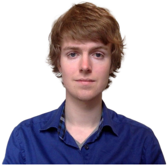
Scalable Exact Inference in Multi-Output Gaussian Processes

Wessel P. Bruinsma[1,2], Eric Perim[2], Will Tebbutt[1], J. Scott Hosking[3,4], Arno Solin[5], Richard E. Turner[1,6]

[1]University of Cambridge, [2]Invenia Labs, [3]British Antarctic Survey, [4]Alan Turing Institute, [5]Aalto University, [6]Microsoft Research

International Conference on Machine Learning 2020

Collaborators



Wessel P. Bruinsma

Eric Perim

Will Tebbutt

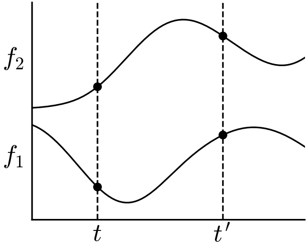
J. Scott Hoskings

Arno Solin

Richard E. Turner

Introduction and Motivation

- Gaussian processes are a powerful and popular probabilistic modelling framework for nonlinear functions.



Central modelling choice:

$$\mathbf{K}(t,t') = \begin{bmatrix} \mathrm{cov}(f_1(t), f_1(t')) & \mathrm{cov}(f_1(t), f_2(t')) \\ \mathrm{cov}(f_2(t), f_1(t')) & \mathrm{cov}(f_2(t), f_2(t')) \end{bmatrix}$$

- Inference and learning: $O(n^3p^3)$ time and $O(n^2p^2)$ memory.
- Often alleviated by exploiting structure in $\mathbf{K}$.

number of outputs

Instantaneous Linear Mixing Model (ILMM)

$$\mathbf{K}(t,t) = \mathbf{I}_m$$

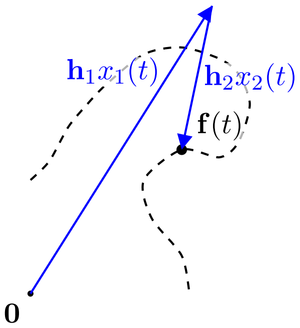
$$\mathbf{x} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}(t,t')),$$

$$\mathbf{f}(t) = \mathbf{h}_1 x_1(t) + \mathbf{h}_2 x_2(t) = \mathbf{H}\mathbf{x}(t),$$

$$\mathbf{y}(t) \sim \mathcal{N}(\mathbf{f}(t), \mathbf{\Sigma}),$$

$\mathbf{x}$: "latent processes",

$\mathbf{H}$: "basis" or "mixing matrix".



- Use $m \ll p$ basis vectors: data lives in "pancake" around $\mathrm{col}(\mathbf{H})$.
- Generalisation of FA to time series setting.
- Captures many existing MOGPs from literature.
- Inference and learning: $O(m^3 n^3)$ instead of $O(p^3 n^3)$.

Inside the ILMM

Key Result

high-dim. observation

$$p\left\{\begin{bmatrix}\bullet\\\bullet\\\vdots\\\bullet\\\bullet\end{bmatrix}\right. \qquad \longmapsto \qquad \left.\begin{bmatrix}\bullet\\\vdots\\\bullet\end{bmatrix}\right\} m\ (\ll p)$$

"projected observation" for $\mathbf{x} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}(t, t'))$

$\mathbf{y}$ $\qquad$ $\mathbf{y}_{\mathsf{proj}} = \mathbf{T}\mathbf{y}$

✗ inference in $p(\mathbf{y})$ $\qquad$ ✓ inference in $p(\mathbf{x})$

noise: $\boldsymbol{\Sigma}$ $\qquad$ projected noise: $\boldsymbol{\Sigma}_{\mathbf{T}}$

Proposition: This is exact!

Key Result (2)

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
\mathbf{Y} \arrow[r, "\text{inference}", "\text{✗ } O(n^3p^3)"'] \arrow[d, "\text{projection}"', "O(nmp)"] & p(\mathbf{f} \mid \mathbf{Y}) \\
\mathbf{TY} \arrow[r, "\text{inference}", "\checkmark\ O(n^3m^3)"'] & p(\mathbf{x} \mid \mathbf{TY}) \arrow[u, "\text{reconstruction}", "O(nmp)"']
\end{tikzcd}
$$

Key Result (3)

$$\log p(\mathbf{Y}) = \overbrace{\log \int p(\mathbf{x}) \prod_{i=1}^{n} \mathcal{N}(\mathbf{T}\mathbf{y}_i \,|\, \mathbf{x}_i, \mathbf{\Sigma}_{\mathbf{T}}) \, \mathrm{d}\mathbf{x}}^{\text{likelihood of projected observations under projected noise}}$$

$$- \underbrace{\frac{1}{2} \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{H}\mathbf{T}\mathbf{y}_i\|_{\mathbf{\Sigma}}^2}_{\text{data ``lost'' by projection (reconstruction error)}} - \underbrace{\frac{1}{2} n \log \frac{|\mathbf{\Sigma}|}{|\mathbf{\Sigma}_{\mathbf{T}}|}}_{\text{noise ``lost'' by projection}} + \text{const.}$$

- Learning $\mathbf{H}$ $\Leftrightarrow$ learning $\mathbf{T}$ $\Leftrightarrow$ learning a transform of the data!
- "Regularisation terms" prevent underfitting.

Key Insight

- Inference in ILMM: condition $\mathbf{x}$ on $\mathbf{Y}_{\text{proj}}$ under noise $\mathbf{\Sigma_T}$.
- Hence,

  if $\mathbf{x}$ are independent under the prior and

  the projected noise $\mathbf{\Sigma_T}$ is diagonal,

  then $\mathbf{x}$ remain independent upon observing data.

$$\downarrow$$

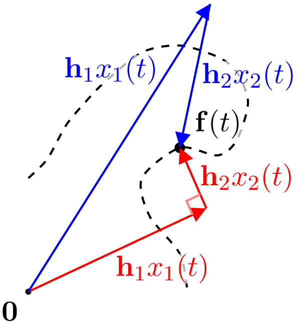
Treat latent processes independently:
condition x_i on $(\mathbf{Y}_{\text{proj}})_{i:}$ under noise $(\mathbf{\Sigma_T})_{ii}$!

- Decouples inference into independent single-output problems.

“Decoupling” the ILMM

Orthogonal ILMM (OILMM)

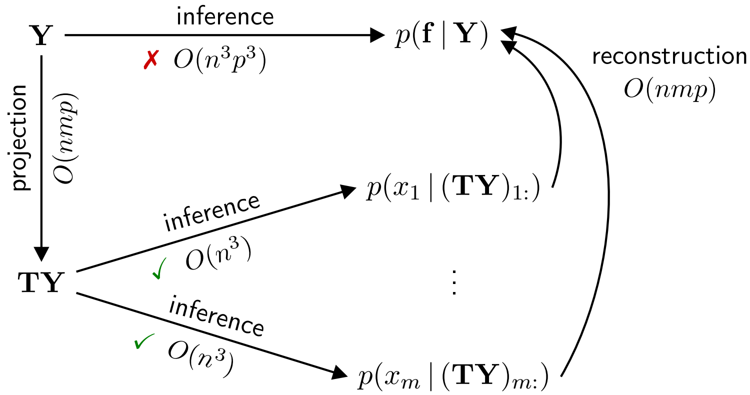
$$\mathbf{x} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}(t, t')),$$
$$\mathbf{f}(t) = \mathbf{H}\mathbf{x}(t)$$
$$= \underset{\text{orthogonal}}{\mathbf{U}}\,\underset{\text{diagonal scaling}}{\mathbf{S}^{\frac{1}{2}}}\,\mathbf{x}(t),$$
$$\mathbf{y}(t) \sim \mathcal{N}(\mathbf{f}(t), \mathbf{\Sigma}).$$



Key property: $\mathbf{\Sigma_T}$ is diagonal!

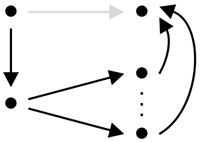
$$\begin{tikzcd}
\text{ILMM} \arrow[r, "\text{orthogonality constraint}"] & \text{OILMM} \\
\text{FA} \arrow[u, "\text{time}", "\text{varying}"'] \arrow[r, "\text{orthogonality constraint}"] & \text{PPCA} \arrow[u, "\text{time}", "\text{varying}"']
\end{tikzcd}$$

Benefits of Orthogonality



- Linear scaling in m!
- Trivially compatible with single-output scaling techniques!

Benefits of Orthogonality (2)



1. Project data and compute proj. noise:

$$\mathbf{Y}_{\mathsf{proj}} = \mathbf{S}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}}\mathbf{Y}, \quad \mathbf{\Sigma}_{\mathbf{T}} = \sigma^{-2}\mathbf{S}^{-1} + \mathbf{D}.$$

2. For $i = 1, \ldots, m$,

   compute the log-probability LML_i of $(\mathbf{Y}_{\mathsf{proj}})_{:i}$ under latent process x_i and observation noise $(\mathbf{\Sigma}_{\mathbf{T}})_{ii}$.

3. Compute the "regularisation term":

$$\text{reg.} = -\frac{n}{2}\log|\mathbf{S}| - \frac{n(p-m)}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\|(\mathbf{I}_p - \mathbf{U}\mathbf{U}^{\mathsf{T}})\mathbf{Y}\|_F^2$$

4. Construct the log-probability of the data Y under the OILMM:

$$\log p(\mathbf{Y}) = \sum_{i=1}^{m} \mathsf{LML}_i + \text{reg.}$$

Complexities of MOGPs

more restrictive ↓ Class	Complexity	
MOGP	$O(p^3 n^3)$	Use single-output scaling techniques to also bring down complexity in n.
ILMM	$O(m^3 n^3)$	
OILMM	$O(mn^3)$ ←	
	$O(mnr^2)$	(r inducing points)
	$O(mnd^3)$	(d-dim. state-space approximation)

Orthogonality gives excellent computational benefits.
But how restrictive is it?

Generality of the OILMM

Definition

An (O)ILMM is separable if $\mathbf{K}(t, t') = k(t, t')\mathbf{I}_m$. Example: ICM.

ILMM versus OILMM:

- Separable case: without loss of generality.
- Non-separable case: only affects correlations through time.
- ILMM can be approximated by an OILMM (in KL) if the right singular vectors of $\mathbf{H}$ are close to unit vectors (in $\|\bullet\|_F$).

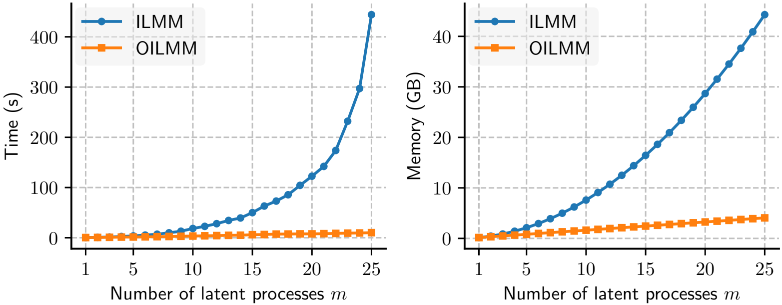
- Separable spatio–temporal GP is an OILMM.
- OILMM gives non-separable relaxation of separable models whilst retaining efficient inference.

Missing Data

- Missing data is troublesome: it breaks orthogonality of $\mathbf{H}$.
- In the paper, we derive a simple and effective approximation.

The OILMM in Practice

Demonstration of Scalability

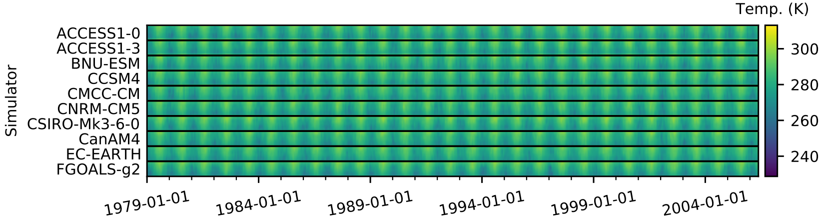


Demonstration of Generality

	EEG		FX	
	PPLP	SMSE	PPLP	SMSE
ILMM	-2.11	0.49	3.39	0.19
OILMM	-2.11	0.49	3.39	0.19

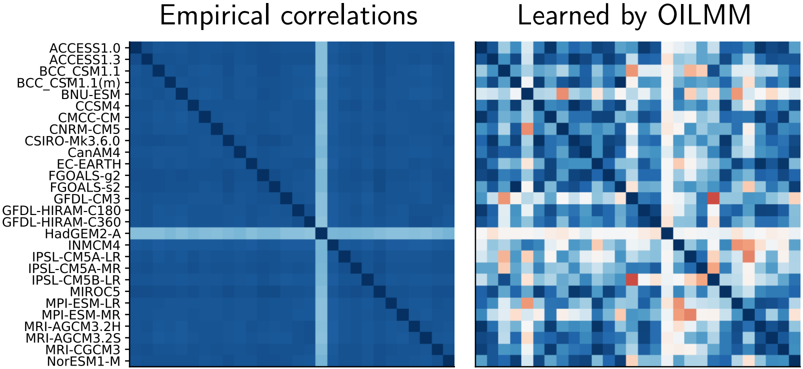
- Near identical performance on two real-world data sets.
- Demonstrates that missing data approximation works well.

Case Study: Climate Simulators



- Jointly model $p_s = 28$ climate simulators at $p_r = 247$ spatial locations and $n = 10\,000$ points in time.
- Equals $p = p_s p_r \approx 7\,\text{k}$ outputs and $pn \approx 70\,\text{M}$ observations.
- Goal: Learn covariance between simulators with $\mathbf{H} = \mathbf{H}_s \otimes \mathbf{H}_r$.
- Use $m = 50$ and inducing points to scale decoupled problems.

Case Study: Climate Simulators (2)



Conclusion

Use projection of the data to accelerate inference in MOGPs with orthogonal bases:

- ✓ Linear scaling in m.
- ✓ Simple to implement.
- ✓ Trivially compatible with single-output scaling techniques.
- ✓ Does not sacrifice significant expressivity.