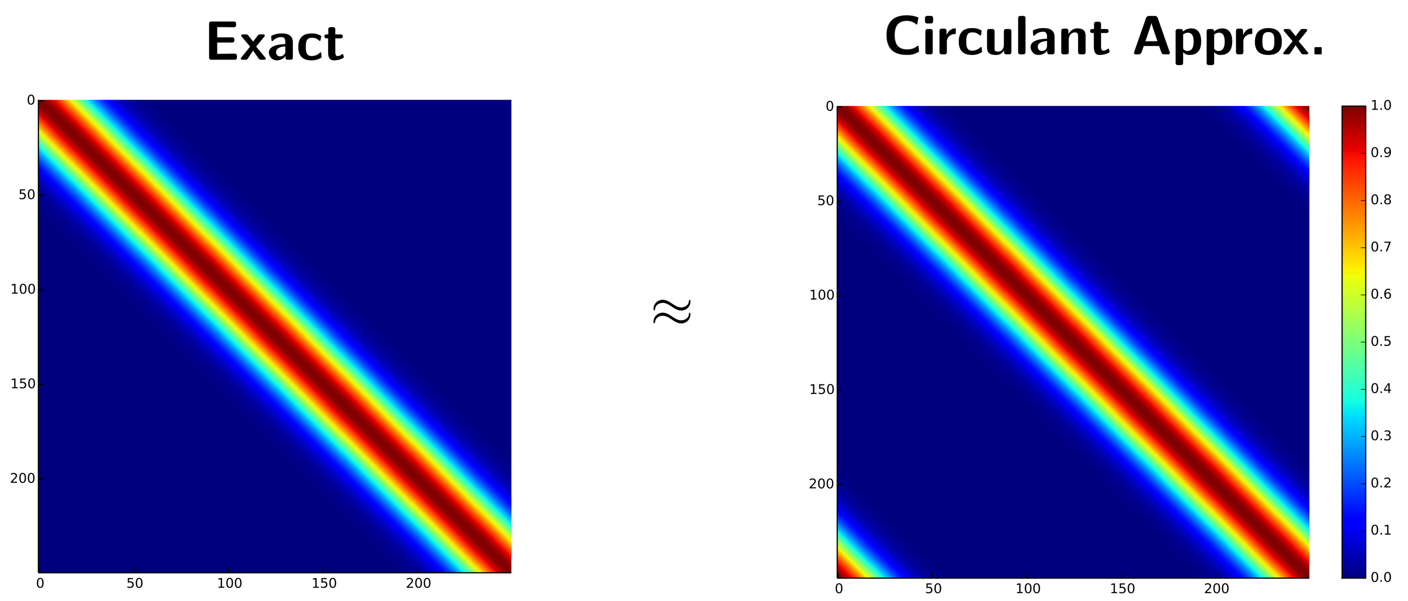


Figure 2: Visualisation of the circulant approximation to an EQ covariance matrix with lengthscale 0.05, computed between data spaced uniformly on $[-0.25, 0.25]$.

Pseudo-Point Approximations

- Arguably, state-of-the-art is VFE approximation [Titsias, 2009].
- N observations, M pseudo-points $\rightarrow K_{D,Z} \in \mathbb{R}^{N \times M}$, $K_{Z,Z} \in \mathbb{R}^{M \times M}$.
- Optimal posterior mean μ_q and covariance Σ_q of pseudo-points are found by maximising

$$L = \log \mathcal{N}\left(\mathbf{y} \,|\, K_{D,Z} K_{Z,Z}^{-1} \mu_q, \beta^{-1}\mathcal{I}\right) - \beta \,\mathrm{tr}(\hat{K}_{D,D} + R_{D,D}) \,/\, 2 \\ - \mathcal{KL}[\mathcal{N}(f_Z \,|\, \mu_q, \Sigma_q) \,||\, \mathcal{N}(f_Z \,|\, 0, K_{Z,Z})],$$

$$\hat{K}_{D,D} := K_{D,D} - K_{D,Z} K_{Z,Z}^{-1} K_{Z,D}, \; R_{D,D} := K_{D,Z} K_{Z,Z}^{-1} \Sigma_q K_{Z,Z}^{-1} K_{Z,D}.$$

- Infeasible for large M, but large M required for complex problems.
- Idea: i) Build **special structure into pseudo-data** to accelerate $K_{Z,Z}$ and ii) **restrict form of** Σ_q to render required computations efficient.

Circulant Pseudo-Point Approximation

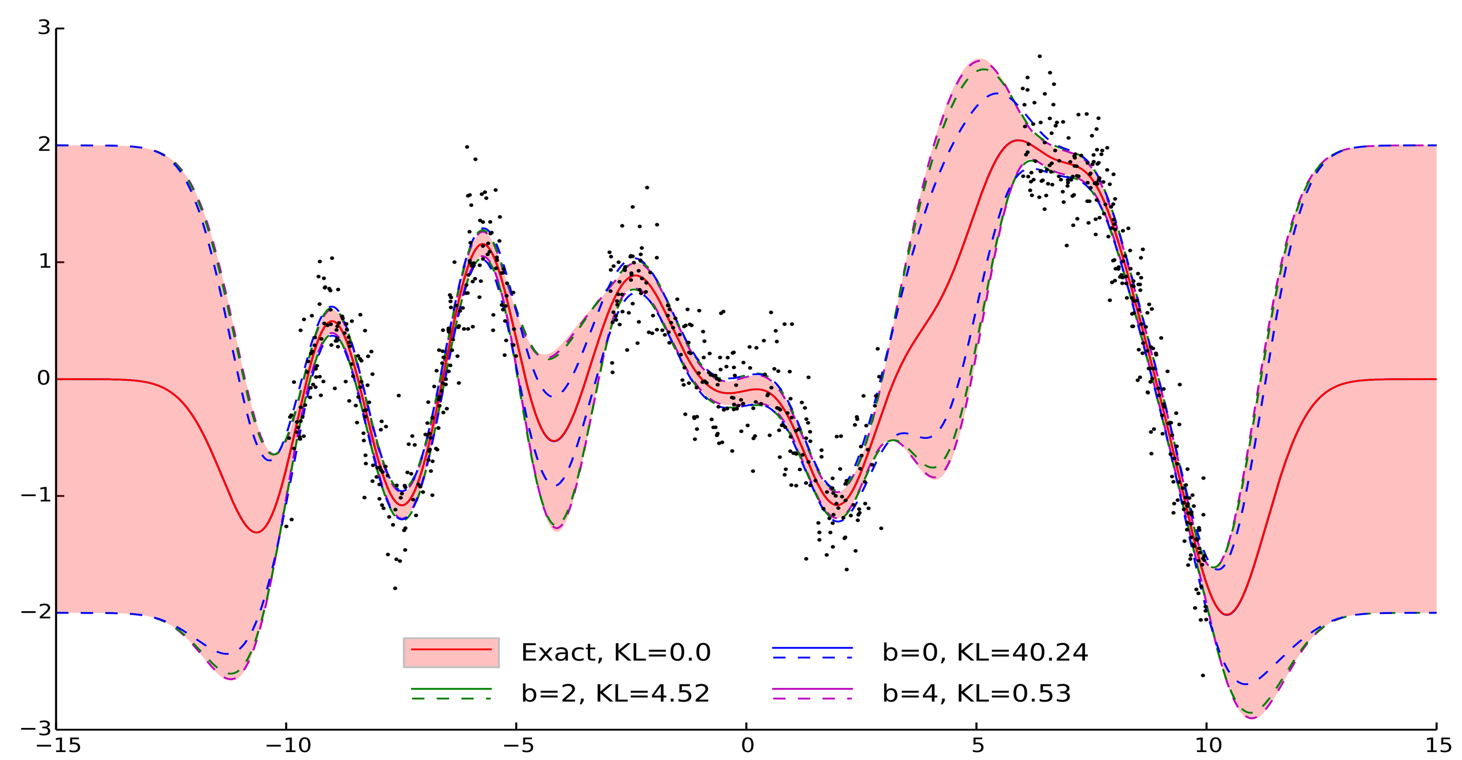


Figure 3: The quality of the approximation depends upon the band-width b of W. $b = 0$ yields a rough approximation, $b = 4$ yields almost exact posterior.

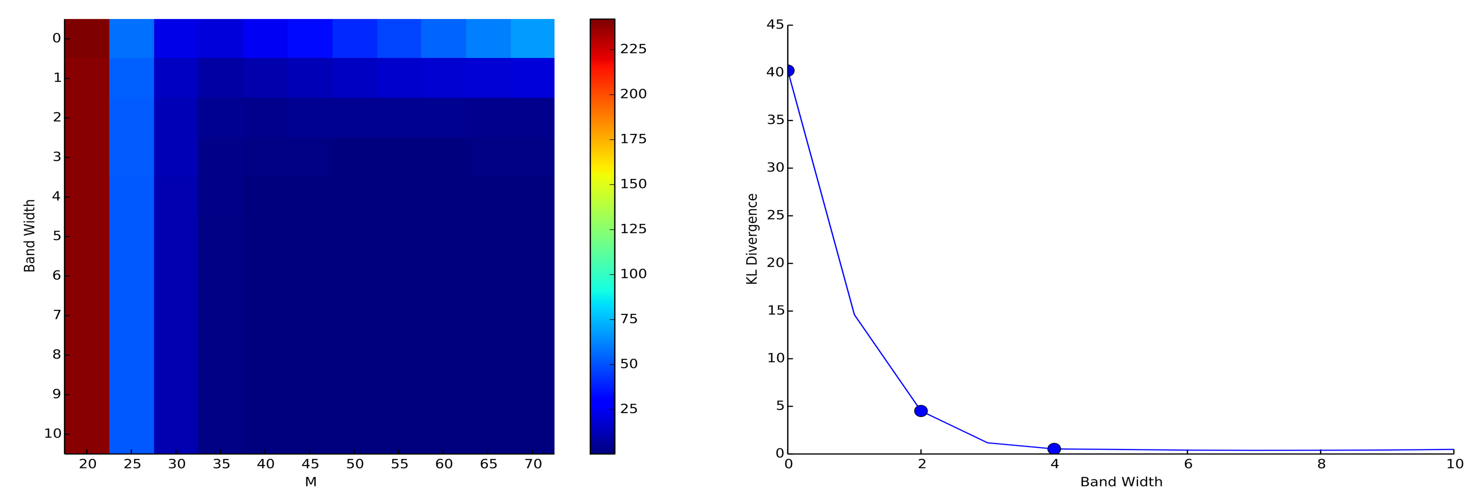
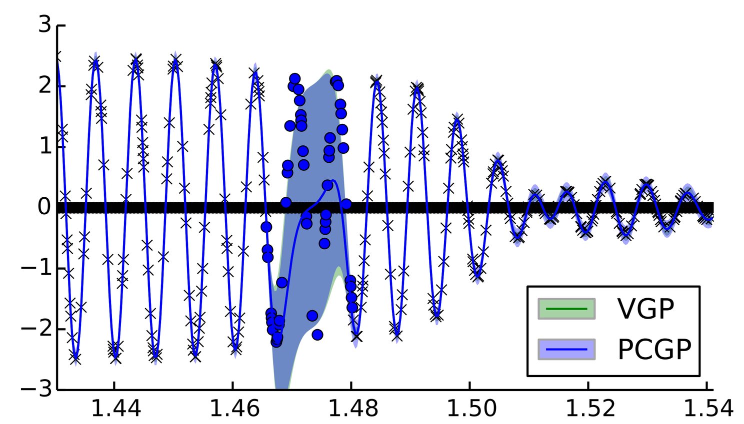


Figure 4: (Left) The KL divergence achieved after 1000 iterations of Adagrad for a range of band widths and numbers of pseudo-data. (Right) The $M = 50$ column of the image on the left, shown for clarity. Highlighted points (circled) correspond to the approximations shown in figure 3.

Circulaunt Pseudo-Point Approximation

- Stationary k and pseudo-points on regular grid $\rightarrow K_{Z,Z} \approx$ circulant.
- Letting pseudo-points extend outside data domain prevents poor posterior prediction at edges of data domain.
- Computing μ_q becomes $\mathcal{O}(NM + M \log M)$ using Linear Conjugate Gradients.
- Parametrise $\Sigma_q = K_{Z,Z}^{\frac{1}{2}} W K_{Z,Z}^{\frac{1}{2}}$ where W **is banded**, band-width b, $K_{Z,Z}^{\frac{1}{2}} := U \Gamma_Z^{\frac{1}{2}} U^{\dagger}$. Gradient w.r.t. W is $\mathcal{O}(NM \log M + Mb)$.



Method	Time (s)	RMSE (in)	RMSE (out)
PCGP	603	9.03×10^{-3}	1.77
VFE	1045	9.02×10^{-3}	1.77

Figure 5: Results on audio sub-band data comprising $N = 20000$ irregularly sampled observations, $M = 10000$ pseudo-points used. 50 observations removed between $t = 1.44$ and $t = 1.46$ are held-out. The reconstruction results are shown in the table. Note despite the narrow band-width $b = 0$, the recovered marginal variances are very similar between the circulant pseudo-point (PCGP) and VFE.

Future Work

- Efficient implementation of operations involving banded matrices.
- Exact solution for W - not straightforward due to banding.
- Exploit approximate circulant structure to represent cross-covariance $K_{D,Z}$ efficiently - crucial to decouple M from N in asymptotic complexity of inference (ie from $\mathcal{O}(NM)$ to $\mathcal{O}(N + M)$) if M grows with N (eg. time series).
- Extend to multi-dimensional input / output domains + non-conjugate likelihoods.

References

[Gray, 2006] Gray, R. M. (2006). *Toeplitz and circulant matrices: A review.* now publishers inc.

[Titsias, 2009] Titsias, M. K. (2009). Variational learning of inducing variables in sparse gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574.