Understanding the properties of sparse Gaussian Process approximations

Will Tebbutt
wct23@cam.ac.uk

Problem Description

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

$$\mathcal{N}\left(\mathbf{y} \,\middle|\, 0, K_{D,D} + \sigma_n^2 \mathcal{I}\right) \tag{1}$$

where $(K_{D,D})_{i,j} = k(x_i, x_j)$, $\sigma_n^2 =$ variance of observation noise.
- Computing $K_{D,D}^{-1}$ is $O(N^3)$ operation $\implies$ infeasible for large N.
- Sparse approximations accelerate inference, $O(NM^2)$, but little work on understanding their properties.
- Analysis directly applicable to regularly-sampled time series. Approximations discussed can also be used to accelerate inference in this case.

Sparse Approximations

State-of-the-art is [Titsias, 2009] - investigation therefore focuses on this.

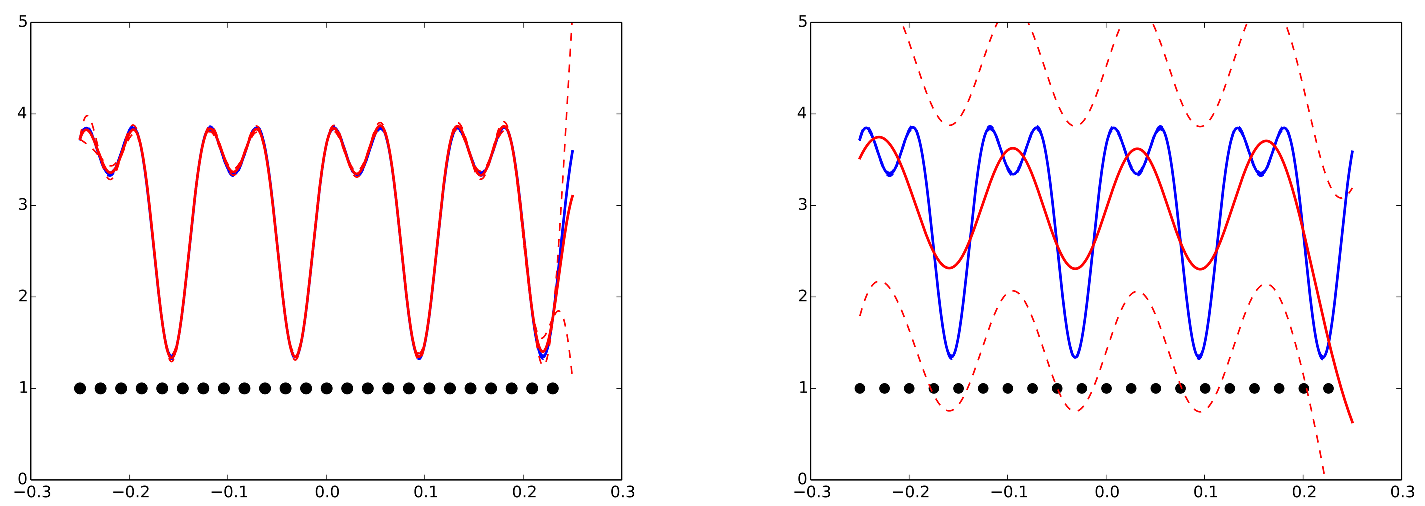

Figure 2: Depiction of speed vs. accuracy trade-off in extreme case. (blue=full GP, red=sparse approx.). (Left: 24 pseudo-data. Right: 20 pseudo data.)

Despite a small change in the number of pseudo-data, a qualitative change in the approximation is observed.

Circulant Approximations to the Covariance Matrix

- If regularly spaced data and stationary k then $K_{D,D}$ is Toeplitz.
- Toeplitz $K_{D,D} \approx$ Circulant, which is easily inverted (see [Gray, 2006]).

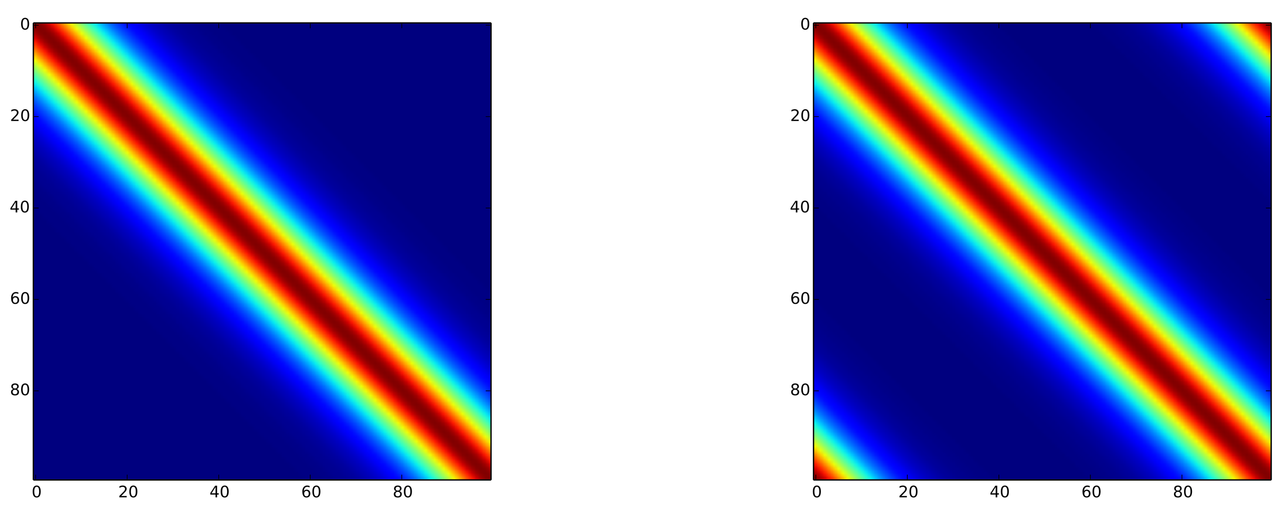

Figure 3: Visualisation of the circulant approximation for an RBF covariance matrix with lengthscale 0.05, computed between data spaced uniformly on $[-0.25, 0.25]$. Left: Exact Toeplitz covariance matrix. Right: Circulant approximation to exact covariance matrix.

Accelerated computations via the Fast Fourier Transform

The posterior mean for a full GP at the observed inputs is

$$m_f = K_{D,D}\left(K_{D,D} + \sigma_n^2 \mathcal{I}\right)^{-1} y. \tag{2}$$

If $K_{D,D}$ is approximately circulant then

$$m_f \approx \mathsf{FT}_D^{\dagger}\left(\Gamma_D + \sigma_n^2 \mathcal{I}\right)^{-1} \mathsf{FT}_D y \tag{3}$$

where the matrix FT_D is the Discrete Fourier Transform (DFT) matrix, $\mathsf{FT}_D^{\dagger}$ is the Inverse DFT matrix and Γ_D is a diagonal matrix whose elements are given by the DFT of the first row of circulant $K_{D,D}$.

Posterior Mean Prediction Error

- Sparse predictive mean $\hat{m}_f$ has same form as full (equation 2).
- Is also approximated as in equation 3. Diagonal of Γ_D truncated to first M elements.
- $$||m_f - \hat{m}_f||_2^2 \approx \sum_{t=M}^{T} |\tilde{y}_t|^2 \left(\frac{\gamma_t}{\gamma_t + \sigma_n^2}\right)^2 \tag{4}$$

where $\tilde{y} := \mathsf{FT}_D y$ and $\{\gamma_t\}_{t=0}^{T-1}$ comprise the diagonal of Γ_D.
- Error a function of lost high-frequency information..
- Sparse approximation accurate if either kernel or data do not contain frequencies higher than those supported by approximation.

Summary and future work

- The properties of sparse approximations can be highly sensitive to the number of pseudo-data.
- Under certain conditions a simple expression is available for the accuracy of a sparse-approximation.
- More experimental validation to be undertaken.

I would like to acknowledge the support of Rich Turner and Thang Bui in the undertaking of this project.

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.