

Exploiting Structure for Fast Kernel Learning

SIAM Data Mining (SDM18)

Trefor W. Evans & Prasanth B. Nair

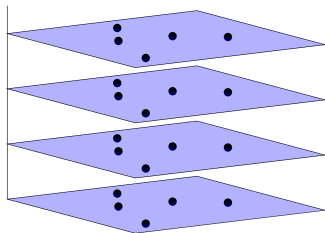
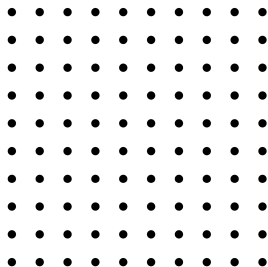
University of Toronto

May 3, 2018

Structured Data

We consider regression (or classification) problems where the training data inputs lie on a grid.

We will call this **structured** data.

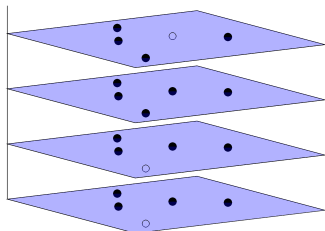
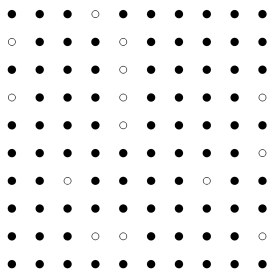


Eg. analysis of images, videos, spatial-temporal fields, sensor networks, or multi-output processes.

Gappy Structured Data

We also consider the practical case where some responses are missing from the structured training set.

We call these missing values **gaps**.



Eg. analysis of images, videos, spatial-temporal fields, sensor networks, or multi-output processes.

Gaussian Processes (GPs)

Given 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

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

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

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

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.

Gaussian Processes (GPs)

Given 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

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

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

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

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

When

- 1 Data is on a grid (with no gaps, $M = N$)
- 2 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

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

$\mathbf{K}_i \in \mathbb{R}^{m \times m}$, $\mathbf{K} \in \mathbb{R}^{M \times M}$ is the covariance between grid points, and $m = \sqrt[d]{M}$ is the number of points along each dimension.

Kronecker Matrix Algebra Merits

Storage of \mathbf{K}

$$\mathcal{O}(M^2) \rightarrow \mathcal{O}(dM^2/d)$$

Matrix-Vector Multiplication with \mathbf{K}

$$\mathcal{O}(M^2) \rightarrow \mathcal{O}(dM^{(d+1)/d})$$

Inverse & Matrix Factorization of \mathbf{K}

$$\mathcal{O}(M^3) \rightarrow \mathcal{O}(dM^3/d)$$

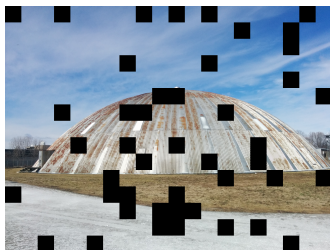
Gappy Structured Data

Training points are still on a grid, however, some responses are missing. Gaps may be caused by missing observations, presence of obstructions or irregular domain boundaries, or data corruption (Gunes et al., 2006; Wilson et al., 2014).

Unfortunately, efficient Kronecker matrix algebra can no longer be used in the presence of these gaps.

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

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



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

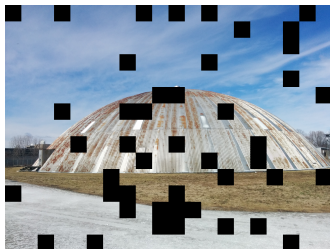
Gappy Structured Data

Training points are still on a grid, however, some responses are missing. Gaps may be caused by missing observations, presence of obstructions or irregular domain boundaries, or data corruption (Gunes et al., 2006; Wilson et al., 2014).

Unfortunately, efficient Kronecker matrix algebra can no longer be used in the presence of these gaps.

$X = \{\mathbf{x}_i\}_{i=1}^N$, 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 (Wilson et al., 2014)

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

Use a conjugate gradient solver to find α

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

which satisfies $(\mathbf{K}_{X,X} + \sigma^2 \mathbf{I}_N) \alpha_X = \mathbf{y}_X$ as the penalty $\gamma \rightarrow \infty$.

$\mathbf{R} \in \mathbb{R}^{M \times M}$ is all zero except $\mathbf{R}_{Z,Z} = \mathbf{I}_L$, and arbitrary numerical values are inserted in the missing entries of $\mathbf{y}_Z \in \mathbb{R}^L$.

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.

Penalize Gaps (Wilson et al., 2014)

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

Use a conjugate gradient solver to find α

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

which satisfies $(\mathbf{K}_{X,X} + \sigma^2 \mathbf{I}_N) \alpha_X = \mathbf{y}_X$ as the penalty $\gamma \rightarrow \infty$.

$\mathbf{R} \in \mathbb{R}^{M \times M}$ is all zero except $\mathbf{R}_{Z,Z} = \mathbf{I}_L$, and arbitrary numerical values are inserted in the missing entries of $\mathbf{y}_Z \in \mathbb{R}^L$.

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 (Our Approach)

Applies a selection matrix, \mathbf{W} to \mathbf{K} allowing algebraic computations to be done on the structured \mathbf{K} matrix.

Ignore Gaps (IG) Formulation

Use a conjugate gradient solver to find α_X

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

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

This method admits fast matrix-vector products like the previous method, and

- requires no user-defined parameters, and
- reduces the linear system size from $M \times M \rightarrow N \times N$

Ignore-Gaps Preconditioner

We develop the following preconditioner for the ignore-gaps (IG) method

$$(\tilde{\mathbf{K}}_{X,X} + \sigma^2 \mathbf{I}_N)^{-1} = \frac{1}{\sigma^2} \left[\mathbf{I}_N - \mathbf{W} \mathbf{Q} \mathbf{S}_p^T (\sigma^2 \mathbf{I}_p + \mathbf{T}_p \mathbf{S}_p \mathbf{Q}^T \mathbf{W}^T \mathbf{W} \mathbf{Q} \mathbf{S}_p^T)^{-1} \mathbf{T}_p \mathbf{S}_p \mathbf{Q}^T \mathbf{W}^T \right],$$

where $\mathbf{S}_p \in \mathbb{R}^{p \times M}$ is a sparse selection matrix such that $\mathbf{S}_p \mathbf{T} \mathbf{S}_p^T = \mathbf{T}_p \in \mathbb{R}^{p \times p}$ is a subset of \mathbf{T} containing the p largest eigenvalues of \mathbf{K} on its diagonal.

This only requires the additional storage and inversion of a matrix of size $p \times p$. After construction, multiplication with this preconditioner costs only $\mathcal{O}(dM^{\frac{d+1}{d}} + p^2)$ time.

Fill Gaps I (Our Approach)

Fill Gaps (FG) Formulation

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

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

- 2 **Solve Structured Problem:** compute

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

which satisfies $(\mathbf{K}_{X,X} + \sigma^2 \mathbf{I}_N) \alpha_X = \mathbf{y}_X$.

Fill Gaps II (Our Approach)

$\mathbf{V} \in \mathbb{R}^{L \times L}$ is a sparse selection matrix such that $\mathbf{VKV}^T = \mathbf{K}_{Z,Z}$, and $\mathbf{Q}, \mathbf{T} \in \mathbb{R}^{M \times M}$ are unitary and diagonal matrices, respectively, formed from the eigen-decomposition of $\mathbf{K} = \mathbf{QTQ}^T$.

This method admits fast matrix-vector products like the PG method, and

- requires no user-defined parameters, and
- reduces the linear system size from $M \times M \rightarrow L \times L$

Stress Testing on Synthetic Video Data

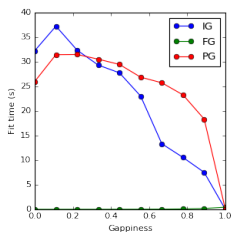
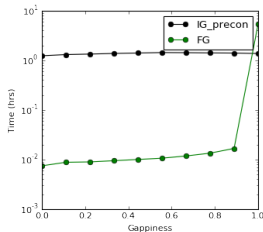
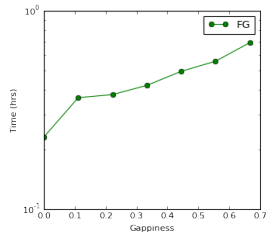
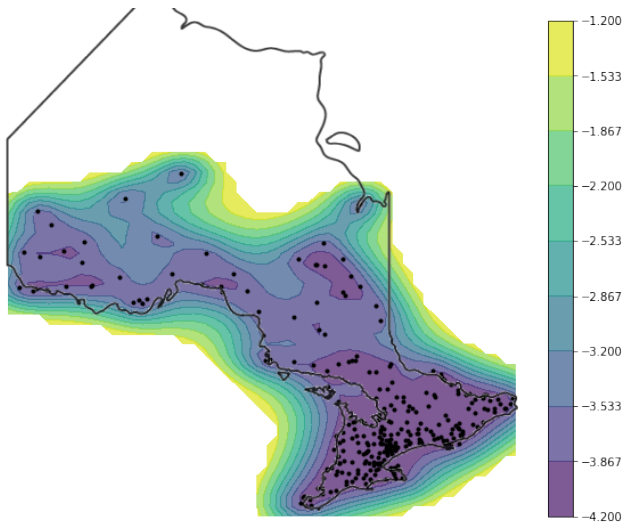
(a) $M = 10,000$ (b) $M = 17 \text{ million}$ (c) $M = 1 \text{ billion}$

Figure: Reconstruction timings comparing the training techniques across a range of gappiness on various problem sizes, M .

The developed algorithms are both faster and more robust than the penalize-gaps (PG) technique.

Ontario Climate Modelling



Log posterior variance of daily temperature ($^{\circ}\text{C}$).

Ontario Climate Modelling

We construct a multi-output GP to model daily minimum and daily maximum temperatures at 291 weather stations over 55 years.

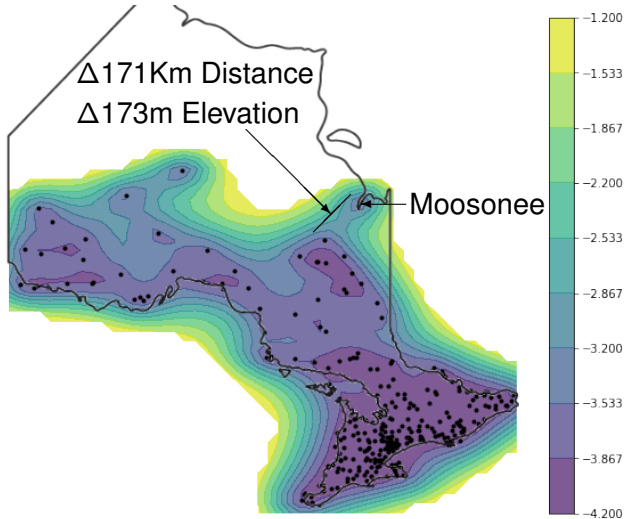
The full grid size is $M=11,928,672$, however, over 6.5 million points are missing and 30% of the remaining points are randomly withheld for testing, giving $N = 3,742,547$ training points; an enormous problem for exact GP modelling.

Ontario Climate Modelling Results

	Run Time (hrs)	RMSE ($^{\circ}\text{C}$)	
		Minimum	Maximum
FG	11.5	2.02	1.45
IG	173.1	2.02	1.45
IG ₁₀₀₀	37.3	2.02	1.45
IG ₃₀₀₀	19.7	2.02	1.45
PG ₁₀₀	221.5	2.02	2.02

Table: Reconstruction time and accuracy of daily maximum and minimum temperatures on the withheld test set using different training techniques for the multi-output GP. PG_# means the penalty $\gamma=\#$ was used and IG_# means a rank $p=\#$ preconditioner was used.

Moosonee Daily Temperature Reconstruction



Log posterior variance of daily temperature ($^{\circ}\text{C}$).

Moosonee Daily Temperature Reconstruction

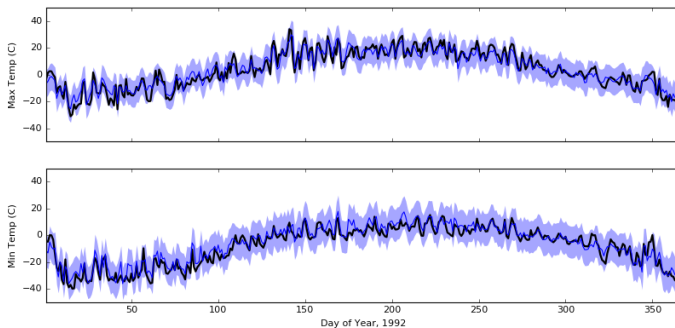


Figure: Reconstructed daily temperature observations for Moosonee in 1992. The black curves show the actual daily maximum (top) and minimum temperatures (bottom) which were both withheld from the model to compute the blue posterior distribution where the mean and three standard deviations (99.7% confidence) are illustrated.

Climate Forecasting

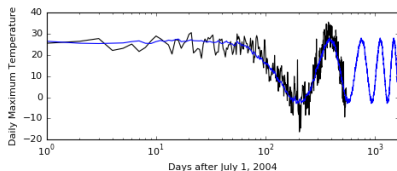


Figure: Forecast of Toronto maximum daily temperature. Actual observations are in black and the posterior mean is in blue. Note that the x-axis is on a log scale. For training, data at the Toronto weather station was withheld for all years and data at all stations after July 1, 2004 was withheld.

Summary

- Developed methods that exploit data structure for fast kernel learning
- Developed methods *faster* and *more robust* than the state-of-the-art
- Novel preconditioner developed that accelerates performance greatly
- Proposed method to infer posterior mean on gaps before training
- Exact GP modelling demonstrated on problems with one-billion points

Thank you



Code available at

https://github.com/treforevans/gp_grid

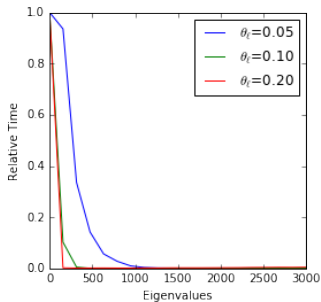
Tutorials coming soon!

Email: trefor.evans@mail.utoronto.ca

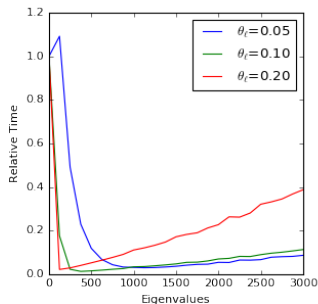
References I

-  Gunes, Hasan, Sirod Sirisup, and George Em Karniadakis (2006). “Gappy data: To Krig or not to Krig?” In: *Journal of Computational Physics* 212.1, pp. 358–382.
-  Wilson, Andrew et al. (2014). “Fast kernel learning for multidimensional pattern extrapolation”. In: *Advances in Neural Information Processing Systems*, pp. 3626–3634.

Preconditioner Experiments



(a) CG Time



(b) Setup & CG Time.

Figure: Results of a preconditioner efficacy study. We use a CG solver to find α_X considering varying values of p , and kernel lengthscales, θ_ℓ . The timings in figure 4a only include the time required to solve the linear system using a CG solver while figure 4b also includes the time required to construct the preconditioner. All timings are presented relative to the time to perform the reconstruction with no preconditioner. We use $M = 10,000$.

Estimated Temporal Kernel

