Exploiting Structure for Fast Kernel Learning

Trefor W. Evans and Prasanth B. Nair

University of Toronto

Overview

We propose two methods for exact Gaussian process (GP) inference and learning on massive image, video, spatial-temporal, or multi-output datasets with missing values (or "gaps") in the observed responses. Both of these novel approaches make extensive use of Kronecker matrix algebra to design massively scalable algorithms which have low memory requirements. We demonstrate exact GP inference for a spatial-temporal climate modelling problem with 3.7 million training points as well as a video reconstruction problem with 1 billion points.

Gaussian Processes (GPs)

Specify a zero mean GP prior for the targets, $\mathbf{y}_X \sim \mathcal{N}(\mathbf{0}_N, \mathbf{K}_{X,X} + \sigma^2\mathbf{I}_N)$, the log marginal likelihood is

$$\log \mathcal{P}(\mathbf{y}_X|\boldsymbol{\theta}, \sigma^2, \boldsymbol{\mathcal{X}}_X) = -\tfrac{1}{2}\log|\mathbf{K}_{X,X} + \sigma^2\mathbf{I}_N| - \tfrac{1}{2}\mathbf{y}_X^T(\mathbf{K}_{X,X} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{y}_X - \tfrac{N}{2}\log(2\pi)$$

If we estimate kernel hyperparameters, $\boldsymbol{\theta}$, σ^2, we obtain the following posterior distribution at a test point $\mathbf{x}_* \in \mathbb{R}^d$

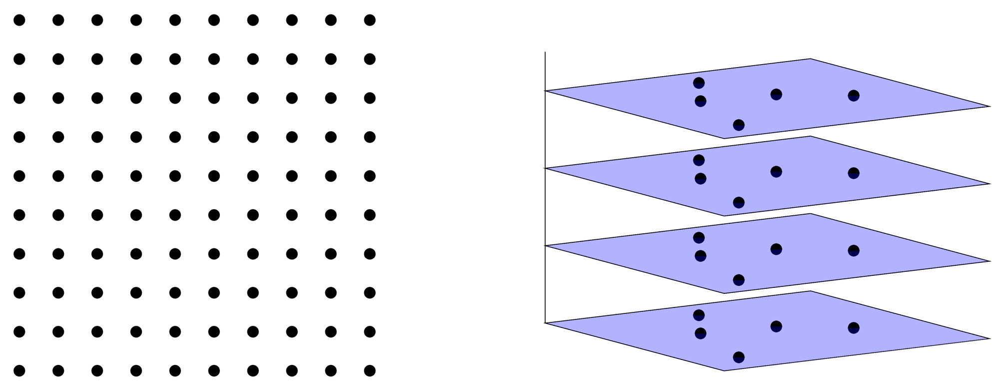
$$y_*|\boldsymbol{\mathcal{X}}_X, \mathbf{x}_* \sim \mathcal{N}(\mathbf{g}_X^T(\mathbf{K}_{X,X} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{y}_X, \;\; k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{g}_X^T(\mathbf{K}_{X,X} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{g}_X)$$

This requires $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ storage!

GPs are typically intractable on large datasets even though their flexibility is most valuable on large scale problems.

Exploiting Structure without Gaps

Consider a regression (or classification) problem where the training data inputs forms a grid. We will call this **structured** data. We can visualize the input distribution of structured data as follows



1. If the data is on a grid (with **no gaps**, $M = N$);
2. and the kernel obeys the product correlation rule, $k(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{d} k_i(x_i, z_i)$,

then the covariance matrix inherits a Kronecker product form

$$\mathbf{K} = \bigotimes_{i=1}^{d} \mathbf{K}_i$$

where $\mathbf{K}_i \in \mathbb{R}^{m \times m}$, $\mathbf{K} \in \mathbb{R}^{M \times M}$ is covariance between grid points, and $m = \sqrt[d]{M}$ is the number of points along each dimension. We can now perform extremely efficient inference by exploiting Kronecker matrix algebra as follows.

Kronecker Matrix Algebra Merits

Storage of K:	$\mathcal{O}(M^2) \to \mathcal{O}(dM^{2/d})$
Matrix-Vector Multiplication with K:	$\mathcal{O}(M^2) \to \mathcal{O}(dM^{(d+1)/d})$
Inverse & Matrix Factorization of K:	$\mathcal{O}(M^3) \to \mathcal{O}(dM^{3/d})$

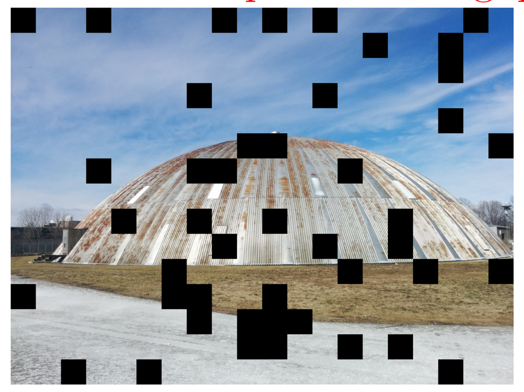
Gaps Destroy Kronecker Product Structure!

In practice, some training data may be missing from the full input grid. These "gaps" may be caused by missing observations or data corruption. Unfortunately, efficient Kronecker matrix algebra can no longer be used in the presence of gaps.

Notation:

$X = \{\mathbf{x}_i\}_{i=1}^{N}$, known response points

$Z = \{\mathbf{x}_i\}_{i=1}^{L}$, missing response points



$\mathbf{K}_{X,X}$ no longer has a Kronecker product form!

Penalize Gaps (PG) Approach (Wilson et al., 2014)

Wilson et al. (2014) approached this problem by using a conjugate gradient solver to find $\boldsymbol{\alpha}$ as follows,

$$\left(\bigotimes_{i=1}^{d} \mathbf{K}_i + \gamma\mathbf{R} + \sigma^2\mathbf{I}_M\right)\boldsymbol{\alpha} = \mathbf{y},$$

which satisfies $\left(\mathbf{K}_{X,X} + \sigma^2\mathbf{I}_N\right)\boldsymbol{\alpha}_X = \mathbf{y}_X$ as the penalty $\gamma \to \infty$. $\mathbf{R} \in \mathbb{R}^{M \times M}$ is all zero except $\mathbf{R}_{Z,Z} = \mathbf{I}_L$, and arbitrary numerical values are inserted in $\mathbf{y}_Z \in \mathbb{R}^L$. However, it is not clear how large the user-defined penalty parameter γ should be *a priori* and given a poor choice, the method will suffer from numerical inaccuracies.

Ignore Gaps (IG) Approach

Applies a selection matrix, $\mathbf{W}$ to $\mathbf{K}$ allowing algebraic computations to be done on the structured $\mathbf{K}$ matrix. Use a conjugate gradient solver to find $\boldsymbol{\alpha}_X$

$$\left(\mathbf{W}\left(\bigotimes_{i=1}^{d} \mathbf{K}_i\right)\mathbf{W}^T + \sigma^2\mathbf{I}_N\right)\boldsymbol{\alpha}_X = \mathbf{y}_X,$$

$\mathbf{W} \in \mathbb{R}^{N \times M}$ is a sparse selection matrix such that $\mathbf{W}\mathbf{K}\mathbf{W}^T = \mathbf{K}_{X,X}$.

- Requires no user-defined parameters (like the PG method), and
- Reduces size of the training problem from $M \times M \to N \times N$

Fill Gaps (FG) Approach

1. **Fill Gaps:** Use a conjugate gradient solver to find $\mathbf{y}_Z$

$$\mathbf{V}\left(\bigotimes_{i=1}^{d} \mathbf{Q}_i\right)\left(\mathbf{T} + \sigma^2\mathbf{I}_M\right)^{-1}\left(\bigotimes_{i=1}^{d} \mathbf{Q}_i^T\right)\mathbf{V}^T\mathbf{y}_Z$$
$$= -\mathbf{V}\left(\bigotimes_{i=1}^{d} \mathbf{Q}_i\right)\left(\mathbf{T} + \sigma^2\mathbf{I}_M\right)^{-1}\left(\bigotimes_{i=1}^{d} \mathbf{Q}_i^T\right)\mathbf{W}^T\mathbf{y}_X,$$

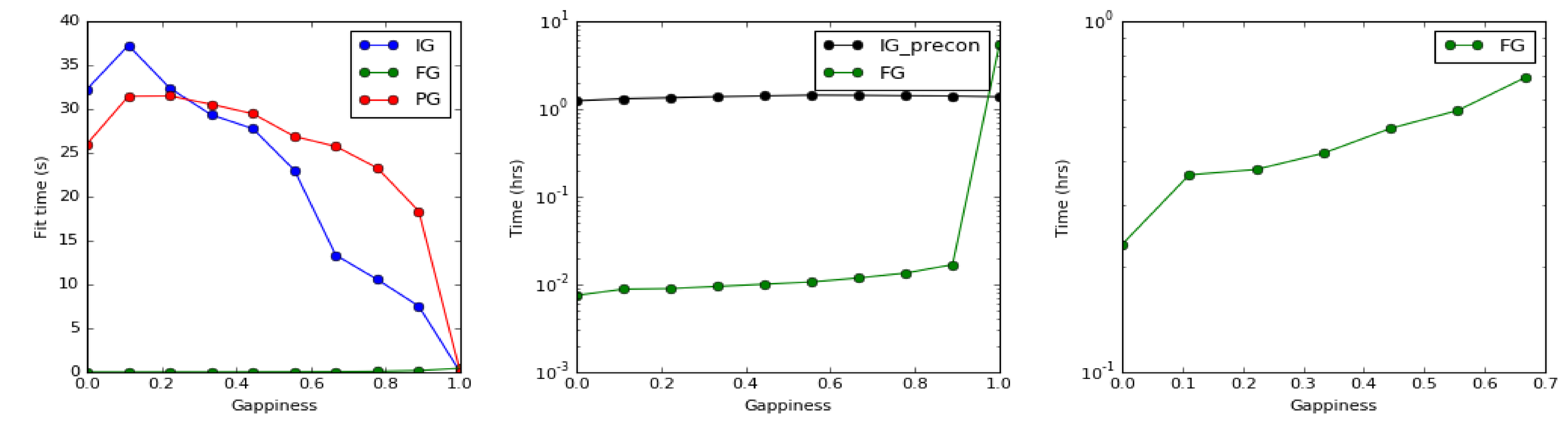
2. **Solve Structured Problem:** compute

$$\boldsymbol{\alpha}_X = \mathbf{W}\left(\bigotimes_{i=1}^{d} \mathbf{Q}_i\right)\left(\mathbf{T} + \sigma^2\mathbf{I}_M\right)^{-1}\left(\bigotimes_{i=1}^{d} \mathbf{Q}_i^T\right)\mathbf{y},$$

$\mathbf{V} \in \mathbb{R}^{L \times L}$ is a sparse selection matrix such that $\mathbf{V}\mathbf{K}\mathbf{V}^T = \mathbf{K}_{Z,Z}$, and $\mathbf{Q}, \mathbf{T} \in \mathbb{R}^{M \times M}$ are the eigenvector and eigenvalue matrices of $\mathbf{K}$, respectively.

- Requires no user-defined parameters (like the PG method), and
- Reduces size of the training problem from $M \times M \to L \times L$

Billion Point Stress Tests



(a) $M = 10,000$ (b) $M = 17$ *million* (c) $M = 1$ *billion*

Figure: Reconstruction timings comparing the training techniques across a range of gappiness on various problem sizes, M, on synthetic video data. Both our approaches are evidently faster and more robust than the existing PG technique which was unsuccessful in the larger studies.

Ontario Climate Modelling

We construct a multi-output GP to model daily minimum and maximum temperatures at 291 Ontario weather stations over 55 years with N =**3,742,547** train points. Both our approaches decreased run-time verses the existing PG technique by more than one order of magnitude.

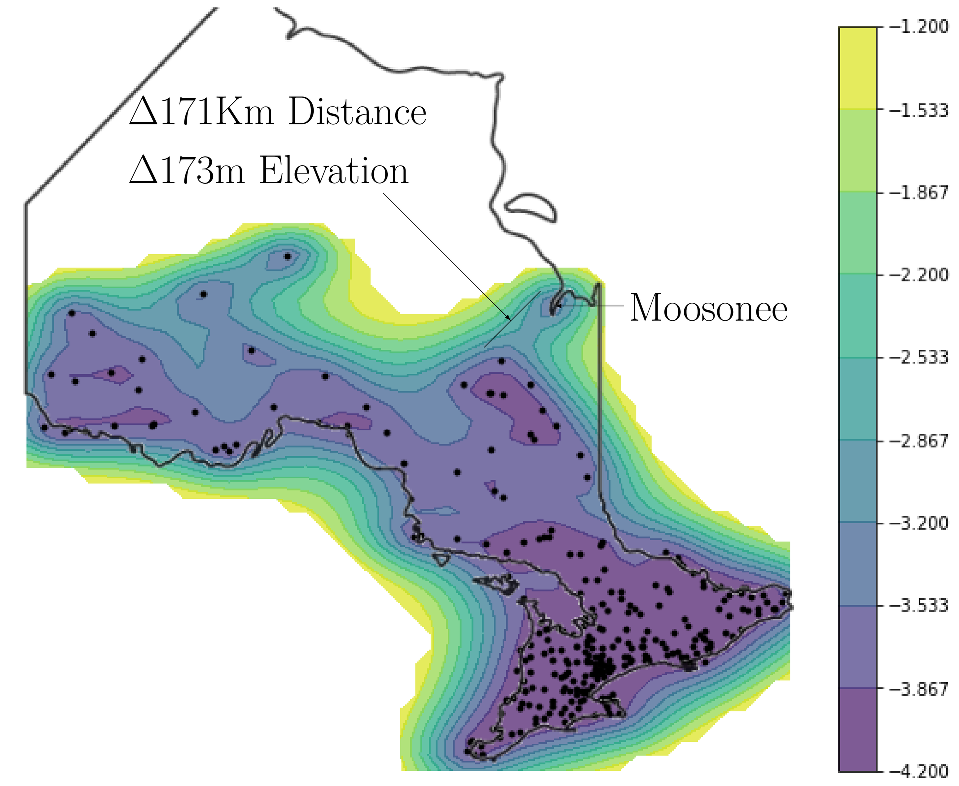


Figure: Log posterior variance of daily temperature ($^\circ C$).

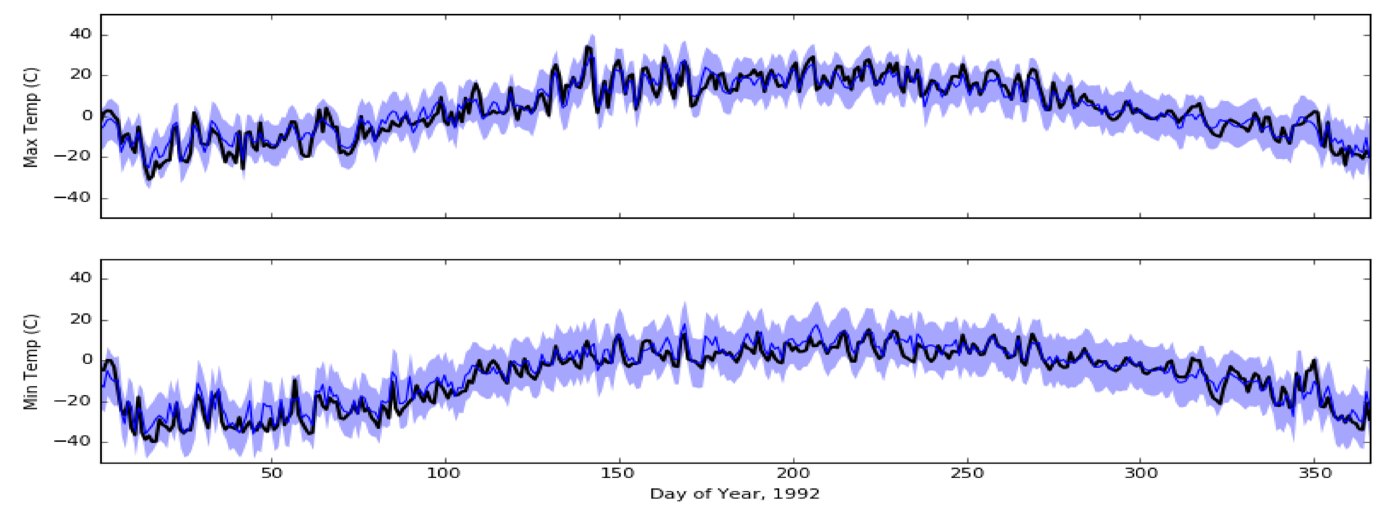


Figure: Reconstructed daily temperature observations for Moosonee in 1992. Actual (black), and posterior mean and 99.7% confidence intervals (blue) are shown.

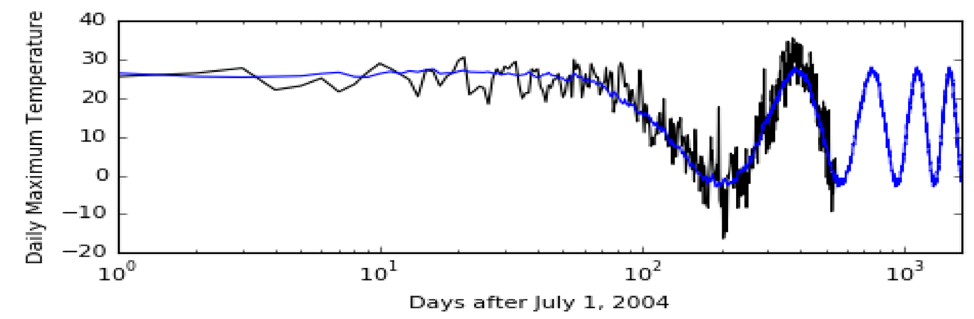


Figure: Forecast of Toronto maximum daily temperature. Actual observations (black) and posterior mean (blue) shown. Toronto training data was removed along with all station data after July 1, 2004.

Acknowledgements: Work funded by an NSERC Discovery Grant and the Canada Research Chairs program.

Wilson, Andrew et al. (2014). "Fast kernel learning for multidimensional pattern extrapolation". In: *Advances in Neural Information Processing Systems*, pp. 3626–3634.