Combining Pseudo-Point and State Space Approximations for Sum-Separable Gaussian Processes

Will Tebbutt[1] **Arno Solin**[2] **Richard E. Turner**[1]

[1]University of Cambridge, UK
[2]Aalto University, Finland

Abstract

Gaussian processes (GPs) are important probabilistic tools for inference and learning in spatio-temporal modelling problems such as those in climate science and epidemiology. However, existing GP approximations do not simultaneously support large numbers of off-the-grid spatial data-points and long time-series which is a hallmark of many applications. Pseudo-point approximations, one of the gold-standard methods for scaling GPs to large data sets, are well suited for handling off-the-grid spatial data. However, they cannot handle long temporal observation horizons effectively reverting to cubic computational scaling in the time dimension. State space GP approximations are well suited to handling temporal data, if the temporal GP prior admits a Markov form, leading to linear complexity in the number of temporal observations, but have a cubic spatial cost and cannot handle off-the-grid spatial data. In this work we show that there is a simple and elegant way to combine pseudo-point methods with the state space GP approximation framework to get the best of both worlds. The approach hinges on a surprising conditional independence property which applies to space–time separable GPs. We demonstrate empirically that the combined approach is more scalable and applicable to a greater range of spatio-temporal problems than either method on its own.

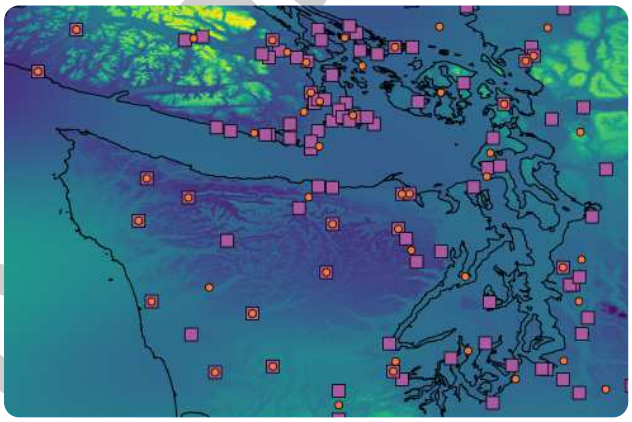


Figure 1: Spatial slice of a large-scale spatio-temporal modelling problem: The posterior mean belief over max temperature (standardised scale, -3 3) on a day in early 2020 around Seattle and Vancouver. Pink squares are weather stations, orange dots are pseudo-points.

1 INTRODUCTION

Large spatio-temporal data containing millions or billions of observations arise in various domains, such as climate science. While Gaussian process (GP) models [Rasmussen and Williams, 2006] can be useful models in such settings, the computational expense of exact inference is typically prohibitive, necessitating approximation. This work combines two classes of approximations with complementary strengths and weaknesses to tackle spatio-temporal problems: pseudo-point [Quiñonero-Candela and Rasmussen, 2005, Bui et al., 2017] and state-space [Särkkä et al., 2013, Särkkä and Solin, 2019] approximations. Fig. 1 shows a single time-slice of a spatio-temporal model for daily maximum temperature, which extrapolates from fixed weather stations, constructed using this technique.

This work hinges on a conditional independence property possessed by separable GPs. This property was identified by O'Hagan [1998], and appears to have gone largely unnoticed within the GP community. This property, in conjunction with the imposition of some structure on the pseudo-point locations, yields a collection of methods for approximate inference algorithm which scale linearly in time, the same as standard pseudo-point methods in space, and which can be implemented straightforwardly by utilising standard Kalman filtering-like algorithms.

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

In particular, we show *(i)* how O'Hagan's conditional independence property can be exploited to significantly accelerate the variational inference scheme of Titsias [2009] for GPs with separable and sum-separable kernels, *(ii)* how this can be straightforwardly combined with the Markov property exploited by state space approximations [Särkkä and Solin, 2019] to obtain an accurate approximate inference algorithm for sum-separable spatio-temporal GPs that scales linearly in time, and *(iii)* how the earlier work of Hartikainen et al. [2011] on this topic is more closely related to the pseudo-point work of Csató and Opper [2002] and Snelson and Ghahramani [2005] than previously realised.

2 SUM-SEPARABLE SPATIO-TEMPORAL GPS

We call a GP *separable across space and time* if its kernel is of the form

$$\kappa((\mathbf{r},\tau),(\mathbf{r}',\tau')) = \kappa^{\mathbf{r}}(\mathbf{r},\mathbf{r}')\,\kappa^{\tau}(\tau,\tau') \tag{1}$$

where $\mathbf{r},\mathbf{r}' \in \mathcal{X}$ are spatial inputs and $\tau,\tau' \in \mathbb{R}$ are temporal inputs. We also call kernels such as κ separable. There is no particular restriction on what we define $\mathcal{X}$ to be – it could be 3-dimensional Euclidean space in the literal sense, or it could be something else, such as a graph or the surface of a sphere. Moreover, we place no restrictions on the form of $\kappa^{\mathbf{r}}$, in particular we do not require it to be separable. Similarly, while the temporal inputs must be in $\mathbb{R}$, it is irrelevant whether this dimension actually corresponds to time or to something else entirely.

This work considers a generalisation of separable GPs that we call *sum-separable across space and time*, or simply *sum-separable*. We call a GP *sum-separable* if it can be sampled by summing samples from a collection of independent separable GPs. Specifically, let $f_p \sim \mathcal{GP}(0,\kappa_p)$, $p = \{1,...,P\}$, be a collection of P independent separable GPs with kernels κ_p, and $f := \sum_{p=1}^{P} f_p$, then f is sum-separable. f has kernel

$$\kappa((\mathbf{r},\tau),(\mathbf{r}',\tau')) = \sum\nolimits_{p=1}^{P} \kappa_p((\mathbf{r},\tau),(\mathbf{r}',\tau'))\,, \tag{2}$$

which is *not* separable, meaning that sum-separable GPs such as f are not generally separable. In fact they are a much more expressive family of models, as they can represent processes which vary on multiple length scales in space and time. Note that these are also distinct from additive GPs [Duvenaud et al., 2011] since each function depends on both space and time.

3 PSEUDO-POINT APPROXIMATIONS

Pseudo-point approximations tackle the scaling problems of GPs by summarising a complete data set through a much smaller set of carefully-chosen uncertain pseudo-observations.

Consider a GP, $f \sim \mathcal{GP}(m,\kappa)$, of which N observations $\mathbf{y} \in \mathbb{R}^N$ are made at locations $\mathbf{x} \in \mathcal{X}^N$ through observation model $p(\mathbf{y}\,|\,\mathbf{f}) = \prod_{n=1}^{N} p(\mathbf{y}_n\,|\,\mathbf{f}_n)$, $\mathbf{f}_n := f(\mathbf{x}_n)$. The seminal work of Titsias [2009], revisited by Matthews et al. [2016], introduced the following approximation to the posterior distribution over f:

$$q(f) = q(\mathbf{u})\,p(f_{\neq\mathbf{u}}\,|\,\mathbf{u})\,, \tag{3}$$

where $\mathbf{u}_m := f(\mathbf{z}_m)$ are the *pseudo-points* for a collection of M *pseudo-inputs* $\mathbf{z}_{1:M}$, and $f_{\neq\mathbf{u}} := f \setminus \mathbf{u}$ are all of the random variables in f except those used as pseudo-points. We assume that $q(\mathbf{u})$ is Gaussian with mean $\mathbf{m}_{\mathbf{u}}^{\mathrm{q}}$ and covariance matrix $\mathbf{C}_{\mathbf{u}}^{\mathrm{q}}$. Subject to the constraint imposed in Eq. (3), this family contains the optimal choice for $q(\mathbf{u})$ if each observation model $p(\mathbf{y}_n\,|\,f(\mathbf{x}_n))$ is Gaussian; moreover, a Gaussian form for $q(\mathbf{u})$ is the de-facto standard choice when $p(\mathbf{y}_n\,|\,f(\mathbf{x}_n))$ is not Gaussian – see e.g. Hensman et al. [2013]. This choice for $q(\mathbf{u})$ yields the following approximate posterior predictive distribution at any collection of test points $\mathbf{x}_*$

$$\begin{aligned} q(\mathbf{f}_*) &= \mathcal{N}(\mathbf{f}_*;\mathbf{C}_{\mathbf{f}_*\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{m}_{\mathbf{u}}^{\mathrm{q}},\mathbf{C})\,, \\ \mathbf{C} &:= \mathbf{C}_{\mathbf{f}_*} - \mathbf{C}_{\mathbf{f}_*\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{C}_{\mathbf{u}\mathbf{f}_*} + \mathbf{C}_{\mathbf{f}_*\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{C}_{\mathbf{u}}^{\mathrm{q}}\Lambda_{\mathbf{u}}\mathbf{C}_{\mathbf{u}\mathbf{f}_*}\,, \end{aligned} \tag{4}$$

where $\Lambda_{\mathbf{u}} := \mathbf{C}_{\mathbf{u}}^{-1}$ is the inverse of the covariance matrix between all pseudo-points, $\mathbf{C}_{\mathbf{f}_*\mathbf{u}}$ is the cross-covariance between the prediction points and pseudo-points under f, and $\mathbf{m}_{\mathbf{u}}$ and $\mathbf{m}_{\mathbf{f}_*}$ are the mean vectors at the pseudo-points and prediction points respectively. For observation model

$$p(\mathbf{y}\,|\,\mathbf{f}) = \mathcal{N}(\mathbf{y};\mathbf{f},\mathbf{S}) \tag{5}$$

where $\mathbf{S} \in \mathbb{R}^{N\times N}$ is a positive-definite diagonal matrix, it is possible to find the optimal $q(\mathbf{u})$ in closed-form:

$$q(\mathbf{u}) \propto \mathcal{N}(\mathbf{y};\mathbf{C}_{\mathbf{f}\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{u},\mathbf{S})\,\mathcal{N}(\mathbf{u};\mathbf{0},\mathbf{C}_{\mathbf{u}}) \tag{6}$$

and the ELBO at this optimum is also closed-form:

$$\begin{aligned} \mathcal{L} = \log\mathcal{N}(\mathbf{y};\mathbf{m}_{\mathbf{f}},\mathbf{C}_{\mathbf{f}\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{C}_{\mathbf{u}\mathbf{f}}+\mathbf{S}) \\ -\frac{1}{2}\mathrm{tr}\big(\mathbf{S}^{-1}(\mathbf{C}_{\mathbf{f}}-\mathbf{C}_{\mathbf{f}\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{C}_{\mathbf{u}\mathbf{f}})\big)\,, \end{aligned} \tag{7}$$

and is known as the *saturated bound*. It can be computed using only $\mathcal{O}\big(NM^2\big)$ operations using the matrix inversion and determinant lemmas

Related Models The optimal $q(\mathbf{u})$ coincides with the exact posterior distribution over $\mathbf{u}$ under an *approximate model* with observation density $\mathcal{N}(\mathbf{y};\mathbf{C}_{\mathbf{f}\mathbf{u}}\Lambda_{\mathbf{u}}\mathbf{u},\mathbf{S})$, and that the first term in $\mathcal{L}$ is the log marginal likelihood under this approximate model. It is well known that this is precisely the approximation employed by Seeger et al. [2003], known as the *Deterministic Training Conditional* (DTC). Despite their

similarities, the DTC log marginal likelihood and the ELBO typically yield quite different kernel parameters and pseudo-inputs when optimised for – while the pseudo-inputs $\mathbf{z}_{1:M}$ are variational parameters in the variational approximation, and therefore not subject to overfitting (see section 2. of Bui et al. [2017]), they are model parameters in the DTC. For this reason, the variational approximation is widely favoured over the DTC.

However, this close relationship between the variational approximation and the DTC is utilised in Sec. 5 to obtain algorithms which combine pseudo-point and state-space approximations in a manner which is both efficient, and easy to implement.

Benefits and Limitations Pseudo-point approximations perform well when many more observations of a GP are made than are needed to accurately describe its posterior. This is often the case for regression tasks where the inputs are sampled independently. In this case the value of M required to maintain an accurate approximation as N increases generally seems not to grow too quickly—indeed Burt et al. [2019] showed that if the inputs $\mathbf{x}_n$ are sampled i.i.d. from a Gaussian, then the value of M required scales roughly logarithmically in N. However, Bui and Turner [2014] noted that this is typically not the case for time series problems, where the interval in which the observations live typically grows linearly in N. Indeed Tobar [2019] showed that the number of the pseudo-points per unit time must not drop below a rate analogous to the Nyquist-Shannon rate if an accurate posterior approximation is to be maintained as N grows. Consequently the number of pseudo-points M required to maintain a good approximation must grow linearly in N, so the cost of accurate approximate inference using pseudo-point methods is really $\mathcal{O}(N^3)$ in this case.

4 STATE SPACE APPROXIMATIONS TO SUM-SEPARABLE SPATIO-TEMPORAL GPS

Many time-series GPs can be augmented with additional latent dimensions in such a way that the marginal distribution over the original process is unchanged, but with the highly beneficial property that conditioning on all D dimensions at any point in time renders past and future time points independent [Särkkä and Solin, 2019]. This augmentation is exact for many GPs, in particular the popular half-integer Matérn family, and a good approximation for others, such as those with exponentiated-quadratic kernels. Consequently, for any collection of T points in time, $\tau_1 < \tau_2 < ... < \tau_T$, the augmented GP forms a D-dimensional Gauss-Markov chain, whose transition dynamics are a function of the kernel of the GP. This means that standard algorithms (similar to Kalman filtering) can be utilised to perform inference under Gaussian likelihoods, thus achieving linear scaling in T. This technique can be extended to separable and sum-separable spatio-temporal GPs for rectilinear grids of inputs, the details of which are as follows.

Separable GPs Let $\bar{f}$ be such an augmentation of f such that the distribution over $\bar{f}(\tau, \mathbf{r}, 1)$ is approximately equal to that of $f(\tau, \mathbf{r})$, and conditioning on all latent dimensions renders $\bar{f}$ Markov in τ. $\bar{f}$ is specified implicitly through a linear stochastic differential equation, meaning that inference under Gaussian observations can be performed efficiently via filtering / smoothing in a Linear-Gaussian State Space Model (LGSSM). Let $\bar{\mathbf{f}}_t$ be the collection of random variables in $\bar{f}$ at inputs given by the Cartesian product between the singleton $\{t\}$, N_T arbitrary locations in space $\mathbf{r}_{1:N_T}$, and all of the latent dimensions $\{1, \ldots, D\}$. Let the kernel of f be separable: $\kappa((\mathbf{r}, \tau), (\mathbf{r}', \tau')) = \kappa^{\mathbf{r}}(\mathbf{r}, \mathbf{r}')\,\kappa^{\tau}(\tau, \tau')$. Any collection of finite dimensional marginals $\bar{\mathbf{f}} := \bar{\mathbf{f}}_{1:T}$, each using the same $\mathbf{r}_{1:N_T}$, form an LGSSM with $N_T D$-dimensional state with dynamics

$$\bar{\mathbf{f}}_t \mid \bar{\mathbf{f}}_{t-1} \sim \mathcal{N}\big([\mathbf{I}_{N_T} \otimes \mathbf{A}_t]\,\bar{\mathbf{f}}_{t-1}, \mathbf{C}^{\mathbf{r}}_{\mathbf{f}} \otimes \mathbf{Q}_t\big) \tag{8}$$

$$\mathbf{H}_{ab} := \mathbf{I}_a \otimes \begin{bmatrix} 1 & \mathbf{0}_{1\times b-1} \end{bmatrix} \tag{9}$$

$$\mathbf{f}_t = \mathbf{H}_{N_T D}\,\bar{\mathbf{f}}_t, \tag{10}$$

$$\mathbf{y}_t \mid \mathbf{f}_t \sim \mathcal{N}(\mathbf{f}_t, \mathbf{S}_t) \tag{11}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{A}_t \in \mathbb{R}^{D\times D}$ and $\mathbf{Q}_t \in \mathbb{R}^{D\times D}$ are functions of κ^{τ}, $\mathbf{Q}_t$ is positive definite, $\mathbf{C}^{\mathbf{r}}_{\mathbf{f}}$ is the covariance matrix associated with $\kappa^{\mathbf{r}}$ and $\mathbf{r}_{1:N_T}$, $\mathbf{0}_{p\times q}$ is a $p \times q$ matrix of zeros, $\mathbf{y}_t$ is the block of $\mathbf{y}$ containing the observations at the t^{th} time, and the diagonal matrix $\mathbf{S}_t$ is the on-diagonal block of $\mathbf{S}$ corresponding to $\mathbf{y}_t$. See Solin [2016] for further details about $\mathbf{A}_t$ and $\mathbf{Q}_t$.

Sum-Separable GPs Let f be the sum-separable GP given by summing over $f_p \sim \mathcal{GP}(0, \kappa_p)$. A state space approximation to f is obtained by constructing a D_p-dimensional state space approximation for each f_p, the finite dimensional marginals of which form an LGSSM

$$\bar{\mathbf{f}}^p_t \mid \bar{\mathbf{f}}^p_{t-1} \sim \mathcal{N}\big([\mathbf{I}_{N_T} \otimes \mathbf{A}^p_t]\,\bar{\mathbf{f}}^p_{t-1}, [\mathbf{C}^{\mathbf{r},p}_{\mathbf{f}} \otimes \mathbf{Q}^p_t]\big) \tag{12}$$

$$\mathbf{f}_t = \sum_{p=1}^{P} \mathbf{H}_{N_T D_p} \bar{\mathbf{f}}^p_t \tag{13}$$

where $\mathbf{A}^p_t$, $\mathbf{Q}^p_t$, and $\mathbf{C}^{\mathbf{r},p}_{\mathbf{f}}$ are defined in the same way as above for each f_p, and $\mathbf{y}_t \mid \mathbf{f}_t$ is again given by Eq. (11). This LGSSM has $N_T \sum_{p=1}^{P} D_p$ latent dimensions, increasing the time and memory needed to perform inference when compared to a separable model, and is the price of a more flexible model.

Benefits and Limitations While this formulation truly scales linearly in T it has two clear limitations, *(i)* all locations of observations must lie on a rectilinear time-space grid if any computational gains are to be achieved; and *(ii)* inference scales cubically in N_T, meaning that inference is

rendered infeasible by time or memory constraints if a large number of spatial locations are observed.

5 EXPLOITING SEPARABILITY TO OBTAIN THE BEST OF BOTH WORLDS

We now turn to the main contribution of this work: combining the pseudo-point and state space approximations. The result is an approximation which is applicable to any sum-separable GP whose time kernels can be approximated by a linear SDE. We do this simply by constructing a variational pseudo-point approximation of the state space approximation to the original process. In cases where the state space approximation is exact, this is similar to constructing an inter-domain pseudo-point approximation [Lazaro-Gredilla and Figueiras-Vidal, 2009] to the original process, where some of the pseudo-points are placed in auxiliary dimensions.

In this section we show that by constraining the pseudo-inputs, approximate inference becomes linear in time.

5.1 THE CONDITIONAL INDEPENDENCE STRUCTURE OF SEPARABLE GPS

O'Hagan [1998] showed that a separable GP $f(\mathbf{r}, \tau)$ has the following conditional independence properties:

$$f(\mathbf{r}, \tau) \perp\!\!\!\perp f(\mathbf{r}', \tau') \mid f(\mathbf{r}, \tau')\,, \tag{14}$$
$$f(\mathbf{r}, \tau) \perp\!\!\!\perp f(\mathbf{r}', \tau') \mid f(\mathbf{r}', \tau)\,. \tag{15}$$

These are explained graphically in Fig. 2. It is straightforward to show (see App. A.1) that this property extends to

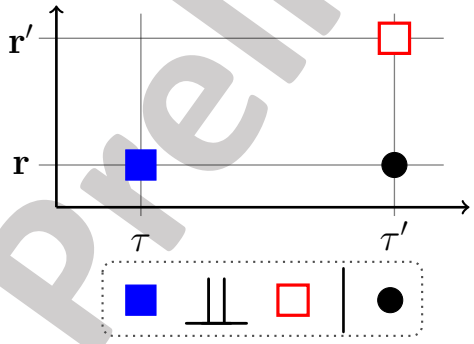


Figure 2: Depiction of the conditional independence property in Eq. (14). The blue square is $f(\mathbf{r}, \tau)$, the red square is $f(\mathbf{r}', \tau')$, and the black circle is $f(\mathbf{r}, \tau')$.

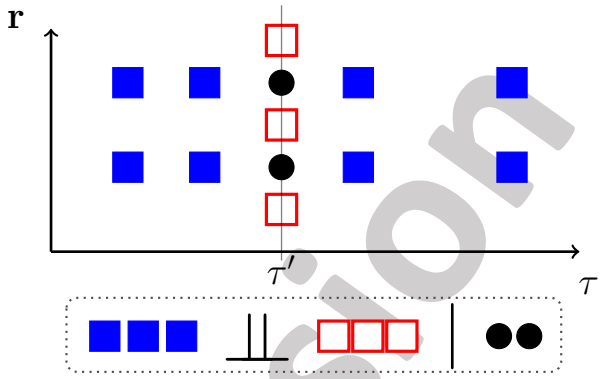


Figure 3: Depiction of the conditional independence property in Eq. (16). The blue squares are $f(\mathcal{R}, \mathcal{T})$, the red squares are $f(\mathcal{R}', \mathcal{T}')$, and the black circles are $f(\mathcal{R}, \mathcal{T}')$.

collections of random variables in f:

$$f(\mathcal{R}, \mathcal{T}) \perp\!\!\!\perp f(\mathcal{R}', \mathcal{T}') \mid f(\mathcal{R}, \mathcal{T}') \quad \text{where} \tag{16}$$
$$f(\mathcal{R}, \mathcal{T}) := \{f(\mathbf{r}, \tau) \mid \mathbf{r} \in \mathcal{R}, \tau \in \mathcal{T}\}$$
$$f(\mathcal{R}', \mathcal{T}') := \{f(\mathbf{r}, \tau') \mid \mathbf{r} \in \mathcal{R}'\}$$
$$f(\mathcal{R}, \mathcal{T}') := \{f(\mathbf{r}, \tau') \mid \mathbf{r} \in \mathcal{R}\}$$

where $\mathcal{R}$ and $\mathcal{R}'$ are sets of points in space, $\mathcal{T}$ is a set of points through time, and $\tau' \in \mathcal{T}$. This conditional independence property is depicted in Fig. 3, and it is this second property that sits at the core of the approximation introduced in the next section.

5.2 COMBINING THE APPROXIMATIONS

We now combine the pseudo-point and state space approximations, and show how a temporal conditional independence property means that the optimal approximate posterior is Markov. This in turn leads to a closed-form expression for the optimum under Gaussian observation models and the existence of a simplified LGSSM in which exact inference yields optimal approximate inference in the original model.

Pseudo-Point Approximation of State Space Augmentation We perform approximate inference in a separable GP f with the kernel in Eq. (1) by applying the standard variational pseudo-point approximation (Sec. 3) to its state space augmentation (Sec. 4) $\bar{f}$:

$$q(\bar{f}) := q(\bar{\mathbf{u}})\, p(\bar{f}_{\neq \bar{\mathbf{u}}} \mid \bar{\mathbf{u}})\,, \quad q(\bar{\mathbf{u}}) = \mathcal{N}(\bar{\mathbf{u}}; \mathbf{m}^{\mathrm{q}}_{\bar{\mathbf{u}}}, \mathbf{C}^{\mathrm{q}}_{\bar{\mathbf{u}}})\,,$$

where the pseudo-points $\bar{\mathbf{u}} = \bar{\mathbf{u}}_{1:T}$ form a rectilinear grid of points in time, space, and *all* of the latent dimensions with the same structure as $\bar{\mathbf{f}}$ in Sec. 4, but replacing $\mathbf{r}_{1:N_T}$ with a collection of M_τ spatial pseudo-inputs, $\mathbf{z}_{1:M_\tau}$, for a total of $TM_\tau D$ pseudo-points. $p(\bar{\mathbf{u}})$ is therefore Markov-through-time with conditional distributions

$$\bar{\mathbf{u}}_t \mid \bar{\mathbf{u}}_{t-1} \sim \mathcal{N}([\mathbf{I}_{M_\tau} \otimes \mathbf{A}_t]\bar{\mathbf{u}}_{t-1}, \mathbf{C}^{\mathbf{r}}_{\mathbf{u}} \otimes \mathbf{Q}_t)\,, \tag{17}$$
$$\mathbf{u}_t := \mathbf{H}_{M_\tau D}\bar{\mathbf{u}}_t. \tag{18}$$

where $\mathbf{C}^{\mathbf{r}}_{\mathbf{u}}$ is the covariance matrix associated with $\kappa^{\mathbf{r}}$ and $\mathbf{z}_{1:M_\tau}$. Note the resemblance to Eq. (8). No constraint is placed on the location of the pseudo-points in space, only that they must remain at the same place for all time points.

Crucially, we now relax the assumption that the inputs associated with $\mathbf{f}$ must form a rectilinear grid. Instead, it is necessary only to require that each observation is made at one of the T times at which we have placed pseudo-points. We denote the number of observations at time t by N_t, and continue to denote by $\mathbf{f}_t$ the set of observations at time t.

Exploiting Conditional Independence Due to O'Hagan [1998]'s conditional independence property, $p\big(\bar{\mathbf{f}}_t \,\big|\, \bar{\mathbf{u}}\big) = p\big(\bar{\mathbf{f}}_t \,\big|\, \mathbf{u}_t\big)$; see App. A for details. Consequently, the reconstruction terms in the ELBO depend only on $\mathbf{u}_t$ as opposed to the entirety of $\bar{\mathbf{u}}$:

$$\mathcal{L} = \sum_{t=1}^{T} r_t - \mathcal{KL}[q(\bar{\mathbf{u}}) \,\|\, p(\bar{\mathbf{u}})]\,, \qquad (19)$$
$$r_t := \mathbb{E}_{q(\mathbf{u}_t)}\left[\mathbb{E}_{p(\mathbf{f}_t \,|\, \mathbf{u}_t)}[\log p(\mathbf{y}_t \,|\, \mathbf{f}_t)]\right]$$

This property alone yields substantial computational savings – only the covariance between $\mathbf{u}_t$ and $\mathbf{f}_t$ need be computed, as opposed to all of $\bar{\mathbf{u}}$ and $\mathbf{f}_t$. Moreover, this means that

$$\mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}}\Lambda_{\bar{\mathbf{u}}} = \begin{bmatrix} \mathbf{B}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{B}_T \end{bmatrix},\ \mathbf{B}_t := \mathbf{C}_{\mathbf{f}_t\mathbf{u}_t}\Lambda_{\mathbf{u}_t}\mathbf{H}_{M_\tau D}. \quad (20)$$

The Optimal Approximate Posterior is Markov As an immediate consequence of Eq. (19), and by the same argument as that made by Seeger [1999], highlighted by Opper and Archambeau [2009], the optimal approximate posterior precision satisfies

$$\Lambda^{\mathrm{q}}_{\bar{\mathbf{u}}} = \Lambda_{\bar{\mathbf{u}}} + \begin{bmatrix} \mathbf{G}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{G}_T \end{bmatrix}, \mathbf{G}_t := -2\nabla_{\mathbf{C}^{\mathrm{q}}_t} r_t. \qquad (21)$$

where $\Lambda^{\mathrm{q}}_{\bar{\mathbf{u}}} := [\mathbf{C}^{\mathrm{q}}_{\bar{\mathbf{u}}}]^{-1}$, and $\mathbf{C}^{\mathrm{q}}_t$ is the t^{th} block on the diagonal of $\mathbf{C}^{\mathrm{q}}_{\bar{\mathbf{u}}}$. Recall that the precision matrix of a Gauss-Markov model is block tridiagonal (see e.g. Grigorievskiy et al. [2017]), so $\Lambda_{\bar{\mathbf{u}}}$ is block tridiagonal. Further, the exact posterior precision of an LGSSM with a Gaussian observation model is given by the sum of this block tridiagonal precision matrix and a block-diagonal matrix with the same block size. $\Lambda^{\mathrm{q}}_{\bar{\mathbf{u}}}$ has precisely this form, so the optimal approximate posterior over $\bar{\mathbf{u}}$ must be a Gauss-Markov chain.

Approximate Inference via Exact Inference in an Approximate Model The above is equivalent to the optimal approximate posterior having density proportional to

$$q(\bar{\mathbf{u}}) \propto \prod_{t=1}^{T} p(\bar{\mathbf{u}}_t \,|\, \bar{\mathbf{u}}_{t-1})\, \mathcal{N}\big(\mathbf{y}^{\mathrm{q}}_t; \bar{\mathbf{u}}_t, \mathbf{G}_t^{-1}\big)\,, \qquad (22)$$

where $\mathbf{y}^{\mathrm{q}}_1, ..., \mathbf{y}^{\mathrm{q}}_T$ are a collection of T surrogate observations, detailed in App. B.1. Thus the optimal $q(\bar{\mathbf{u}})$ is given by exact inference in an LGSSM. Moreover, Ashman et al. [2020] (App. A) show that $\mathbf{G}_t$ can be written as a sum of N_t rank-1 matrices.

Solution for Gaussian Observation Models Under a Gaussian observation model, the optimal approximate posterior is given by the exact posterior under the DTC observation model, as discussed in section Sec. 3. Eq. (20) means that the DTC observation model can be written as

$$\mathcal{N}(\mathbf{y}; \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}}\Lambda_{\bar{\mathbf{u}}}\bar{\mathbf{u}}, \mathbf{S}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{y}_t; \mathbf{B}_t\bar{\mathbf{u}}_t, \mathbf{S}_t)\,. \qquad (23)$$

In conjunction with $p(\bar{\mathbf{u}})$, this yields the required LGSSM.

This LGSSM can be exploited both to perform approximate inference and compute the saturated bound in linear time, repurposing existing code – see App. B.2. This LGSSM also makes it clear, for example, how to employ the parallelised inference procedures proposed by Särkkä and García-Fernández [2020] and Loper et al. [2020] within this approximation.

Sum-Separable Models Extending this approximation to sum-separable processes is similar to the standard state space approximation. The resulting LGSSM is

$$\bar{\mathbf{u}}^p_t \,|\, \bar{\mathbf{u}}^p_{t-1} \sim \mathcal{N}\big([\mathbf{I}_{M_\tau} \otimes \mathbf{A}^p_t]\, \bar{\mathbf{u}}^p_{t-1}, [\mathbf{C}^{\mathbf{r},p}_{\mathbf{u}} \otimes \mathbf{Q}^p_t]\big) \qquad (24)$$
$$p(\mathbf{y}_t \,|\, \bar{\mathbf{u}}_t) = \mathcal{N}(\mathbf{y}_t; \textstyle\sum_{p=1}^{P} \mathbf{B}^p_t \bar{\mathbf{u}}^p_t, \mathbf{S}_t)\,.$$
$$\mathbf{B}^p_t := \mathbf{C}_{\mathbf{f}^p_t\mathbf{u}^p_t}\Lambda_{\mathbf{u}^p_t}\mathbf{H}_{M_\tau D_p}.$$

Note the resemblance to Eq. (12).

Efficient Inference in the Conditionals The structure present in each $\mathbf{B}^p_t$ can be used to accelerate inference. In particular note that $\mathbf{H}_{M_\tau D_p}$ has size $M_\tau \times DM_\tau$ while $\mathbf{C}^p_{\mathbf{f}_t\mathbf{u}_t}\Lambda^p_{\mathbf{u}_t}$ is $N_t \times M_\tau$. Certainly $M_\tau \leq DM_\tau$ and typically $M_\tau < N$, so this linear transformation forms a bottleneck. App. F explores this property, and shows how to exploit it to accelerate inference.

Computational Complexity The total number of flops required to compute the saturated ELBO is $T(DM_\tau)^3 + D^3M^2_\tau + M^2_\tau \sum_{t=1}^{T} N_t$ to leading order. This is a great deal fewer when T is large than the $M^3 + M^2N = M^3_\tau T^3 + M^2_\tau T^2 N$ required if the bound is computed naively. Similar improvements are achieved when making posterior predictions.

Utilising Other Pseudo-Point Approximations The conditional independence property exploited to develop the variational approximation in this section also shines new light on the work of Hartikainen et al. [2011]. In the specific case of their equation 5, in which the observation model is (adopting their notation) $p(\mathbf{y}_k \,|\, \mathbf{x}_k) =$

$\mathcal{N}(\mathbf{y}_k; [\mathbf{I}_N \otimes \mathbf{H}]\mathbf{x}_k, \mathbf{S}_t)$, they perform approximate inference in $p(\bar{\mathbf{u}})$ using the modified observation model

$$\tilde{p}(\mathbf{y}_t \mid \bar{\mathbf{u}}_t) := \mathcal{N}\left(\mathbf{y}_t; \mathbf{C}_{\mathbf{f}_t\bar{\mathbf{u}}_t}\Lambda_{\bar{\mathbf{u}}_t}\bar{\mathbf{u}}_t, [\tilde{\mathbf{C}}_\mathbf{y}]_t\right),$$
$$[\tilde{\mathbf{C}}_\mathbf{y}]_t := \mathrm{diag}(\mathbf{C}_{\mathbf{f}_t} - \mathbf{C}_{\mathbf{f}_t\bar{\mathbf{u}}_t}\Lambda_{\bar{\mathbf{u}}_t}\mathbf{C}_{\bar{\mathbf{u}}_t\mathbf{f}_t}) + \mathbf{S}_t$$

which is inspired by the well-known FITC [Csató and Opper, 2002, Snelson and Ghahramani, 2005] approximation. However, due to O'Hagan [1998]'s conditional independence property, this is equivalent to

$$\tilde{p}(\mathbf{y} \mid \bar{\mathbf{u}}) := \mathcal{N}\left(\mathbf{y}; \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}}\Lambda_{\bar{\mathbf{u}}}\bar{\mathbf{u}}, \tilde{\mathbf{C}}_\mathbf{y}\right),$$
$$\tilde{\mathbf{C}}_\mathbf{y} := \mathrm{diag}(\mathbf{C}_\mathbf{f} - \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}}\Lambda_{\bar{\mathbf{u}}}\mathbf{C}_{\bar{\mathbf{u}}\mathbf{f}}) + \mathbf{S}.$$

While Hartikainen et al. [2011] did not actually consider the Gaussian observation model in their work, it is clear from the above that they would have utilised *exactly* the FITC approximation applied to $\bar{f}$ had they done so.

Bui et al. [2017] showed that both FITC and VFE can be viewed as edge cases of the Power EP algorithm introduced by Minka [2004]. Consequently the equivalent approximate model generalised both that of FITC and VFE – only $\tilde{\mathbf{C}}_\mathbf{y}$ is changed from FITC: let $\alpha \in [0, 1]$, then

$$\tilde{\mathbf{C}}_\mathbf{y} := \alpha\,\mathrm{diag}(\mathbf{C}_\mathbf{f} - \mathbf{C}_{\mathbf{f}\bar{\mathbf{u}}}\Lambda_{\bar{\mathbf{u}}}\mathbf{C}_{\bar{\mathbf{u}}\mathbf{f}}) + \mathbf{S}.$$

In short, most standard pseudo-point approximations can be straightforwardly combined with state space approximations for sum-separable spatio-temporal GPs in the manner that we propose due to the conditional independence property.

Relationship with Other Approximation Techniques There are several existing methods that could be used to scale GPs to large spatio-temporal problems beyond those already considered – each method makes different assumptions about the kinds of problems considered, therefore making different trade-offs relative to ours.

The popular Kronecker-product methods for separable kernels explored by Saatçi [2012] are unable to handle heteroscedastic observation noise or missing data, scale cubically in time, and require observations to lie on a rectilinear grid. Our approach suffers none of these limitations.

Wilson and Nickisch [2015] introduced a pseudo-point approximation they call *Structured Kernel Interpolation* (SKI) which is closely-related to the Kronecker-product methods, but removes many of their constraints. In particular, SKI places pseudo-points on a grid across all input dimensions, and utilises them to construct a sparse approximation to the prior covariance matrix over the data – crucially it is local in the sense that the approximation to the covariance between the pseudo-points and any given point depends only on a handful of pseudo-points. SKI covers the domain in a regular grid of points, which results in exponential growth in

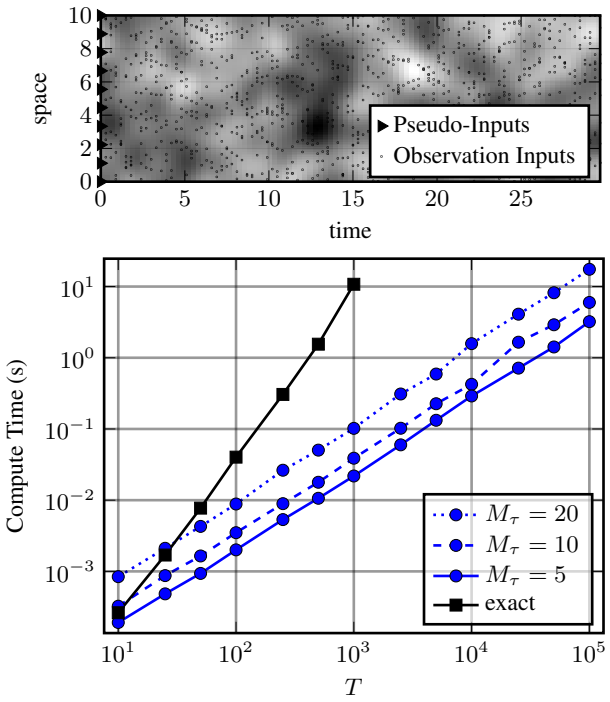


Figure 4: Arbitrary Spatial Locations. Top: Locations of (pseudo-)inputs for $M_\tau = 10$. 10 locations in space chosen randomly at each time point. Bottom: Time to compute ELBO vs performing exact inference. ELBO tight for $M_\tau = 20$; see Fig. 11.

the number of pseudo-points as the number of dimensions grows. So, while this approximation scales very well in low-dimensional settings, it does not scale to input domains comprising more than a few dimensions. Moreover, to exploit this grid structure, separability across all dimensions is required. Gardner et al. [2018] alleviates this exponential scaling problem, but still require that the kernel be separable across all dimensions if their approximation is to be applied. Our approach does not suffer from this constraint as only the time dimension must be covered by pseudo-points – there are no constraints on their spatial locations. Naturally, that we do not perform similar approximations to SKI across the spatial dimensions means that our method will have the standard set of limitations experienced by all pseudo-point methods as the number of points in space grows. In short, the two classes of method are applicable to different kinds of spatio-temporal problems. They take somewhat orthogonal approaches to approximate inference, so combining them by utilising SKI across the spatial dimensions could offer the benefits of both classes of approximation in situations where SKI is applicable to the spatial component.

Similarly, approximations based on the relationship between GPs and Stochastic Partial Differential Equations [Whittle, 1963, Lindgren et al., 2011] could be combined with this work to improve scaling in space when the spatial kernel

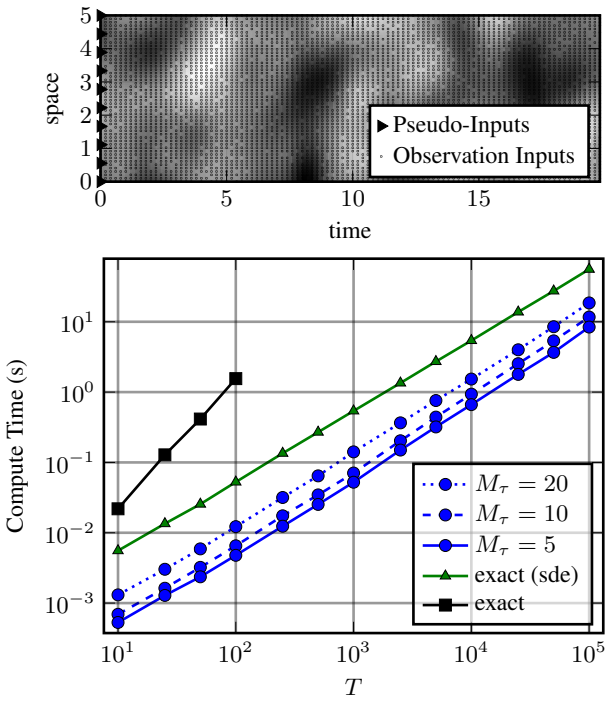


Figure 5: Grid-with-Missings. Top: Locations of (pseudo-)inputs – note the grid structure with 50 observations per time point, of which 5 are missing. Bottom: Time to compute ELBO vs LML naively and via state space methods (*sde*). ELBO tight for $M_\tau = 20$; see Fig. 11.

is in the Matérn family. In low-dimensional settings other standard inter-domain pseudo-point approximations such as those of Hensman et al. [2017], Burt et al. [2020], and Dutordoir et al. [2020] could be applied.

6 EXPERIMENTS

We view the proposed approximation to be a useful contribution if it is able to outperform the vanilla state space approximation (Sec. 4), which is a strong baseline for the tasks we consider. To that end, we benchmark inference against synthetic data in Sec. 6.1, on a large-scale temperature modeling task to which both the vanilla and pseudo-point state space approximations can feasibly be applied (Sec. 6.2), and finally to a problem to which it is completely infeasible to apply the vanilla state space approximation (Sec. 6.3). We do not compare directly against the vanilla pseudo-point approximations of Titsias [2009] and Hensman et al. [2013]. As noted in Sec. 3, they are asymptotically no better than exact inference for problems with long time horizons.

6.1 BENCHMARKING

We first conduct two simple proof-of-concept experiments on synthetic data with a separable GP to verify our proposed method. In both experiments we consider quite a large temporal extent, but only moderate spatial, since we expect the proposed method to perform well in such situations – if the spatial extent of a data set is very large relative to the characteristic spatial variation, pseudo-point methods will struggle and, by extension, so will our method. App. E.1 contains additional details on the setup used, and App. E.1.1 contains the same experiments for a sum-separable model.

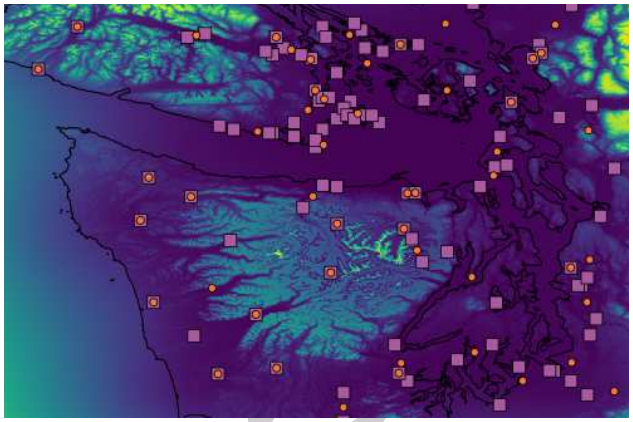


Figure 6: Posterior std. dev. counterpart to Fig. 1. The colour scale (0 1.75) is relative, pink squares are weather stations, and orange dots pseudo-points.

Arbitrary Spatial Locations Fig. 4 (top) shows how inputs were arranged for this experiment; at each time 10 spatial locations were sampled uniformly between 0 and 10, so $N = 10T$. The spatial location of pseudo-inputs are regular between 0 and 10. When using pseudo-points, we are indeed able to achieve substantial performance improvements relative to exact inference by utilising the state space methodology, while retaining a tight bound.

Grid-with-Missings Fig. 5 (top) shows how (pseudo) inputs were arranged for this experiment for $M_\tau = 10$; the same 50 spatial locations are considered at each time point, but 5 of the observations are dropped at random, for a total of $N_t = 45$ observations per time point – our largest case therefore involves $N = 4.5 \times 10^6$ observations. The timing results show that we are able to compute a good approximation to the LML using roughly a third of the computation required by the standard state space approach to inference.

6.2 CLIMATOLOGY DATA

The Global Historical Climatology Network (GHCN) [Menne et al., 2012] comprises daily measurements of a variety of meteorological quantities, going back more than 100 years. We combine this data with the NASA Digital Elevation Model [NASA-JPL, 2020] to model the daily maximum temperature in the region $(47°, -127°)$ and $(49°, -122°)$, which contains 99 weather stations. We utilise all data in this region since the year 2000, training on 90% (331522) and

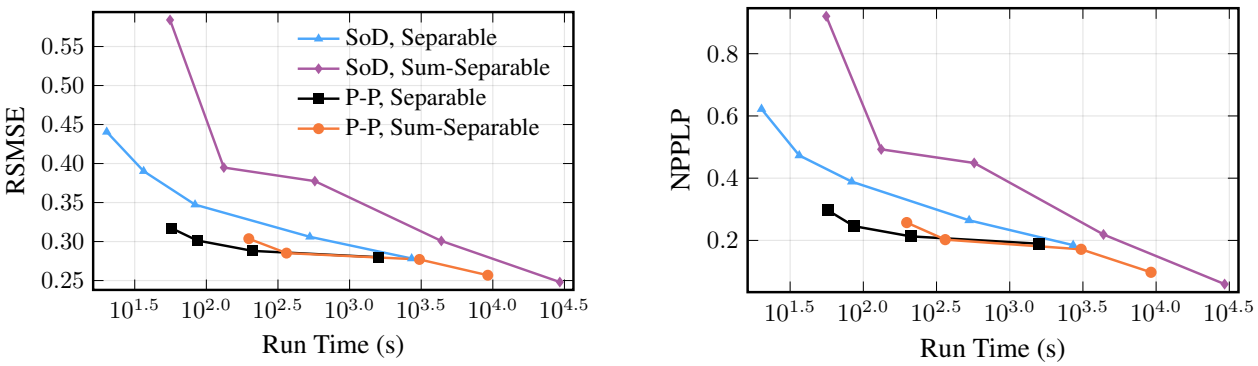
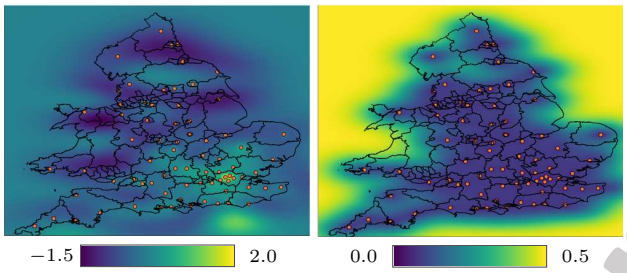


Figure 7: Test Root Standardised Mean-Squared Error (RSMSE) and Negative Posterior Predictive Log Probability (NPPLP). Marked points on Pseudo-Point curves used $M \in \{5, 10, 20, 50\}$ moving from left to right – similarly for SoD markers, with the addition of $M = 99$, corresponding to learning with the exact LML. Larger M improves performance, but time taken to train is increased. Sum-Separable models take longer to train than Separable but can produce better results.



(a) Mean (b) Std. dev.

Figure 8: Apartment price posterior mean and standard deviation on a day near the end of 2020. Pseudo-point locations picked using K-means and marked with orange dots.

testing on 10% (36835) of the data. This experiment was conducted on a workstation with a 3.60 GHz Intel i7-7820X CPU (8 cores), and 46 GB of 3000 MHz DDR3 RAM.

Two models were utilised: a simple separable model with a Matérn-$\frac{5}{2}$ kernel over time, and Exponentiated Quadratic over space, and a sum of two such kernels with differing length scales and variances. Additional details in App. E.2.

Fig. 7 compares a simple subset-of-data (SoD) approximation, which is exact when $M = 99$, with the pseudo-point (P-P) approximation developed in this work. The results demonstrate that *(i)* the pseudo-point approximation has a more favourable speed-accuracy trade-off than the SoD, offering near exact inference in less time for a separable kernel, and *(ii)* a sum-separable model offers substantially improved results over a separable in this scenario.

6.3 APARTMENT PRICE DATA

Property sales data by postcode across England and Wales are provided by HM Land Registry [2014]. There are over 10^6 unique postcodes in England and Wales, of which a tiny proportion contain a sale on a given day. Consequently this data set has essentially arbitrary spatial locations at each point in time, which our approximation can handle, but which renders the vanilla state-space method infeasible.

Table 1: Performance on apartment price data. $M_\tau = 75$.

	RSMSE	NPPLP
Separable	0.658	2920
Sum-Separable	0.618	192

We follow a similar procedure to Hensman et al. [2013], cross-referencing postcodes against a separate database [Camden, 2015] to obtain latitude-longitude coordinates, which we regress against the standardised logarithm of the price. However, we train on 843766 of the 1687536 apartment sales between 2010 and 2020, and test on the remainder. We again consider a separable and sum-separable GP that are similar to those in Sec. 6.2, but the temporal kernel is Matérn-$\frac{3}{2}$. More detail in App. E.3.

Table 1 again demonstrates that a sum-separable model is able to capture more useful structure in the data than the separable model; Fig. 8 shows the variability and uncertainty in the prices on an arbitrarily chosen day.

7 DISCUSSION

This work shows that pseudo-point and state space approximations can be directly combined in the same model to effectively perform approximate inference and learning in sum-separable GPs, and ties up loose ends in the theory related to combining these models. This is important in spatio-temporal applications, where the model admits a form of an arbitrary-dimensional (spatial) random field with dynamics over a long temporal horizon. Experiments on synthetic and real-world data show that this approach enables a favourable

trade-off between computational complexity and accuracy.

Standard approximations for non-Gaussian observation models, such as those discussed by Wilkinson et al. [2020], Chang et al. [2020], and Ashman et al. [2020], can be applied straightforwardly within our approximation. Our method represents the simplest point in a range of possible approximations. As such there are several promising paths forward to achieve further scalability beyond simply utilising hardware acceleration, including *(i)* applying the estimator developed by Hensman et al. [2013] to our method to utilise mini batches of data, *(ii)* embedding the infinite-horizon approximation introduced by Solin et al. [2018] to trade off some accuracy for a substantial reduction in the computational complexity of our approximation, *(iii)* removing the constraint that observations must appear at the same time as pseudo-points by utilising the method developed by Adam et al. [2020].

Code `github.com/JuliaGaussianProcesses/TemporalGPs.jl` contains an implementation of the approximation developed in this work.

`github.com/willtebbutt/PseudoPointStateSpace-UAI-2021` contains code built on top of TemporalGPs.jl to reproduce the experiments.

Author Contributions

WT conceived the idea, implemented models, and ran the experiments. All authors helped develop the idea, write the paper, and devise experiments.

Acknowledgements

We thank Adrià Garriga-Alonso, Wessel Bruinsma, Matt Ashman, and anonymous reviewers for invaluable feedback. Will Tebbutt is supported by Deepmind and Invenia Labs. Arno Solin acknowledges funding from the Academy of Finland (grant id 324345). Richard E. Turner is supported by Google, Amazon, ARM, Improbable, Microsoft Research and EPSRC grants EP/M0269571 and EP/L000776/1.

References

Vincent Adam, Stefanos Eleftheriadis, Artem Artemev, Nicolas Durrande, and James Hensman. Doubly sparse variational Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2874–2884. PMLR, 2020.

Matthew Ashman, Jonathan So, Will Tebbutt, Vincent Fortuin, Michael Pearce, and Richard E Turner. Sparse Gaussian process variational autoencoders. *arXiv preprint arXiv:2010.10177*, 2020.

Thang D Bui and Richard E Turner. Tree-structured Gaussian process approximations. In *Advances in Neural Information Processing Systems 27*, pages 2213–2221. Curran Associates, Inc., 2014.

Thang D Bui, Josiah Yan, and Richard E Turner. A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18(1):3649–3720, 2017.

David Burt, Carl Edward Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 862–871. PMLR, 2019.

David R Burt, Carl Edward Rasmussen, and Mark van der Wilk. Variational orthogonal features. *arXiv preprint arXiv:2006.13170*, 2020.

Open Data Camden. National Statistics Postcode Lookup UK Coordinates. `https://opendata.camden.gov.uk/Maps/National-Statistics-Postcode-Lookup-UK-Coordina 77ra-mbbn`, 2015. [Online; accessed January-2021].

Paul E Chang, William J Wilkinson, Mohammad Emtiyaz Khan, and Arno Solin. Fast variational learning in state-space Gaussian process models. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

Jiahao Chen and Jarrett Revels. Robust benchmarking in noisy environments. *arXiv e-prints*, art. arXiv:1608.04295, Aug 2016.

Lehel Csató and Manfred Opper. Sparse On-Line Gaussian Processes. *Neural computation*, 14(3):641–668, 2002.

Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse gaussian processes with spherical harmonic features. In *International Conference on Machine Learning*, pages 2793–2802. PMLR, 2020.

David Duvenaud, Hannes Nickisch, and Carl Edward Rasmussen. Additive Gaussian Processes. *Advances in Neural Information Processing Systems*, 24:226–234, 2011.

Jacob Gardner, Geoff Pleiss, Ruihan Wu, Kilian Weinberger, and Andrew Wilson. Product Kernel Interpolation for Scalable Gaussian Processes. In *International Conference on Artificial Intelligence and Statistics*, pages 1407–1416. PMLR, 2018.

Alexander Grigorievskiy, Neil Lawrence, and Simo Särkkä. Parallelizable sparse inverse formulation Gaussian processes (SpInGP). In *International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2017.

Jouni Hartikainen, Jaakko Riihimäki, and Simo Särkkä. Sparse spatio-temporal Gaussian processes with general likelihoods. In *International Conference on Artificial Neural Networks*, pages 193–200. Springer, 2011.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 282–290. AUAI Press, 2013.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360, 2015.

James Hensman, Nicolas Durrande, and Arno Solin. Variational fourier features for Gaussian processes. *The Journal of Machine Learning Research*, 18(1):5537–5588, 2017.

HM Land Registry. Price Paid Data. `https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads`, 2014. [Online; accessed January-2021].

Michael Innes. Don't unroll adjoint: Differentiating ssa-form programs. *CoRR*, abs/1810.07951, 2018. URL `http://arxiv.org/abs/1810.07951`.

Mohammad Khan and Wu Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Artificial Intelligence and Statistics*, pages 878–887, 2017.

Mohammad Emtiyaz Khan and Didrik Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *2018 International Symposium on Information Theory and Its Applications (ISITA)*, pages 31–35. IEEE, 2018.

Miguel Lazaro-Gredilla and Anibal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems*, pages 1087–1095, 2009.

Finn Lindgren, Håvard Rue, and Johan Lindström. An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Jackson Loper, David Blei, John P Cunningham, and Liam Paninski. General linear-time inference for Gaussian processes on one dimension. *arXiv preprint arXiv:2003.05554*, 2020.

Alexander G. de G. Matthews, James Hensman, Richard E. Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 231–239. PMLR, 2016.

Matthew J Menne, Imke Durre, Russell S Vose, Byron E Gleason, and Tamara G Houston. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7): 897–910, 2012.

Thomas Minka. Power EP. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

Patrick Kofod Mogensen and Asbjørn Nilsen Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):615, 2018. doi: 10.21105/joss.00615.

NASA-JPL. NASADEM Merged DEM Global 1 arc second V001 [Data set]. NASA EOSDIS Land Processes DAAC., 2020. URL `https://doi.org/10.5067/MEaSUREs/NASADEM/NASADEM_HGT.001`.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

Anthony O'Hagan. A Markov property for covariance structures. *Statistics Research Report*, 98:13, 1998.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6 (Dec):1939–1959, 2005.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Yunus Saatçi. *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge, 2012.

Simo Särkkä and Ángel F. García-Fernández. Temporal parallelization of Bayesian smoothers. *IEEE Transactions on Automatic Control*, 2020.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

Matthias Seeger. Bayesian methods for support vector machines and gaussian processes. Technical report, Univeristy of Edinburgh, 1999.

Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2003.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264. MIT Press, 2005.

Arno Solin. *Stochastic differential equation methods for spatio-temporal Gaussian process regression*. PhD thesis, Aalto University, 2016.

Arno Solin, James Hensman, and Richard E Turner. Infinite-horizon Gaussian processes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3490–3499, 2018.

Michalis K Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574. PMLR, 2009.

Felipe Tobar. Band-limited Gaussian processes: The sinc kernel. In *Advances in Neural Information Processing Systems*, pages 12749–12759. Curran Associates, Inc., 2019.

Peter Whittle. Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2): 974–994, 1963.

William J Wilkinson, Paul E Chang, Michael Riis Andersen, and Arno Solin. State space expectation propagation: Efficient inference schemes for temporal Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International Conference on Machine Learning*, pages 1775–1784, 2015.