2 Layers Suffice: Learning using Compound Tensor-Variate & Scalar-Variate Stochastic Processes

Dalia Chakrabarty DALIA.CHAKRABARTY@BRUNEL.AC.UK
Department of Mathematics
Brunel University London
Uxbridge, Middlesex UB8 3PH, UK

Kangrui Wang KWANG@TURING.AC.UK
The Alan turing Institute
British Library
London NW1 2DB

Editor:

Abstract

Prediction of a random variable, at test data on another associated variable, follows from the learning of the (generally high-dimensional) functional relationship between the two variables. Such learning often comprises modelling this sought (tensor-valued in general) function with a (high-dimensional) stochastic process - typically, a tensor-variate Gaussian Process (GP). We review three ways of learning covariance matrices of the resulting tensor Normal likelihood, including kernel parametrisation of its covariance matrices. Notwithstanding prevalent deep-learning techniques that treat the number of layers of the learning scheme as a chosen parameter, here we prove that two layers suffice, even when data on the observable is distributed discontinuously, as long as a completely generic covariance kernel function is employed. Such generality manifests in modelling each kernel hyper-parameter as a random function of the sample path of the tensor-variate GP, where each such function can be modelled with a scalar-variate GP, that is proven to be stationary. Thus our learning strategy includes an outer-layer of a non-stationary tensor-variate GP, that is compounded with multiple scalar-variate, stationary GPs in the inner-layer, and we implement Bayesian inference throughout. As an aside, we advance the (Generalised Wishart) generative process of a non-stationary, temporally-evolving covariance matrix. An empirical illustration of this dual-layered learning method is made to a real astronomical dataset, and model checking undertaken.

Keywords: Compound Tensor-variate Scalar-variate GPs, Covariance Kernel parametrisation, Lipschitz continuity, Deep learning

1. Introduction

Real-world applications often demand learning the functional relationship between a random variable $\boldsymbol{S}$, and another variable $\boldsymbol{V}$ that bears influence on $\boldsymbol{S}$, s.t. we can state $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$, where this inter-variable functional relation $\boldsymbol{\xi}(\cdot)$ that we seek to learn, can itself be modelled as random. The ulterior aim behind the uncertainty-included learning of $\boldsymbol{\xi}(\cdot)$, is the uncertainty-included prediction of values of either variable, at which noise-included test data on the other has been realised. The sought function can be modelled as a realisation from an adequately chosen stochastic process, (where uncertainties in the learning are explained by such a generative process, in addition to the noise in the training data). Typically, $\boldsymbol{S}$ is a system parameter vector, and the observed variable $\boldsymbol{V}$ can be tensor-valued in general, such that data comprising its multiple measurements, is hypercuboidally-shaped.

Hypercuboidally-shaped data show up in multiple real-world applications, including those reported by Mardia and Goodall (1993); Bijma et al. (2005); Werner et al. (2008); Theobald and Wuttke (2008); Barton and Fuhrmann (1993). For example, in computer vision, the image of one person might be a matrix of dimensions $a \times b$, i.e. image has a resolution of a pixels by b pixels. Then, repetition across n persons, inflates such image data to a cuboidally-shaped dataset. Examples of handling high-dimensional datasets within computer vision exist (Dryden et al., 2009; Fu, 2016; Pang et al., 2016; Wang, 2011; Qiang and Fei, 2011). In health care, the p number of health parameters of n patients, when charted across k time-points, again generates a high-dimensional data, which gets further enhanced, if the experiment involves tracking for changes across ℓ groups of n patients each, where each such group is identified by the level of intervention (Chari et al., 2010a; Clarke et al., 2008; Oberg et al., 2015; Chari et al., 2010b; Sarkar, 2015; Wang et al., 2015; Fan, 2017). Again, in ecological datasets, there could be n spatial locations at each of which, p traits of k species could be tracked, giving rise to a high-dimensional data (Leitao et al., 2015; Warton, 2011; Dunstan et al., 2013).

It is a shortcoming of traditional modelling strategies, that these groupings in the data are treated as independent – or for that matter, even the variation in parameter values of any group across varying time points, is ignored, and a mere snapshot of each group is considered, one at a time. In this article, we focus on methodologies that permit the consideration of parameters across all relevant levels of measurement, within one integrated framework, to enable the learning of correlations across all such levels, thus permitting the prediction of the system parameter vector, with meaningful uncertainties, and avoid information loss associated with categorisation of data.

While discussing the generic methodologies that help address the problem of learning the inter-variable relationship $\boldsymbol{\xi}(\cdot)$, given general hypercuboidally-shaped data, we focus on such learning when this data displays discontinuities. Then, the generative tensor-variate Gaussian Process (GP) of this function $\boldsymbol{\xi}(\cdot)$, is ascribed a non-stationary covariance function, Acknowledgement of non-stationarity in correlation learning is not new (Paciorek and Schervish, 2004). In some approaches, transformation of the input variable is suggested to accommodate non-stationarity (Sampson and Guttorp, 1992; Snoek et al., 2014; Schmidt and OHagan, 2003). When faced with learning the dynamically varying covariance structure of time-dependent data, others have resorted to learning such a covariance, using Generalised Wishart Process (Wilson and Ghahramani, 2011). In another approach, latent parameters that bear information on non-stationarity, have been modelled with GPs and learnt simultaneously with the sought function (Tolvanen et al., 2014), while others have used multiple GPs to capture the non-stationarity (Gramacy, 2005; Heinonen et al., 2016).

What is currently missing, is a template for including non-stationarity in high-dimensional data, via a flexible and generic model of the correlation structure of the generative stochastic process underlying the sought function, s.t. this correlation structure adapts to the discontinuities of the function sampled from this process. We prove the need for modelling the correlation as dependent on this process' sample path in Section 3.1, and show that this can be undertaken by modelling each length scale of the correlation structure of this high-dimensional stochastic process, as a new random function of the sample path of the process. This random function then, can in turn be modelled as a realisation from a scalar-variate stochastic process, such as a scalar-variate GP (Section 2). Interestingly, each such random function is proven to be continuous, by rephrasing it as a function of the discrete-valued time step variable that the aforementioned sample path is generated at. Hyperparameters of this function's generative scalar-variate process are then rendered data-driven constants (Section 3.2), that we learn. Illustration of the same, using Bayesian inference techniques (Metropolis-within-Gibbs) is given below. The covariance structure of the resulting likelihood is then dependent on this time step variable, and its probability distribution at any time, is presented (as a Generalised Wishart distribution) in Section 4.

Thus, in this paper we forward a method that performs uncertainty-included learning of a high-dimensional functional relationship between the system parameter $\boldsymbol{S}$ and a high-dimensional observable $\boldsymbol{V}$, given discontinuities in the hypercuboidally-shaped data that comprises measurements on $\boldsymbol{V}$, by nesting multiple lower-dimensional, stationary Gaussian Processes, within a tensor-variate,non-stationary GP (Section 2 and Section 3), with inference based on the Metropolis-within-Gibbs technique, as discussed in Section 6. In this paradigm, learning of the sought functional relation $\boldsymbol{\xi}(\cdot)$ is then double-layered, in which multiple scalar-variate GPs inform a high-dimensional (tensor-variate) GP. To contextualise the implications of these results to contemporary deep learning strategies, we include the proof (in Section 5) that no more than 2 such layers in the learning strategy are needed.

These results are then empirically illustrated on a cuboidally-shaped, real-world dataset (Section 7, Section 8); one reason for choosing to work with this particular high-dimensional dataset is that results of its analysis exist in the literature, and comparison (in Section 9.3), of results of the application undertaken here, to those in the literature, showcases the methodology that is forwarded in this paper. The ulterior interest in prediction of the system parameter values, (at which test data on the observable is measured), is undertaken for this application (Section 9, Section 8). Convergence diagnostics of chains run with the Metropolis-within-Gibbs algorithm are presented within these sections, and further diagnostics are included in Appendix A. Additionally, flexibility in the design of the presented model, permits both inverse and forward predictions; this flexibility is exploited to predict new data at chosen system parameter values, given the learnt model, to permit model checking, by comparing such generated data against the empirically observed data (Appendix B).

2. Model

Let system parameter vector $\boldsymbol{S} \in \mathcal{X} \subseteq \mathbb{R}^d$, affect, or be affected by observable $\boldsymbol{V}$, where $\boldsymbol{V}$ is ($k-1$-th ordered) tensor-valued in general, i.e. $\boldsymbol{V} \in \mathcal{Y} \subseteq \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_{k-1}}$, $m_i \in \mathbb{N}, \forall\, i = 1, \ldots, k-1$. That $\boldsymbol{S}$ and $\boldsymbol{V}$ exist in a state of relatedness, is expressed by: $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$ where $\boldsymbol{\xi} : \mathcal{X} \subseteq \mathbb{R}^d \longrightarrow \mathcal{Y} \subseteq \mathbb{R}^{m_1 \times m_2 \times \ldots \times m_{k-1}}$. While this equation expresses the functional relationship between a r.v. called $\boldsymbol{S}$ and another called $\boldsymbol{V}$, when we seek realisations or values of either r.v. at which test measurement on the other r.v. is recorded, we invoke measurement errors in $\boldsymbol{V}$ and $\boldsymbol{X}$, as well as uncertainty in the learnt $\boldsymbol{\xi}(\cdot)$ given noise in the training data. In other words, information about measurement error in either r.v. can be considered to be subsumed into the unknown form of $\boldsymbol{\xi}(\cdot)$, that we aim to learn.

Now, it may appear that the model $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$ is less preferred over the model that expresses the relationship between $\boldsymbol{S}$ and $\boldsymbol{V}$ as $\boldsymbol{S} = \boldsymbol{f}(\boldsymbol{V})$, because in the latter model, the task of learning the unknown functional relationship is easier, in light of $\boldsymbol{f}(\cdot)$ being vector-valued, and therefore lower-dimensional than the tensor-valued function $\boldsymbol{\xi}(\cdot)$. However, difficulty in learning the functional relation between 2 r.v.s increases with dimensionality of the input variable, more than that of the output variable. Thus, the $\boldsymbol{f}(\cdot)$ suggested above is harder to learn than $\boldsymbol{\xi}(\cdot)$. Hence we persist with $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$ as our model equation. The reason for this dependence of difficulty of functional learning, ties in with the kernel-based parametrisation of the covariance structure of the stochastic process that generates the sought function $\boldsymbol{f}(\cdot)$; higher-dimensional kernel functions render functional learning more difficult. We will review this issue in Section 7.

Definition 1 *We define functional relationship $\boldsymbol{\xi}(\cdot)$, between vector-valued $\boldsymbol{S}$, and the $k-1$-th ordered tensor-valued r.v. $\boldsymbol{V}$, as a tensor-valued function, with $\prod\limits_{i=1}^{k-1} m_i$-number of component functions, each of which is a function of vector $\boldsymbol{S}$, with these component functions correlated to each other.*

Ultimately, we want to predict the value of either r.v. ($\boldsymbol{V}$ or $\boldsymbol{S}$), at which a new or test data on the other variable is observed. For example, the inverse prediction of the value $\boldsymbol{s}^{(test)}$ of $\boldsymbol{S}$, at which test data $\boldsymbol{v}^{(test)}$ on $\boldsymbol{V}$ is realised, is given as $\boldsymbol{s}^{(test)} = \boldsymbol{\xi}^{(-1)}(\boldsymbol{V})\big|_{\boldsymbol{V}=\boldsymbol{v}^{(test)}}$, within the conventional paradigm. Such a scheme however does not allow for easy propagation of the uncertainty in learning $\boldsymbol{\xi}(\cdot)$, into the prediction of $\boldsymbol{S}$, or of incorporation of measurement noise in $\boldsymbol{V}$, in the prediction. These problems are supplemented by the obvious concerns related to the inversion of the learnt function; computational complexity of a prediction in this conventional framework increases with dimensionality. Another crucial drawback of these methods – irrespective of dimensionality of the sought function – is that there is no organic way of quantifying the smoothness of the sought $\boldsymbol{\xi}(\cdot)$ directly from the data. Parametric approaches are additionally deficient in high-dimensions. As $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$, the function $\boldsymbol{\xi}(\cdot)$ has the same dimensionality as the tensor-valued $\boldsymbol{V}$ variable. Given the training data $\mathbf{D} := \{(\boldsymbol{s}_i, \boldsymbol{v}_i)\}_{i=1}^{n}$, the correlation between a pair of component functions of this tensor-valued $\boldsymbol{\xi}(\cdot)$ function, computed at 2 given design values of $\boldsymbol{S}$, is the same as that between the corresponding components of $\boldsymbol{V}$. Thus, the function $\boldsymbol{\xi}(\cdot)$ that we seek, will have to acknowledge all such data-driven correlation constraints. However, parametric fitting methods (such as fitting with splines, etc) cannot convey correlation information to the sought function. These concerns are mitigated by modelling the sought function as a stochastic realisation from a generative Process.

So we choose the sought tensor-valued $\boldsymbol{\xi}(\cdot)$ function to be a random realisation from a tensor-variate Gaussian Process(GP). Then by definition of the GP, joint probability density of n realisations of a sampled tensor-valued $\boldsymbol{\xi}(\cdot)$, is given by the correspondingly high-dimensional equivalent of the Multivariate Normal, namely a Tensor Normal density.

Thus, the joint probability of n realisations of the sampled function $\boldsymbol{\xi}(\cdot)$, at the n design points $\boldsymbol{s}_1, \ldots \boldsymbol{s}_n$ follows the k-variate Tensor Normal distribution (Kolda and Bader, 2009;

Richter et al., 2008; McCullagh, 1987; Manceur and Dutilleul, 2013):

$$[\boldsymbol{\xi}(\boldsymbol{s}_1), \ldots, \boldsymbol{\xi}(\boldsymbol{s}_n)] \sim \mathcal{TN}(\boldsymbol{M}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k),$$

where mean of this density is a k-th ordered mean tensor $\boldsymbol{M}$ of dimensions $m_1 \times \ldots \times m_k$, and $\boldsymbol{\Sigma}_j$ is the $m_j \times m_j$-dimensional, j-th covariance matrix; $j = 1, \ldots, k$. Here $\boldsymbol{\xi}(\boldsymbol{S})$ outputs the variable $\boldsymbol{V}$, so that the joint probability above is the joint of the n values of $\boldsymbol{V}$ that comprise the training data $\mathbf{D}$. In other words, probability of the data comprising measurements of $\boldsymbol{V}$ is Tensor Normal, i.e. likelihood of the Process parameters, given the data, is a Tensor Normal density.

Definition 2 *Likelihood of model parameters $\boldsymbol{M}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k$, given data $\mathbf{D}$, is the k-variate Tensor Normal density:*

$$\mathcal{L}(\boldsymbol{M}, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_k | \mathbf{D}) \propto \exp(-\|(\mathbf{D}_{\boldsymbol{V}} - \boldsymbol{M}) \times_1 \boldsymbol{A}_1^{-1} \times_2 \boldsymbol{A}_2^{-1} ... \times_k \boldsymbol{A}_k^{-1}\|^2/2), \qquad (1)$$

where n observed values of the $k-1$-th dimensional tensor-valued $\boldsymbol{V}$ are collated to form the k-th ordered tensor $\mathbf{D}_{\boldsymbol{V}}$. Dependence on $\boldsymbol{S}$ in the RHS of Equation 1, is borne by the covariance matrices. Here $\boldsymbol{A}_j$ is the unique square-root of the positive definite covariance matrix $\boldsymbol{\Sigma}_j$, i.e. $\boldsymbol{\Sigma}_j = \boldsymbol{A}_j \boldsymbol{A}_j^T$.

One way to compute the square root of a matrix, is to use Cholesky decomposition[1].

The notation $\times_j$ in Equation 1 presents the j-mode product of a matrix and a tensor (Oseledets, 2011).

Then in the Bayesian approach, this likelihood can be employed in Equation 1 to write the joint posterior probability density of the mean tensor and covariance matrices, given the data. But before doing that, we identify those parameters – if any – that can be estimated in a pre-processing stage of the inference, in order to reduce the computational burden of inference. Also, it would be useful to find ways of (kernel-based) parametrisation of the sought covariance matrices, thereby reducing the number of parameters that we need to learn. To this effect, we estimate the mean tensor $\boldsymbol{M} \in R^{m_1 \times m_2 \ldots \times m_k}$, as the sample mean $\overline{\boldsymbol{v}}$ of the sample $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$, However, if necessary, the mean tensor itself can be regarded as a random variable and learnt from the data (Chakrabarty et al., 2015), The modelling of the GP covariance structure, is discussed in the following subsection.

Our aim is to predict, subsequent to the functional learning.

Remark 3 *To perform Bayesian inverse prediction of value $\mathbf{s}^{(test)}$ of the input variable $\boldsymbol{S}$, at which test data $\mathbf{v}^{(test)}$ on $\boldsymbol{V}$ is realised, we*

— *sample from the posterior probability density of $\mathbf{s}^{(test)}$ given test data $\mathbf{v}^{(test)}$, and (modal) values of parameters of the Tensor Normal likelihood, subsequent to the MCMC-based inference on the marginals of each such unknown given the training data, or,*

— *sample from the joint posterior probability density of $\mathbf{s}^{(test)}$ and all other unknowns parameters of the Tensor Normal likelihood, given training, as well as test data.*

[1] As Hoff (1997); Manceur and Dutilleul (2013) suggest, a k-th ordered random tensor $\boldsymbol{\Sigma} \in R^{m_1 \times m_2 \ldots \times m_k}$ can be decomposed to a k-th ordered tensor $\boldsymbol{Z}$ and k number of covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k$ by Tucker decomposition, (Hoff et al., 2011; Manceur and Dutilleul, 2013; Kolda and Bader, 2009), according to $\boldsymbol{\Sigma} = \boldsymbol{Z} \times_1 \boldsymbol{\Sigma}_1 \times_2 \boldsymbol{\Sigma}_2 ... \times_k \boldsymbol{\Sigma}_k$,

Computational speed of the first approach is higher, as marginal distributions of GP parameters are learnt separately, namely, before the prediction exercise is undertaken. When the training data is small, or if the training data is not representative of the test data, prediction of $\boldsymbol{s}^{(test)}$ via the second method may affect the learning of the GP parameters.

2.1 Three ways of learning covariance matrices

When possible, covariance matrices of the GP that is invoked to model the sought function $\boldsymbol{\xi}(\cdot)$, are kernel-parametrised. Let the ij-th element of p-th covariance matrix $\boldsymbol{\Sigma}_p^{(m_p \times m_p)}$ be $\sigma_{ij}^{(p)}$; $j, i = 1, \ldots, m_p$, $p \in \{1, \ldots, k\}$.

Definition 4 *Let covariance matrix $\boldsymbol{\Sigma}_p = [\sigma_{ij}^{(p)}]$, bear information about covariance between the i-th and j-th "slice"s of the k-th ordered data tensor $\boldsymbol{D_V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{m_p})$, where the $m_1 \times \ldots \times m_{p-1} \times m_{p+1} \times \ldots \times m_k$-dimensional i-th "slice" of data tensor $\boldsymbol{D_V}$ is the measured value $\boldsymbol{v}_i$ of the $k-1$-th ordered tensor-valued r.v. $\boldsymbol{V}$, where the i-th slice is realised at the i-th design point $\boldsymbol{s}_i$.*

A simple model of the covariance between the i-th and j-th slices of data $\boldsymbol{D_V}$ suggest that $\sigma_{ij}^{(p)}$ is a decreasing function $K_p(\cdot, \cdot)$ of $\parallel \boldsymbol{s}_i - \boldsymbol{s}_j \parallel$, where $\parallel \cdot \parallel_2$ is the L_2 norm. Then $K_p(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the covariance kernel function, computed at the i-th and j-th values of input variable $\boldsymbol{S}$. In such a model, the number of distinct unknown parameters involved in the learning of $\boldsymbol{\Sigma}_p$ reduces from $m_p(m_p + 1)/2$, to the number of hyper-parameters that parametrise the kernel function $K_p(\cdot, \cdot)$.

However, kernel parametrisation is not always possible.

–Firstly, such parametrisation may cause information loss and this may not be acceptable (Aston and Kirch, 2012).

–More fundamentally, we can undertake kernel parametrisation of covariance between a pair of relevant realisations of the output variable ($\boldsymbol{V}$), only when information exists on values of the input space variable $\boldsymbol{S}$, at which a realisation of the output variable is identified. When such information is unattainable,

–we can learn elements of the covariance matrix directly using MCMC; such direct learning of all distinct elements of $\boldsymbol{\Sigma}_p$ is feasible, as long as total number of all unknowns learnt by MCMC $\lesssim 200$. If $\boldsymbol{\Sigma}_p$ is even higher-dimensional,

–we can use a "plugin estimate" for each element of the covariance matrix $\boldsymbol{\Sigma}_p$. Such a plugin estimate can be computed as follows. We collapse each of the m_p number of $k-1$-th ordered tensor-shaped slices of the data, onto the q-th axis in the space $\mathscr{Y}$ of $\boldsymbol{V}$, where $q \in \{1, \ldots, k-1\}$. This will reduce each slice to a m_q-dimensional vector, so that the empirical estimate of $\sigma_{ij}^{(p)}$, is the covariance computed using the i-th and j-th such m_q-dimensional vectors.

Indeed such an empirical estimate of any covariance matrix is easily generated, but it indulges in linearisation amongst the different dimensionalities of the observable $\boldsymbol{V}$, causing loss of information about the covariance structure amongst the components of these high-dimensional slices. This approach is inadequate when the sample size is small because then, the plugin estimate will tend to be incorrect; indeed discontinuities and steep gradients in the data, especially in small-sample and high-dimensional data, will render such estimates of the covariance structure incorrect. Importantly, such an approach does not leave any scope for identifying the smoothness in the function $\boldsymbol{\xi}(\cdot)$ that represents the functional relationship between the input and output variables. Lastly, uncertainties in the estimated GP covariance structure, remain inadequately known.

Remark 5 *Covariance matrices of the tensor normal likelihood density that represents the probability of the data given model parameters (i.e. the likelihood function), can be obtained using training data, as*
– kernel parametrised, or as
– empirically-estimated, or as
– learnt directly using MCMC-based inference schemes.

An accompanying computational worry is the inversion of any covariance matrix; for a covariance matrix that is an $m_p \times m_p$-dimensional matrix, the order for matrix inversion is well known to be $\mathcal{O}(m_p^3)$ (Knuth, 1997).

3. Kernel parametrisation

Definition 6 *When kernel parametrisation of a covariance matrix is undertaken, i.e. we set* $\boldsymbol{\Sigma}_p \equiv [\sigma_{ij}^{(p)}] = [K_p(\boldsymbol{s}_i, \boldsymbol{s}_j)]$, *different forms of the covariance kernel* $K_p(\cdot,\cdot)$ *can be used, such as the simple Squared Exponential (SQE):*

$$K_p(\boldsymbol{s}_i, \boldsymbol{s}_j) := A_0 \left[\exp\left(-(\boldsymbol{s}_i - \boldsymbol{s}_j)^T \boldsymbol{Q}^{-1}(\boldsymbol{s}_i - \boldsymbol{s}_j)\right)\right], \quad \forall i, j = 1, \ldots, d, \tag{2}$$

where $\boldsymbol{Q}^{(d\times d)}$ *is a diagonal matrix, the diagonal elements of which are the length scale hyperparameters* $\ell_1, \ldots, \ell_d \in \mathbb{R}_{>0}$*, where* ℓ_c *is the length scale that we need to move along the* c*-th direction in input space* $\mathcal{X}$*, for correlation to fade by a threshold factor; here* $c = 1, \ldots, d$*, i.e. there are* d*-directions in the space* $\mathcal{X}$ *that hosts the input variable* $\boldsymbol{S}$*. Then* $\boldsymbol{Q}^{-1}$ *is also diagonal, with the diagonal elements given as* $\dfrac{1}{\ell_1}, \ldots, \dfrac{1}{\ell_d}$*, where* $q_c := 1/\ell_c$ *is the smoothness hyperparameter along the* c*-th direction in* $\mathcal{X}$*,* $c \in \{1, \ldots, d\}$*. We learn these* d *unknown parameters from the data. Here* A_0 *is the global amplitude, that is subsumed as a scale factor, in one of those covariance matrices, distinct elements of which are learnt directly using MCMC.*

Remark 7 *The model of the covariance kernel used in Definition 6 avoids using amplitude parameters that depend on the locations at which covariance is computed, i.e. Definition 6 avoids the model:* $K(\boldsymbol{s}_i, \boldsymbol{s}_j) := a_{ij}\left[\exp\left(-(\boldsymbol{s}_i - \boldsymbol{s}_j)^T \boldsymbol{Q}^{-1} \quad (\boldsymbol{s}_i - \boldsymbol{s}_j)\right)\right]$*. Instead, the model that is advanced, is endowed with a global amplitude* A_0*. This helps avoid learning a very large number* $(d(d+1)/2)$ *of amplitude parameters* a_{ij} *directly from MCMC.*

A loose interpretation of this amplitude modelling is that we have scaled all local amplitudes a_{ij} using the global factor $= \max\limits_{ij}\{a_{ij}\}$, and these scaled local amplitudes (that are ≤ 1), are then subsumed into the argument of the exponential in the definition of the SQE kernel function, s.t. reciprocal of the correlation length scales, that are originally interpreted as the elements of the diagonal matrix $\boldsymbol{Q}^{-1}$, are now interpreted as the smoothing parameters modulated by such local amplitudes. This interpretation is loose, since the same smoothness parameter cannot accommodate all scaled local amplitudes$\in (0, 1]$, for all $\boldsymbol{s}_i - \boldsymbol{s}_j$.

3.1 Including non-stationarity, by modelling each hyperparameter of the covariance kernel using a Stochastic Process

It is possible that the distribution of the tensor-valued r.v. $\boldsymbol{V}$ across the space of the input r.v. $\boldsymbol{S}$ is s.t., decline in the correlation between realisations $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ of $\boldsymbol{V}$, can no longer be modelled by a declining function of the Euclidean distance $\parallel \boldsymbol{s}_i - \boldsymbol{s}_j \parallel$; $i, j = 1, \ldots, n$. Then the function $\boldsymbol{\xi}(\boldsymbol{S})$ sampled from the high-dimensional GP, may not be continuous. However $\boldsymbol{v}_i$ is a measured value of $\boldsymbol{V}$, so $\boldsymbol{v}_i$ is always finite $\forall i = 1, \ldots, n$.

Measured values $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ of the output variable $\boldsymbol{V}$ in a training data set, though always finite – implying $\boldsymbol{\xi}(\boldsymbol{s}_i)(= \boldsymbol{v}_i)$ is always finite, at the i-th design point $\boldsymbol{s}_i$ – the distribution of $\boldsymbol{v}$ across values of $\boldsymbol{S}$ can be s.t. the function sampled from the GP given this training data, is rendered Lipschitz continuous or discontinuous. Indeed, the continuity response of the sample function of the GP in such contexts, is adequately captured by the Lipschitz condition. This motivates us to address continuity-related questions raised below, in light of Lipschitz continuity.

Definition 8 *When the data on the output variable $\boldsymbol{V}$ measured at the different locations in input space, is s.t. no sampled function is Lipschitz continuous, we refer to such a distribution to imply there exist discontinuities in the data.*

Below, we prove results that reflect on the nature of Lipschitz continuity of the sampled function, given different types of data distributions, and on the effect of the same on hyperparameters of the correlation function of the kernel parametrised covariance matrix of the GP. We show below that when the distribution of $\boldsymbol{v}$ across $\boldsymbol{s}$, renders the sampled function, not Lipschitz continuous, the applicability of stationary covariance kernels (such as that invoked in Equation 2) is challenged. When a function $\boldsymbol{\xi}(\cdot)$ sampled from the tensor-variate GP, is not Lipschitz continuous, it implies that similarity between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ does not imply similarity between $\boldsymbol{\xi}(\boldsymbol{s}_i)$ and $\boldsymbol{\xi}(\boldsymbol{s}_j)$, $\forall \boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{X}$. Therefore, then it is wrong to adopt a stationary definition of the correlation between the function at pairs of points in its domain, (as in Equation 2), since a stationary kernel function employs the same length scale hyperparameters $\ell_1, \ldots, \ell_d$, (and global amplitude A_0).

Put alternatively, if in the general case of discontinuous data distribution, we still wish to persist with a parametric form of kernel parametrisation suggested in this equation, then we will need to re-interpret the hyperparameters of such a parametric form as adapting to discontinuities in the data. Thus, for a choice of an SQE kernel, the length-scales $\ell_1, \ldots, \ell_d$ can no longer be treated as data-driven constants, but then we will need to model ℓ_c as adapting to the discontinuities in the data $\forall c = 1, \ldots, d$. A generic way of ensuring this is to model ℓ_c as a function of the sample-path of the high-dimensional GP that is invoked to model $\boldsymbol{\xi}(\cdot)$, $\forall c = 1, \ldots, d$.

That *discontinuities in the data* defined above, can be accounted for, by modelling the kernel hyperparameters as dependent on the sample function of this high-dimensional GP, follows from Lemma 13 according to which, if a function sampled from the high-dimensional GP is discontinuous, there is a lack of universality in values of the correlation hyperparameters of the kernel-parametrised correlation matrix of this high-dimensional GP, where by "universal" hereon, is implied: a constant, (albeit unknown), irrespective of the form of the sampled function. On the other hand, Lemma 12 states that continuity in the data implies universality of these hyperparameter values. Lemma 13 uses the conclusion summarised in Remark 10 that follows from Theorem 9, which states that for a random tensor-valued function – that is sampled from a tensor-variate GP – to be Lipschitz continuous, the correlation length scale hyperparameters of this GP need to depend on the form of this sampled function. Thus, Remark 10 summarises that in the absence of discontinuities in data distribution, GP that $\boldsymbol{\xi}(\cdot)$ is sampled from, can be stationary. In the presence of *discontinuities in the data* however, any tensor-valued $\boldsymbol{\xi}(\cdot)$ sampled from a tensor-variate GP will be rendered a continuous function, if this GP is modelled as non-stationary, by ensuring that hyperparameters of its correlation vary with the sampled function.

Theorem 9 *Given the model $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$, with $\boldsymbol{S} \in \mathcal{X} \subseteq \mathbb{R}^d$; tensor-valued r.v. $\boldsymbol{V} \in \mathcal{Y}$; and the tensor-valued function $\boldsymbol{\xi}(\cdot)$ modelled as a random realisation from a high-dimensional Gaussian Process (GP), a sample function $\boldsymbol{\xi}(\cdot)$ of this GP is a Lipschitz-continuous map over the bound set $\mathcal{X}$, if the vector $\boldsymbol{q}$ of correlation hyperparameters of the correlation structure of the GP is s.t. each element of $\boldsymbol{q}$ is $\boldsymbol{\xi}$-dependent, i.e.*

$$\boldsymbol{q}(\boldsymbol{\xi}) = (q_1(\boldsymbol{\xi}), \ldots, q_d(\boldsymbol{\xi}))^T \in \mathbb{R}^d.$$

This correlation structure is defined s.t. absolute value of correlation between $\boldsymbol{\xi}(\boldsymbol{s}_1)$ and $\boldsymbol{\xi}(\boldsymbol{s}_2)$ is

$$|corr(\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2))| := K\left(\langle(\boldsymbol{s}_1 - \boldsymbol{s}_2), \boldsymbol{q}\rangle^2\right), \quad \forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X},$$

with

$$K(\boldsymbol{s}_1, \boldsymbol{s}_2) := \exp\left[-\langle(\boldsymbol{s}_1 - \boldsymbol{s}_2), \boldsymbol{q}\rangle^2\right].$$

Proof Here, $\boldsymbol{\xi}(\cdot)$ is a sample function of a high-dimensional GP. For $\boldsymbol{S} \in \mathcal{X}$, where $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$, and $\boldsymbol{V} \in \mathcal{Y}$, we recall that $\boldsymbol{\xi} : \mathcal{X} \longrightarrow \mathcal{Y}$ is defined to be Lipschitz-continuous map, if

$$d_{\mathcal{Y}}(\boldsymbol{\xi}(\boldsymbol{s}_1) - \boldsymbol{\xi}(\boldsymbol{s}_2)) \leq L_{\boldsymbol{\xi}} d_{\mathcal{X}}(\boldsymbol{s}_1, \boldsymbol{s}_2), \quad \forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X}, \tag{3}$$

–where, $L_{\boldsymbol{\xi}} \in \mathbb{R}$ is the infinum over all values that permit inequation 3 to hold for this given $\boldsymbol{\xi}$, (and it is the Lipschitz constant for $\boldsymbol{\xi}(\cdot)$),
–where $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are metric spaces.

Let metric $d_{\mathcal{X}}(\cdot, \cdot)$ be the L_2 norm:

$$d_{\mathcal{X}}(\boldsymbol{s}_1, \boldsymbol{s}_2) :=\parallel \boldsymbol{s}_1 - \boldsymbol{s}_2 \parallel, \quad \forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X},$$

and the metric $d_{\mathcal{Y}}(\boldsymbol{\xi}(\cdot), \boldsymbol{\xi}(\cdot))$ be defined as (square root of the logarithm of) the inverse of the correlation:

$$d_{\mathcal{Y}}(\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2)) := \sqrt{-\log|corr(\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2))|}, \quad \forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X},$$

–where correlation being a measure of affinity, $\log|1/corr(\cdot, \cdot)|$, transforms this affinity into a squared distance for this correlation model; so the transformation $\sqrt{\log|1/corr(\cdot, \cdot)|}$ to a metric is undertaken;
–and the given kernel-parametrised correlation is:

$$|corr(\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2))| := \exp[-\langle(\boldsymbol{s}_1 - \boldsymbol{s}_2), \boldsymbol{q}\rangle^2], \quad \forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X},\ \boldsymbol{q} \in \mathbb{R}^d,$$

so that

$$d_{\mathcal{Y}}(\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2)) = \langle(\boldsymbol{s}_1 - \boldsymbol{s}_2), \boldsymbol{q}\rangle.$$

Then for the map $\boldsymbol{\xi}$ to be Lipschitz-continuous, from inequation 3, we require:

$$\sum_{i=1}^{d} q_i^2(\boldsymbol{s}_1^{(i)} - \boldsymbol{s}_2^{(i)})^2 \leq L_{\boldsymbol{\xi}}^2 \sum_{i=1}^{d}(\boldsymbol{s}_1^{(i)} - \boldsymbol{s}_2^{(i)})^2, \tag{4}$$

where the vector of correlation hyperparameters, $\boldsymbol{q} = (q_1, \ldots, q_d)^T$.

Let

$$q_{max} := \max(q_1, \ldots, q_d). \tag{5}$$

Then q_{max} exists for finite correlation hyperparameters. Let

$$(q_i^{/})^2 := \left(\frac{q_i}{q_{max}}\right)^2 \leq 1, \forall i = 1, \ldots, d.$$

Inequation 4 is valid, if we choose to define the $\boldsymbol{\xi}$-specific constant $L_{\boldsymbol{\xi}}$ as:

$$L_{\boldsymbol{\xi}}^2 = q_{max}^2, \tag{6}$$

since $(q_i^{/})^2 \leq 1$.
Thus a sample function $\boldsymbol{\xi}(\cdot)$ of the given high-dimensional GP, is rendered Lipschitz-continuous by the choice suggested in equation 6.
But this equation also implies that q_{max} varies with the form of $\boldsymbol{\xi}$, since the LHS of this equation is ξ-dependent. So q_{max} is $\boldsymbol{\xi}$-dependent by this choice.
Then recalling definition, q_{max} from Equation 5, it follows that in general, q_i is $\boldsymbol{\xi}$-dependent, $\forall i = 1, \ldots, d$.
Thus, any sample function $\boldsymbol{\xi}(\cdot)$ of the given high-dimensional GP, is rendered Lipschitz-continuous, for $\boldsymbol{\xi}$-dependent q_i. □

Remark 10 *Theorem 9 states that if a universal value of $L_{\boldsymbol{\xi}}$ can allow for inequation 3 to hold for distinct realisations of the random function $\boldsymbol{\xi}(\cdot)$ (as distinct samples taken from a high-dimensional GP), then LHS of Equation 6 is independent of the sampled forms of $\boldsymbol{\xi}(\cdot)$. This implies, RHS of this equation is independent of $\boldsymbol{\xi}(\cdot)$ too, i.e. the maxima q_{max} amongst components of the correlation hyperparameter vector, is independent of the form of the sampled function. On the other hand, if a universal $L_{\boldsymbol{\xi}}$ is not valid for all sampled functions, and LHS of Equation 6 is dependent on the sample function, then q_{max} is $\boldsymbol{\xi}$-dependent.*

We anticipate sample function $\boldsymbol{\xi}(\cdot)$ of the high-dimensional GP, to be locally or globally discontinuous, as delineated in Definition 11.

Definition 11 *The Lipschitz continuity of a sample function $\boldsymbol{\xi}(\cdot)$ can be s.t.*

Case(I) $\forall \boldsymbol{s}_i \in \mathcal{X}$, $\exists \boldsymbol{s}_2 \in \mathcal{X}$, $\boldsymbol{s}_i \neq \boldsymbol{s}_2$, s.t. $\nexists$ finite Lipschitz constant $L_{\boldsymbol{\xi}}^{(i,2)} > 0$, for which $d_{\mathcal{Y}}(\boldsymbol{\xi}(\boldsymbol{s}_i) - \boldsymbol{\xi}(\boldsymbol{s}_2)) \leq L_{\boldsymbol{\xi}}^{(i,2)} d_{\mathcal{X}}(\boldsymbol{s}_i, \boldsymbol{s}_2)$. Here the bounded set $\mathcal{X} \subset \mathbb{R}^d$.

Case(II) $\forall \boldsymbol{s}_i \in \mathcal{X}$, $\exists \boldsymbol{s}_2, \boldsymbol{s}_3 \in \mathcal{X}$, *with* $\parallel \boldsymbol{s}_2 - \boldsymbol{s}_i \parallel \neq \parallel \boldsymbol{s}_3 - \boldsymbol{s}_i \parallel$, *s.t.* $d_{\mathcal{Y}}(\boldsymbol{\xi}(\boldsymbol{s}_i) - \boldsymbol{\xi}(\boldsymbol{s}_2)) \leq L_{\boldsymbol{\xi}}^{(i,2)} d_{\mathcal{X}}(\boldsymbol{s}_i, \boldsymbol{s}_2)$, *but* $d_{\mathcal{Y}}(\boldsymbol{\xi}(\boldsymbol{s}_i) - \boldsymbol{\xi}(\boldsymbol{s}_3)) \leq L_{\boldsymbol{\xi}}^{(i,3)} d_{\mathcal{X}}(\boldsymbol{s}_i, \boldsymbol{s}_3)$, *and* $L_{\boldsymbol{\xi}}^{(i,2)} \neq L_{\boldsymbol{\xi}}^{(i,3)}$. *In such a case, the Lipschitz constant used for the sample function* $\boldsymbol{\xi}(\cdot)$ *is defined to be*

$$L_{\boldsymbol{\xi}} = \max\{L_{\boldsymbol{\xi}}^{(i,j)}\}_{i \neq j; \boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{X}}. \tag{7}$$

Lemma 12 *Let* n *realisations of a random function sampled from a high-dimensional GP, be elements of the set* $\{\boldsymbol{\xi}_1(\cdot), \ldots, \boldsymbol{\xi}_n(\cdot)\}$, *where* $\forall i = 1, \ldots, n$, $\boldsymbol{\xi}_i(\cdot)$ *is s.t. it is*
–either globally Lipschitz, or is as described in Case II of Definition 11,
–and Case I (of Definition 11) is not true.
Then $\exists$ *a universal correlation hyperparameter vector that parametrises the GP that sample functions* $\boldsymbol{\xi}_i$, $i = 1, \ldots, n$ *are sampled from.*

Proof $\forall i \in \{1, \ldots, n\}$, $\exists$ a finite Lipschitz constant $L_{\boldsymbol{\xi}_i}$ defined as in Equation 7, for $\boldsymbol{\xi}_i(\cdot)$. Then $\forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X}$, $\exists$ a finite $L_{max} > 0$, where

$$L_{max} := \max_{\boldsymbol{\xi}}\{L_{\boldsymbol{\xi}_1}, L_{\boldsymbol{\xi}_2}, \ldots, L_{\boldsymbol{\xi}_n}\}, \tag{8}$$

i.e. $\exists$ a finite Lipschitz constant for all forms of the function sampled from the high-dimensional GP,
$\Longrightarrow \exists$ a universal correlation hyperparameter vector that parametrises the GP that all n sample functions are sampled from (by Remark 10 on Theorem 9), where such universality refers to the same value of this vector, independent of the sample function from this GP. By Theorem 9, (Equation 5), the maxima of the d components of this universal correlation hyperparameter vector, is then L_{max}. $\square$

Thus we see from Lemma 12 that adherence to Lipschitz continuity of all forms of the function $\boldsymbol{\xi}(\cdot)$ sampled from the high-dimensional GP, implies that a universal correlation length-scale vector can describe the GP that all of these sampled functions are selected from. Equivalently, in absence of *discontinuities in the data* that comprises measured values of the observable $\boldsymbol{V}$, a universal correlation length-scale vector can describe the correlation structure of the GP that the function $\boldsymbol{\xi}(\cdot)$ is sampled from.

Next we examine the implication of sampled function(s) that is (are) not Lipschitz continuous.

Lemma 13 *Let* n *forms of a random function sampled from a high-dimensional GP, be elements of the set* $\{\boldsymbol{\xi}_1(\cdot), \ldots, \boldsymbol{\xi}_n(\cdot)\}$, *s.t.* $\exists \boldsymbol{\xi}_i(\cdot) \in \{\boldsymbol{\xi}_1(\cdot), \ldots, \boldsymbol{\xi}_n(\cdot)\}$ *that is not Lipschitz continuous.*
Then correlation hyperparameters are dependent on the sample path $\boldsymbol{\xi}(\cdot)$ *of this GP.*

Proof $\exists \boldsymbol{\xi}_i(\cdot) \in \{\boldsymbol{\xi}_1(\cdot), \ldots, \boldsymbol{\xi}_n(\cdot)\}$ that is not Lipschitz continuous.
$\Longrightarrow$ sampled function $\boldsymbol{\xi}_k(\cdot)$ is defined by Case I of Definition 11.
$\Longrightarrow \nexists$ a universal (i.e. constant) L_{max}, irrespective of the sample function of this high dimensional GP.
Then by Remark 10 that follows from Theorem 9, the maxima q_{max} of the d components of the hyperparameter vector (of the correlation structure of the GP that the sample functions are modelled by), is no longer universal, but dependent on the sample function of this

high-dimensional GP.
$\Longrightarrow$ components of the correlation hyperparameter vector are $\boldsymbol{\xi}$-dependent in general, from Equation 5. $\square$ Lemma 13 then suggests that *discontinuities in the data* that comprise measured values of $\boldsymbol{V}$, would imply that the correlation hyperparameter of the high-dimensional GP that the random function $\boldsymbol{\xi}(\cdot)$ is sampled from, is dependent on the sample function of this GP.

Remark 14 *Above, $q_1, \ldots, q_d$ are hyperparameters of the correlation kernel; they are interpreted as the reciprocals of the length-scales $\ell_1, \ldots, \ell_d$, i.e. $\ell_i = 1/q_i, \forall i = 1, \ldots, d$.*

Remark 15 *If the map $\boldsymbol{\xi} : \mathcal{X} \longrightarrow \mathcal{Y}$ is Lipschitz-continuous, (i.e. if hyperparameters $q_1, \ldots, q_d$ are $\boldsymbol{\xi}$-dependent, by Theorem 9), then by Kerkheim's Theorem (Kerkheim, 1994), $\boldsymbol{\xi}$ is differentiable almost everywhere in $\mathcal{X} \subset \mathbb{R}^d$; this is a generalisation of Rademacher's Theorem to metric differentials (see Theorem 1.17 in Hajlasz (2014)). However, in our case, the function $\boldsymbol{\xi}(\cdot)$ is not necessarily differentiable given discontinuities in the data on the observable $\boldsymbol{V} \in \mathcal{Y}$, and therefore, is not necessarily Lipschitz.*

3.2 Re-interpretation of dependence of correlation hyperparameters on sample function, as dependence on a discrete time index

Theorem 9 and Lemma 13 negate usage of a universal value of the correlation length scale of the correlation structure of the high-dimensional GP that $\boldsymbol{\xi}(\cdot)$ is a sample function of, in anticipation of *discontinuities in the data*, and this leads us to model these correlation hyperparameters as dependent on the sample function $\boldsymbol{\xi}(\cdot)$ of this GP. However, one random sample is taken from this GP, in each step (or iteration) of the iterative (Bayesian) inference scheme that is undertaken. We employ this fact, to re-interpret the variability of the correlation length parameters.

Proposition 16 *For $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$, with $\boldsymbol{S} \in \mathcal{X} \subseteq \mathbb{R}^d$ and $\boldsymbol{V} \in \mathcal{Y} \subseteq \mathbb{R}^{(m_1 \times \ldots \times m_k)}$,*

$$|corr(\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2))| := \exp\left[-\langle(\boldsymbol{s}_1 - \boldsymbol{s}_2), \boldsymbol{q}(\boldsymbol{\xi})\rangle^2\right], \quad \forall \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X},$$

where $\boldsymbol{\xi}(\cdot)$ is a sample function of a tensor-variate GP. In this updated model, c-th component $q_c = 1/\ell_c$ of correlation hyperparameter $\boldsymbol{q}(\boldsymbol{\xi})$ is modelled as randomly varying with the sample function, $\boldsymbol{\xi}(\cdot)$, of the tensor-variate GP, $\forall c = 1, \ldots, d$.
In the iterative inference that is undertaken, one sample function of the tensor-variate GP is generated in every iteration

$$\begin{aligned}
&\Longrightarrow q_c, \textit{ is dependent on sample function of the GP, where a sample is} \\
&\quad\ \textit{generated per iteration} \\
&\equiv q_c \textit{ is dependent on iteration number } T \in \{0, 1, \ldots, t_{max}\} \subset \mathbb{Z}_{\geq 0}, \\
&\Longrightarrow \textit{We model} \quad \ell_c = g_c(t), \quad c = 1, \ldots, d,
\end{aligned}$$

where this scalar-valued random function $g_c : \{0, 1, \ldots, t_{max}\} \subset \mathbb{Z}_{>0} \longrightarrow \mathbb{R}_{\geq 0}$, is modelled as a realisation from a scalar-variate GP.

It is important to remember that the scalar-variate GP that $g_c(\cdot)$ is sampled from, is independent of the GP that $g_{c'}(\cdot)$ is sampled from; $c \neq c'; c, c' = 1, \ldots, d$. Thus, the distinction between these scalar-variate GPs that generate the iteration-number dependent functions along the d different directions of input space $\mathcal{X}$, will be brought out via the dependence in the notation for the sample function of the c-th scalar-variate GP, as $g_c(\cdot)$, $c = 1, \ldots, d$.

In addition, the correlation structure of the generative scalar-variate GP – that generates sample functions along a given direction in input space – can vary. In other words, values of parameters that define the (SQE) kernel used to parametrise the scalar-variate GP that generates sample functions along the c-th direction in input space, may vary. This variation can happen with time, if the correlation structure along one/multiple directions in input space, is varying with time; a temporally-evolving data that comprises measurements of the random variable $\boldsymbol{V}$, where *discontinuities in the data* are present, can cause such a variation. Parameters that define the covariance kernel include the amplitude A and scale δ. Thus, scalar-valued functions sampled from GPs that are distinguished by values of the amplitude A and length-scale hyperparameter δ of the invoked SQE covariance kernel – even for a given c value – should be marked by these descriptor variables: $A > 0$ and $\delta > 0$.

Proposition 17 *We restate relationship between iteration number T and correlation length scale hyperparameter ℓ_c in the c-th direction in input space as:*

$$\ell_c = g_{c,\boldsymbol{x}}(t), \quad \textit{where vector of descriptor variables is} \quad \boldsymbol{X} := (A, \delta)^T, \quad \textit{with}$$

–A_c the amplitude variable of the SQE-looking covariance function of the scalar-variate GP that $g_{c,\boldsymbol{x}}(\cdot)$ is a realisation of. A_c takes the value $a_c \geq 0$;
–δ_c the length scale variable of the SQE-looking covariance function of the scalar-variate GP that $g_{c,\boldsymbol{x}}(\cdot)$ is a realisation of; $\delta_c \in \mathbb{R}_{>0}$.
Then the scalar-variate GPs that $g_{c,\boldsymbol{x}}(\cdot)$ and $g_{c,\boldsymbol{x}^/}(\cdot)$ are sampled from, have distinct correlation functions for $\boldsymbol{x} \neq \boldsymbol{x}^/$. Here $c = 1, \ldots, d$.

Definition 18 *Value of correlation length scale hyperparameter ℓ_c, at the t-th iteration, acknowledges information on only the past t_0 number of iterations:*

$$\begin{aligned} \ell_c &= g_{c,\boldsymbol{x}}(t - t^{'}), \quad \textit{if } t \geq t_0, \; c = 1, \ldots, d; \; t^{'} = 1, \ldots, t_0, \\ \ell_c &= \ell_c^{(const)}, \quad \textit{if } t = 0, 1, \ldots, t_0 - 1, \quad c = 1, \ldots, d, \end{aligned} \tag{9}$$

where $\ell_c^{(const)}$ is an unknown constant that we learn from the data, during the first t_0 iterations.

As $g_{c,\boldsymbol{x}}(t)$ is a realisation from a scalar-variate GP, the joint probability distribution of t_0 number of values of the function $g_{c,\boldsymbol{x}}(t)$, at a given $\boldsymbol{x} = (a, \delta)^T$, is Multivariate Normal, with t_0-dimensional mean vector $\boldsymbol{M}_{c,\boldsymbol{x}}$ and $t_0 \times t_0$-dimensional covariance matrix $\boldsymbol{\Psi}_{c,\boldsymbol{x}}$. This follows from the definition of a scalar-variate GP from which t_0 samples have been generated. Thus, the joint probability of the t_0 number of samples from this Process is

$$[g_{c,\boldsymbol{x}}(t-1), \ldots, g_{c,\boldsymbol{x}}(t-2), g_{c,\boldsymbol{x}}(t-t_0)] \sim \mathcal{MN}(\boldsymbol{M}_{c,\boldsymbol{x}}, \boldsymbol{\Psi}_{c,\boldsymbol{x}}). \tag{10}$$

Definition 19 *t_0 is the number of iterations that we look back at, to collect the dynamically-varying "look back-data" $\mathbf{D}_{c,t}^{(orig)} := \{\ell_{c,t-t_0}, \ldots, \ell_{c,t-1}\}$ employed to learn parameters of the scalar-variate GP that $g_{c,\boldsymbol{x}}(\cdot)$ is modelled with.*

- *The mean vector $\boldsymbol{M}_{c,\boldsymbol{x}}$ is empirically estimated as the mean of the dynamically varying look back-data, s.t. at the t-th iteration it is estimated as a t_0-dimensional vector with each component $\hat{m}_{c,\boldsymbol{x}}^{(t)} := [\ell_{c,t-t_0} + \ldots + \ell_{c,t-1}]/t_0$.*
- *$t_0 \times t_0$-dimensional covariance matrix is dependent on the iteration-number and this is now acknowledged in the notation to state:*

$$\boldsymbol{\Psi}_{c,\boldsymbol{x}}(t) = \left[a_c \exp\left(-\frac{(t_i - t_j)^2}{\delta_c^2} \right) \right], \; i,j = t-1, \ldots, t-t_0.$$

In the t-th iteration, upon the empirical estimation of the mean as given above in Definition 19, the mean is subtracted from the "look back-data" $\mathbf{D}_{c,t}^{(orig)}$ so that the subsequent mean-subtracted look back-data is $\mathbf{D}_{c,t} := \{\ell_{c,t-t_0} - \hat{m}_{c,\boldsymbol{x}}^{(t)}, \ldots, \ell_{c,t-1} - \hat{m}_{c,\boldsymbol{x}}^{(t)}\}$. It is indeed this mean-subtracted sample that is used.

Definition 20 *In light of this declared usage of the mean-subtracted "look back-data" $\mathbf{D}_{c,t}$, we update the likelihood over what is declared in Equation 10, to:*

$$[g_{c,\boldsymbol{x}}(t-1), \ldots, g_{c,\boldsymbol{x}}(t-2), g_{c,\boldsymbol{x}}(t-t_0)] \sim \mathcal{MN}(\underset{\sim}{0}, \boldsymbol{\Psi}_{c,\boldsymbol{x}}(t)), \quad \forall c = 1, \ldots, d. \tag{11}$$

Then Equation 11 states that the joint probability of the t_0 realisations of the correlation length-scale parameter ℓ_c – that informs on how quickly correlation is fading along the c-th direction in the input space of the r.v. $\boldsymbol{S}$, in the tensor-variate GP that $\boldsymbol{\xi}(\cdot)$ is sampled from, in each iteration – is a Multivariate Normal density, parametrised by a zero mean vector, and an iteration step-number dependent, i.e. a time-dependent correlation matrix, (which is additionally, chosen to be SQE kernel-parametrised, distinguished by amplitude and length-scale hyperparameters that are the components of $\boldsymbol{X}$).

As we have seen, the correlation structure of a GP that generates sample functions along a given direction in input space, may be evolving with time – as parametrised by evolution in the amplitude and scale parameters of its correlation structure. This motivates the need for undertaking study of temporally evolving correlation structures, and in particular, an application may seek to make inference on the parameters that determine the temporal evolution of such a correlation structure. The distribution of such a time-varying correlation matrix, is advanced in the next section.

4. Temporally-evolving covariance matrix

To summarise the modelling strategies discussed above, we model $\boldsymbol{\xi}(\cdot)$ as a random realisation from a high-dimensional, tensor-variate GP, s.t. each hyperparameter of the kernel-parametrised correlation matrix of the resulting Tensor Normal likelihood, is modelled as an unknown function of the iteration step-number, i.e. of the discrete-valued time variable T. Here, each such unknown function is modelled as a realisation from a scalar-variate GP that has a temporally-evolving, (i.e. T-dependent) covariance structure, resulting in a Multivariate Normal likelihood, with a time-dependent covariance matrix. In Theorem 21, we identify the generating stochastic process of this time-dependent matrix-valued covariance function, as a Generalised Wishart Process.

Theorem 21 *The dependence on time T, of the dynamically varying covariance matrix $\boldsymbol{\Psi}_{c,\boldsymbol{x}}(T)$ of the Multivariate Normal likelihood in Equation 11, at iteration number $t \geq t_0$, is modelled by a Generalised Wishart Process ($\mathcal{GWP}$):*

$$\boldsymbol{\Psi}_{c,\boldsymbol{x}}(t) \sim \mathcal{GWP}(d, \boldsymbol{G}_c, k(\cdot,\cdot)), \quad \textit{where}$$

– t_0 is the number of iterations we look back to;
– $k(\cdot,\cdot)$ is the covariance kernel parametrising the covariance function of the scalar-variate GP that generates the scalar-valued function $g_{c,\boldsymbol{x}}(\cdot)$, at the vector $\boldsymbol{x} = (a_c, \delta_c)^T$ of descriptor variables, s.t. $k(t_i, t_j) = \exp\left(-\frac{(t_i - t_j)^2}{\delta_c^2}\right)$, $\forall t_i, t_j = t-1, \ldots, t-t_0$;
– $\boldsymbol{G}_c$ is a positive definite square scale matrix $\boldsymbol{G}_c$ of dimensionality t_0, containing the amplitudes of this covariance function;
– $c = 1, \ldots, d$, with the space $\mathcal{X}$ of input variable $\boldsymbol{S}$ d-dimensional.

Proof The covariance kernel $k(\cdot,\cdot)$ that parametrises the covariance function of the scalar-variate GP that generates $g_{c,\boldsymbol{x}}(t)$, is s.t. $k(t_i, t_i)$=1 $\forall i = 1, \ldots, t_0$.

At each time point, i.e. at each iteration, a new value of the vector $\boldsymbol{x}_c$ of descriptor variables in the c-th direction in the space $\mathcal{X}$ of the input variable $\boldsymbol{S}$, is generated, s.t. in the $t - t_i$-th iteration, it is $\boldsymbol{x}_{c,i} = (a_{c,i}, \delta_{c,i})^T$; $t - t_i = t-1, \ldots, t-t_0$

$$\Longrightarrow \text{at} \quad T = t, \quad \{g_{c,\boldsymbol{x}_1}(t), \ldots, g_{c,\boldsymbol{x}_{t_0}}(t)\} \quad \text{is a sample of the random variable} \quad g_{c,\boldsymbol{x}}(t).$$

Now, $corr(g_{c,\boldsymbol{x}}(t - t_i), g_{c^{/},\boldsymbol{x}^{/}}(t - t_j)) = k(t_i, t_j)\delta(c, c^{/})\delta(\boldsymbol{x}, \boldsymbol{x}^{/})$, where $\delta(\cdot,\cdot)$ is the Delta function.
$\Longrightarrow$ sample estimate of $Cov(g_{c,\boldsymbol{x}}(t - t_i), g_{c,\boldsymbol{x}}(t - t_j))$ is

$$Cov(g_{c,\boldsymbol{x}}(t-t_i), g_{c,\boldsymbol{x}}(t-t_j)) = \sum_{k=1}^{t_0} a_{c,i} a_{c,j} g_{c,\boldsymbol{x}_k}(t - t_i) g_{c,\boldsymbol{x}_k}(t - t_j), \quad \forall t-t_i, t-t_j = t-1, \ldots, t-t_0,$$

which is the ij-th element of matrix $\boldsymbol{\Psi}_{c,\boldsymbol{x}}(t)$. This definition of the plugin estimate of the covariance holds, since mean of the r.v. $g_{c,\boldsymbol{x}}(t)$ is 0, as we have sampled the function $g_{c,\boldsymbol{x}}(\cdot)$ from a zero-mean scalar-variate GP.

$$\text{Let} \quad \boldsymbol{g}_{c,\boldsymbol{x}_k}(t) := (g_{c,\boldsymbol{x}_k}(t - t_1), \ldots, g_{c,\boldsymbol{x}_{t_k}}(t - t_0))^T, \quad k = 1, \ldots, t_0.$$

Let $\boldsymbol{G}_c$ be a $t_0 \times t_0$-dimensional diagonal matrix, the i-th diagonal element of which is $a_{c,i}^2$. Then factorising the scale matrix $\boldsymbol{G}_c = \boldsymbol{L}_{G_c} \boldsymbol{L}_{G_c}^T$, $\boldsymbol{L}_{G_c}$ is diagonal with the i-th diagonal element $a_{c,i}$; $i = 1, \ldots, t_0$. This is defined for every $c \in \{1, \ldots, d\}$.

Then at iteration number $T = t$, we define the current covariance matrix

$$\boldsymbol{\Psi}_{c,\boldsymbol{x}}(t) := \sum_{k=1}^{t_0} \boldsymbol{L}_{G_c} \left(\boldsymbol{g}_{c,\boldsymbol{x}_k}(t)\right)^T \boldsymbol{g}_{c,\boldsymbol{x}_k}^T(t) \boldsymbol{L}_{G_c}^T.$$

Then $\boldsymbol{\Psi}_{c,\boldsymbol{x}}(t)$ is distributed according to the Wishart distribution w.p, $\boldsymbol{G}_c$ and d (Eaton, 1990), i.e. the dynamically-varying covariance matrix is:

$$\boldsymbol{\Psi}_{c,\boldsymbol{x}}(t) \sim \mathcal{GWP}(d, \boldsymbol{G}_c, k(\cdot,\cdot)).$$

□

Remark 22 *If interest lies in learning the covariance matrix at any time point, we could proceed to inference here from, in attempt of the learning of the unknown parameters of this $\mathcal{GWP}$, given the lookback-data $\mathbf{D}_{c,t}$. Our learning scheme then would then involve compounding a Tensor-Variate GP and a $\mathcal{GWP}$.*

The above would be a delineated route to recover the temporal variation in the correlation structure of time series data (as studied, for example by Wilson and Ghahramani (2011)).

Remark 23 *In our study, the focus is on discontinuities in data, where such data is also high-dimensional, and on learning the relationship $\boldsymbol{\xi}(\cdot)$ between the observable $\boldsymbol{V}$ that generates such data, and the system parameter $\boldsymbol{S}$–with the ulterior aim being parameter value prediction. So learning the time-varying covariance matrix $\Psi(t)$ is not the focus of our method development.*

We want to learn $\boldsymbol{\xi}(\cdot)$ given training data $\mathbf{D}$. The underlying motivation is to sample a new $g_{c,\boldsymbol{x}}(\cdot)$ from a scalar-variate GP, at new values of $a_1, \ldots, a_d, \delta_1, \ldots, \delta_d$, to subsequently sample a new tensor-valued function $\boldsymbol{\xi}(\cdot)$, from the tensor-normal GP, at a new value of its d-dimensional correlation length scale hyperparameter vector $\boldsymbol{\ell}$.

5. 2-layers suffice

It may be argued that just as we ascribe stochasticity to the length scales $\ell_1, \ldots, \ell_d$ that parametrise the correlation structure of the tensor-variate GP that models $\boldsymbol{\xi}(\cdot)$, we need to do the same to the descriptor variables a, δ that parametrise the correlation structure of the scalar-variate GP that generates $g_{c,\boldsymbol{x}}(t)$. Following this argument, we would need to hold a, δ – or at least model the scale δ – to be dependent on the sample path of the scalar-variate GP, i.e. set δ to be dependent on $g_{c,\boldsymbol{x}}(\cdot)$.

However, we show below that a global choice of δ is possible irrespective of the sampled function $g_{c,\boldsymbol{x}}(\cdot)$, given that $g_{c,\boldsymbol{x}} : \{t-1, \ldots, t-t_0\} \subset \mathbb{Z}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is always continuous (a standard result). In contrast, the function $\boldsymbol{\xi}(\cdot)$ not being necessarily Lipschitz (see Remark 15), implies that the correlation kernel hyperparameters q_c, are $\boldsymbol{\xi}$-dependent, $\forall c = 1, \ldots, d$.

The argument posed above within Section 5, is indeed motivating scrutiny of the number of layers in the learning strategy that needs to be included. Deep learning in general suggests multiple such layers, but in the last paragraph, sufficiency of 2 layers is being indicated. Below we address this immediate concern for limiting the layering of our learning scheme to only 2.

Theorem 24 *Given $\ell_c = g_{c,\boldsymbol{x}}(t)$, with $T \in \mathcal{N} \subset \mathbb{Z}_{\geq 0}$ and $\ell_c \in \mathbb{R}$, the map $g_{c,\boldsymbol{x}} : \mathbb{Z}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is a Lipschitz-continuous map, $\forall c = 1, \ldots, d$. Here $\mathcal{N} := \{t - t_1, \ldots, t - t_0\}$*

The proof of this theorem is a standard proof of the basic result that the map from the set of natural numbers to reals, is continuous.

Proof Distance $d_{t_{1,2}}$ between $t_1, t_2 \in \mathcal{N}$ is $|t_1 - t_2|$.

Similarly, distance $d_{g_{1,2}}$ between $g_{c,\boldsymbol{x}}(t_1), g_{c,\boldsymbol{x}}(t_2) \in \mathbb{R}_{\geq 0}$ is $|g_{c,\boldsymbol{x}}(t_1) - g_{c,\boldsymbol{x}}(t_2)|$.

$$\text{Assume} \quad \frac{d_{g_{1,2}}}{M} > d_{t_{1,2}}, \quad M > 0, \;\; M \text{ is finite } \forall t_1, t_2 \in \mathcal{N}.$$

$$\text{Consider} \quad \frac{d_{g_{1,2}}}{M} = 1/2 \Longrightarrow |t_1 - t_2| < 1/2 \quad \text{by our assumption,}$$

i.e. for this choice of the LHS of the assumed inequation, the only solution for $|t_1 - t_2| < 1/2$ is $t_1 = t_2$, as $t_1, t_2 \in \mathbb{Z}_{\geq 0}$.

But $t_1 = t_2 \Longrightarrow g_{c,\boldsymbol{x}}(t_1) = g_{c,\boldsymbol{x}}(t_2)$ for injective $g_{c,\boldsymbol{x}}(\cdot)$, i.e. LHS of the assumed inequation is then 0.

This is a contradiction (contradicts our choice of 1/2 for the LHS).

$\therefore$ our assumption is wrong,

$\Longrightarrow$, the correct inequation is:

$$\frac{d_{g_{1,2}}}{M} \leq d_{t_{1,2}}, \quad M > 0, \;\; M \text{ is finite } \forall t_1, t_2 \in \mathcal{N},$$

i.e.

$$|g_{c,\boldsymbol{x}}(t_1) - g_{c,\boldsymbol{x}}(t_2)| \leq M|t_1 - t_2|, \quad M > 0, \;\; M \text{ is finite } \forall t_1, t_2 \in \mathcal{N} \subset \mathbb{Z}_{\geq 0}$$

i.e. $g_{c,\boldsymbol{x}}(\cdot)$ is Lipschitz continuous. □

Now that we have proved continuity of any iteration-number dependent function that outputs the correlation hyperparameter q_c along the c-th direction of input space, what follows is proof of existence of a universal scale hyperparameter of the correlation structure of the scalar-variate GP from which any such function is sampled from, where "universality" as defined above, implies existence of a unique scale hyperparameter irrespective of the difference amongst the sampled functions of this GP, for this given direction in input space (Theorem 26). This result in fact follows from Theorem 25, that proves the existence of a unique minima amongst the Lipschitz constants that define the Lipschitz continuity of the sample functions of this scalar-variate GP, where the continuity of any such sample function is already established by Theorem 24.

Theorem 25 *For any sampled function $g_{c,\boldsymbol{x}} : \mathcal{N} \longrightarrow \mathbb{R}_{\geq 0}$ realised from a scalar-variate GP that has a covariance function that is kernel-parametrised with an SQE kernel function, parametrised by amplitude and scale hyperparameters, the Lipschitz constant that defines the Lipschitz-continuity of $g_{c,\boldsymbol{x}}(\cdot)$, is $g_{c,\boldsymbol{x}}$-dependent, and is given by the reciprocal of the scale hyperparameter, s.t. the set of t_0 values of scale hyperparameters, for each of the t_0 samples of $g_{c,\boldsymbol{x}}(\cdot)$ taken from the scalar-variate GP, admits a finite minima.*

Proof For $\ell_c = g_{c,\boldsymbol{x}}(T)$, $g_{c,\boldsymbol{x}} : \mathcal{N} \subset \mathbb{Z}_{\geq 0} \longrightarrow \mathcal{G} \subset \mathbb{R}_{\geq 0}$ is a Lipschitz-continuous map, (Theorem 24), with $T \in \mathcal{N}$ and $\ell_c \in \mathcal{G}$. ($\mathcal{N}$ is defined in Theorem 24). Distance between any $t - t_1, t - t_2 \in \mathcal{N}$ is given by metric

$$d_{\mathcal{N}}(t - t_1, t - t_2) := |t_1 - t_2|.$$

Distance between $g_{c,\boldsymbol{x}}(t-t_1)$ and $g_{c,\boldsymbol{x}}(t-t_2)$ is given by metric

$$d_{\mathcal{G}}(g_{c,\boldsymbol{x}}(t-t_1), g_{c,\boldsymbol{x}}(t-t_2)) := \sqrt{-\log|corr(g_{c,\boldsymbol{x}}(t-t_1), g_{c,\boldsymbol{x}}(t-t_2))|},$$

s.t. $d_{\mathcal{G}}(g_{c,\boldsymbol{x}}(t-t_1), g_{c,\boldsymbol{x}}(t-t_2)) \geq 0$, and is finite (since $t-t_1, t-t_2$ live in a bound set, and $g_{c\boldsymbol{x}}(\cdot)$ is continuous). The parametrised model of the correlation is

$$|corr(g_{c,\boldsymbol{x}}(t-t_1), g_{c,\boldsymbol{x}}(t-t_2))| := K\left(\frac{(t_1-t_2)^2}{\delta_g^2}\right) \equiv \exp\left[-\frac{(t_1-t_2)^2}{\delta_g^2}\right],$$

s.t. $|corr(g_{c,\boldsymbol{x}}(t-t_1), g_{c,\boldsymbol{x}}(t-t_2))| \in (0,1]$, where $\delta_g > 0$ is the scale hyperparameter.

Now, Lipschitz-continuity of $g_{c,\boldsymbol{x}}(\cdot)$ implies

$$d_{\mathcal{G}}(g_{c,\boldsymbol{x}}(t-t_1), g_{c,\boldsymbol{x}}(t-t_2)) \leq L_g d_{\mathcal{N}}(t-t_1, t-t_2), \tag{12}$$

where the Lipschitz constant L_g is $g_{c,\boldsymbol{x}}$-dependent (Theorem 9). As $d_{\mathcal{N}}(t-t_1, t-t_2) \equiv |t_1-t_2| \leq t_0$, where t_0 is a known finite integer, and as $d_{\mathcal{G}}(\cdot,\cdot)$ is defined as $|t_1-t_2|/\delta_g$, $\delta_g > 0$ (using definition of $d_{\mathcal{G}}(\cdot,\cdot)$), L_g exists for t_1, t_2, and is finite. We get

$$L_g = \frac{1}{\delta_g}. \tag{13}$$

As $t-t_1, t-t_2$ is any point in $\mathcal{N}$, L_g exists for all points in $\mathcal{N}$.

Let set $\boldsymbol{L} := \{L_{g_1}, \ldots, L_{g_{t_0}}\}$, where L_{g_i} defines the Lipschitz-continuity condition (inequation 12) for the i-th sample function $g_i(\cdot)$ from a scalar-variate GP.

$$\exists L_{max} := \max_g[\boldsymbol{L}] = \max_g\{L_{g_1}, \ldots, L_{g_{t_0}}\}, \quad \text{where} \quad L_{max} > 0 \quad \text{and is finite.}$$

Thus, L_{max} is a Lipschitz constant that defines the Lipschitz continuity for any sampled function in $\{g_{c,\boldsymbol{x}}(t-t_1), \ldots, g_{c,\boldsymbol{x}}(t-t_0)\}$, at any iteration number t in a chain of finite and known number of iterations.

Then by Equation 13, $\exists\delta > 0$, s.t.

$$\delta := \max_g\left\{\frac{1}{\delta_{g_1}}, \ldots, \frac{1}{\delta_{g_{t_0}}}\right\} = \min_g\{\delta_{g_1}, \ldots, \delta_{g_{t_0}}\}; \quad \text{where } \delta_{g_i} > 0 \forall i = 1, \ldots, t_0.$$

Here $L_{g_i} = \dfrac{1}{\delta_{g_i}}$; $i = 1, \ldots, t_0$. □

Theorem 26 *Given $\ell_c = g_{c,\boldsymbol{x}}(t)$, where $g_{c,\boldsymbol{x}} : \mathcal{N} \longrightarrow \mathcal{G}$ is a Lipschitz-continuous function, sampled from a scalar-variate GP, the covariance function of which, computed at any 2 points $t-t_1, t-t_2$ in the input space $\mathcal{N}$, is kernel parametrised as*

$$Cov(t_1, t_2) = a_c K\left(\frac{(t_1-t_2)^2}{\delta_c^2}\right) \equiv a_c\left(\exp\left[-\frac{(t_1-t_2)^2}{\delta_c^2}\right]\right),$$

where (a_c, the amplitude hyperparameter and) the scale hyperparameter of this kernel is δ_c that is independent of the sample function $g_{c,\boldsymbol{x}}(\cdot)$; $c \in \{1, \ldots, d\}$.

Proof By Theorem 25, $\delta_c := \min\limits_{g_c}\{\delta_{g_{c,1}}, \ldots, \delta_{g_{c,n}}\}$ exists for any $c \in \{1, \ldots, d\}$. Then the scalar-variate GP that models the sample function $g_{c,\boldsymbol{x}}(\cdot)$ has a covariance kernel that is marked by the finite scale hyperparameter δ_c, independent of the sample function. □

That a universal scale hyperparameter δ that is independent of the sample path can define the covariance kernel of the scalar-variate GP that $g_{c,\boldsymbol{x}}(\cdot)$ is sampled from, owes to the fact that any such sample function $g_{c,\boldsymbol{x}}(\cdot)$ is continuous given the nature of the map (from a subset of integers to reals). However, when the sample function from a GP is not continuous, (such as $\boldsymbol{\xi}(\cdot)$ that is modelled with the tensor-variate GP discussed above), a set of values of the sample function-dependent scale hyperparameter(s) of the covariance kernel of the corresponding GP, will not admit a minima, and therefore, in such a case, a global scale hyperparameter cannot be ascribed to the covariance kernel of the generating GP. This is why we need to retain the correlation length scale hyperparameter ℓ_c to be dependent on the tensor-valued sample function $\boldsymbol{\xi}(\cdot)$, but the scale hyperparameter δ_c is no longer dependent on the scalar-valued sample function $g_{c,\boldsymbol{x}}(\cdot)$. In other words, we do not require to add any further layers to our learning strategy, than the two layers discussed.

We now revisit the issue of opting to express the relation between $\boldsymbol{V}$ and $\boldsymbol{S}$ as $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$, over the model in which $\boldsymbol{S} = \boldsymbol{f}(\boldsymbol{V})$, that we examined earlier in Section 2. As we had said in that section, learning the vector-valued function $\boldsymbol{f}(\cdot)$ within a Stochastic Process-based approach would suggest the need to learn many more length-scale hyperparameters of the kernel parametrised covariance matrix of the ensuing Matrix Normal likelihood; for example, a matrix-valued $\boldsymbol{V}$ with dimensionality $m_1 \times m_2$, would imply that $m_1 m_2$ length-scale hyperparameters are relevant to the kernel-parametrised covariance function. This is more than the d hyperparameters that we would need to learn if the model $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$ is pursued, as long as $d < m_1 m_2$. This alternative model gets even more discouraged when we decide to model the hyperparameters as unknown functions of the iteration number.

5.1 Learning with Compound Tensor-Variate & Scalar-Variate GPs

We find inference defined by a sequential sampling from the scalar-variate GPs (for each of the d directions of input space), followed by that from tensor-variate GP, directly relevant to the purpose at hand. Here, learning involves a Compound tensor-variate and multiple scalar-variate GPs; we refer below to such a Compound Stochastic Process, as a "$nested-GP$" model.

Remark 27 *As δ_c, a_c are not stochastic, hereon, we absorb the dependence of the function $g(\cdot)$ on the direction index, via the descriptor parameters, and refer to this function as $g_{\boldsymbol{x}_c}(t)$; $c = 1, \ldots, d$.*

Definition 28 *$Nested-GP$ model:*
for $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$,

$$\boldsymbol{\xi}(\cdot) \sim \textit{tensor-variate GP},$$

s.t. joint probability of n observations of $k-1$-th ordered tensor-valued variable $\boldsymbol{V}$ (that comprise training data $\mathbf{D}$), is k-th ordered Tensor Normal, with k covariance matrices–which are empirically estimated, or learnt directly using MCMC, or kernel parametrised, s.t. length scale parameter $\ell_1, \ldots, \ell_d$ of this covariance kernel, is each modelled as a dynamically *varying function* $\ell_c = g_{\boldsymbol{x}_c}(t)$*, where*

$$g_{\boldsymbol{x}_c}(t) \sim \ c - th \textit{ scalar-variate GP},$$

$\Longrightarrow$*joint probability of the last* t_0 *observations of* ℓ_c *(that comprise "lookback data"* $\mathbf{D}_{c,t}$*), is Multivariate Normal, the covariance function of which is parametrised by a kernel indexed by the c-th, stationary descriptor parameter vector* $\boldsymbol{x}_c = (a_c, \delta_c)^T$*, where* a_c *is the amplitude and* δ_c *the scale-length hyperparameter of the SQE-looking covariance kernel;* $c = 1, \ldots, d$*.*

Definition 29 $Nonnested - GP$ *model:*
for $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$*,*

$$\boldsymbol{\xi}(\cdot) \sim \textit{tensor-variate GP},$$

s.t. joint probability of observations of $\boldsymbol{V}$ *is k-th ordered Tensor Normal, with k covariance matrices–which are empirically estimated, or learnt directly using MCMC, or kernel parametrised, s.t. length scale parameter* $\ell_1, \ldots, \ell_d$ *of this covariance kernel, is each treated as a stationary unknown. All learning is undertaken using training data* $\mathbf{D}$*.*

6. Inference

Bayesian inference with Metropolis-within-Gibbs lends itself readily to this high-dimensional inferential exercise that relies on sequential updating of the relevant parameters. Below, we suggest inference that permits learning in the nested and non-nested models. Here $\theta^{(t\star)}$ indicates the proposed value of parameter θ in the t-th iteration, while $\theta^{(t)}$ refers to the value that is current in the t-th iteration, for any parameter θ that is being learnt.

- $Nested - GP$:

  1. In $t > t_0$-th iteration, propose amplitude and scale-length of c-th scalar-variate GP as:
  $$a_c^{(t\star)} \sim \mathcal{TN}(a_c^{(t-1)}, 0, v_a^{(c)}), \quad \forall c = 1, \ldots, d,$$
  $$\delta_c^{(t\star)} \sim \mathcal{N}(\delta_c^{(t-1)}, 0, v_\delta^{(c)}), \quad \forall c = 1, \ldots, d,$$
  where $\mathcal{N}(\cdot)$ is Normal, and $\mathcal{TN}(\cdot, 0, \cdot)$ is a Truncated Normal density left-truncated at 0.
  $v_a^{(c)}, v_\delta^{(c)}$ refer to constant, experimentally-chosen variances.
  2. As length scale hyperparameter $\ell_c = g_{\boldsymbol{x}_c}(t) \sim GP(0, \exp\left(-(\cdot - \cdot)^2/2\delta_c^2\right))$, probability of the current lookback data $\mathbf{D}_{c,t}$ given parameters of this c-th scalar-variate GP, is Multivariate Normal with mean vector $\mathbf{0}$ and a current covariance matrix $\boldsymbol{\Psi}_c^{(t-1)} := \left[a_c^{(t-1)} \exp\left(-\frac{(t_i - t_j)^2}{2(\delta_c^{(t-1)})^2}\right)\right]$; $t_i, t_j = t-1, \ldots, t-t_0$. Similarly, the likelihood of the proposed parameters can be defined. These enter computation of the acceptance ratio in the first block of Metropolis-within-Gibbs.
  3. At the updated parameters δ_c, a_c, at $T = t$, length scale hyperparameters $\ell_1, \ldots, \ell_d$ are rendered Normal variates s.t.
  $$\ell_c^{t\star} \sim \mathcal{N}(\ell_c^{(t-1)}, a_c^{(t\star)}),$$

under a Random Walk paradigm, when the mean of this Gaussian distribution is the current value of the ℓ_c parameter; $\forall c = 1, \ldots, d$.

4. The proposed and current values of $\ell_1, \ldots, \ell_d$ inform on the acceptance ratio in the 2nd block of our inference, along with other, directly learnt parameters, of the covariance structure of the tensor-variate GP that $\boldsymbol{x}(\cdot)$ is sampled from.

- $Nonnested - GP$:

  1. In the first block of Metropolis-within-Gibbs, $\ell_1, \ldots, \ell_d$ are updated, once proposed as Normal variates, with experimentally chosen constant variance of the respective proposal density.
  2. Updating of directly-learnt elements of relevant covariance matrices is undertaken in the 2nd block, and the acceptance ratio that invokes the tensor-normal likelihood, is computed to accept/reject these proposed values, at the ℓ_c variable values that are updated in the first block of Metropolis-within-Gibbs.

7. Application

7.1 Background

Astronomical theory tells us that if a set of stars is evolved from some primordial time, in the disk of the Milky Way galaxy, then the velocity vectors of the stars that land in an identified region of the Milky Way disk, at the current time, are affected by the feature parameters of the Milky Way. Thus, if the velocity vectors of these stars are collated to form the velocity matrix $\boldsymbol{V}$, then $\boldsymbol{V}$ is affected by the feature parameter vector $\boldsymbol{S}$ of the Galaxy. So we model that $\boldsymbol{V}$ and $\boldsymbol{S}$ exist in a relationship, and express this by stating $\boldsymbol{V} = \boldsymbol{\xi}(\boldsymbol{S})$, where $\boldsymbol{\xi}(\cdot)$ is an unknown function. Focusing on the immediate neighbourhood of the Sun in the Milky Way disk, the velocity matrix of a sample of our stellar neighbours was observed by the *Hipparcos* satellite, though we have no measurement on the value of $\boldsymbol{S}$ at which this observed value of $\boldsymbol{V}$ was recorded. On the other hand, astronomical simulations (Chakrabarty, 2007) can generate the value of the matrix $\boldsymbol{V}$, at a chosen value of $\boldsymbol{S}$, though astronomical simulations/models cannot predict the value of $\boldsymbol{S}$ at which a value of $\boldsymbol{V}$ is realised, i.e. reliable estimation/astronomical-modelling of the feature parameters of the Milky Way is not possible, given the observed velocity matrix that comprises velocity information of some of our stellar neighbours. Then to undertake the prediction of this value of $\boldsymbol{S}$, we aim to first learn $\boldsymbol{\xi}(\cdot)$ – by modelling it as a random realisation from a GP – and then undertake the inverse prediction of the value of $\boldsymbol{S}$ at which the observed value of $\boldsymbol{V}$ (test data) is realised. The training data used to learn $\boldsymbol{\xi}(\cdot)$ is provided by astronomical simulations that generate the value $\boldsymbol{V} = \boldsymbol{v}_i$ at the i-th design point $\boldsymbol{S} = \boldsymbol{s}_i$; $i = 1, \ldots, n$.

In these astronomical simulations, the sought Milky Way feature parameters include the angular speed of the bar – which is an elongated (triaxially-shaped) structure built of stars, that rotates, pivoted at the centre of the Galaxy, affecting the dynamics of the Milky Way. In fact, the spiral pattern of the Milky Way is considered rotating at a fraction of the bar's angular speed, where this fraction is fixed (to about one-third) in the astronomical base model used to undertake these simulations. The second sought feature parameter is the angular distance of the long-axis of the bar, to the line that joins the Sun to the Milky Way centre, as this angle also affects Milky Way dynamics. All other Milky Way parameters are input as known constants in the base astronomical model that is used to undertake these simulations. In fact, galactic astronomy allows for the existence of a known relation between the bar angular speed at the radial distance of the Sun from the centre of the Milky Way (discussed in Section 9.4). So, the sought Milky Way feature parameters are: (bar rotational frequency ≡) radial location S_1 of the Sun from the Milky Way centre, and the angular separation S_2 of the Sun from a fiduciary axis in the Milky Way disk, namely the bar long-axis.

Data on values of velocity matrix $\boldsymbol{V}$, generated at the n different values of the solar location vector $\boldsymbol{S} = (S_1, S_2)^T$ in the astronomical simulations, are s.t. there are *discontinuities in the data* distribution of the generated $\boldsymbol{V}$ values in the space of $\boldsymbol{S}$, as displayed in Figure 8 of Chakrabarty (2007).

7.2 Details

In this application, the aim is to learn location vector of the Sun in the Milky Way modelled as a 2-dimensional disk. The training data $\mathbf{D}$ is cuboidally-shaped, and is of dimensionalities $m_1 \times m_2 \times m_3$, where dimension of a stellar velocity vector is $m_1 = 2$; number of stars for which velocity vectors are generated at each design point is $m_2 = 50$; number of design points is $m_3 \equiv n = 216$, i.e. the 3-rd ordered tensor $\boldsymbol{D}_{\boldsymbol{V}}$ comprises of $n = 216$ matrices of dimension 50×2, where i-th value of the matrix-variate observable $\boldsymbol{V}^{(50\times 2)}$ is realised at i-th value of system parameter vector $\boldsymbol{S}$, s.t. $\mathbf{D} = \{(\boldsymbol{s}_i, \boldsymbol{v}_i)\}_{i=1}^{n}$. The 3rd-ordered tensor $\boldsymbol{D}_{\boldsymbol{V}}^{(m_1\times m_2\times n)} = (\boldsymbol{v}_1, \vdots, \ldots, \vdots \boldsymbol{v}_n)$

Numerical simulations are conducted with $n = 216$ different astronomical models of the Galaxy, with each such model of the Galaxy distinguished by a value of the Milky Way feature parameter vector $\boldsymbol{S} \in \mathbb{R}^d$, d=2 (Chakrabarty, 2007). Thus, $\boldsymbol{V} = \boldsymbol{v}_i$ at the i-th design point $\boldsymbol{s}_i$, $i = 1, \ldots, 216$.

In particular, there exists the test data $\boldsymbol{v}^{(test)}$ that comprises the m_1 = 2-dimensional velocity vectors of the 50 identified, stellar neighbours of the Sun, as measured by the Hipparcos satellite (Chakrabarty, 2007). It is the same 50 stars for which velocity vectors are simulated at each design point. However, we do not know the real Milky Way feature parameter vector $\boldsymbol{s}^{(test)}$ at which $\boldsymbol{V} = \boldsymbol{v}^{(test)}$ is realised.

Chakrabarty (2007) generated the training data by first placing a regular 2-dimensional polar grid over a chosen annulus in an 2-dimensional astronomical model of the MW disk. In the centroid of each grid cell, an observer is then placed. There are n grid cells; so there are n observers placed in this grid, such that the i-th observer measures velocities of m_{2i} stars that land in that grid cell, at the end of a simulated evolution of a sample of stars that are evolved in this model of the MW disk, under the influence of the feature parameters that mark this MW model. The m_{2i} number of stars are indexed by their location with respect to the observer inside the grid cell, and a random sample of $m_2 = 50$ stars from this collection of m_{2i} stars is taken; $i = 1, \ldots, n = 216$. Thus, each observer records a matrix (or sheet) of 2-dimensional velocity vectors of m_2 stars. The test data measured by the Hipparcos satellite is then the 217-th sheet, except we are not aware of the value of $\boldsymbol{S}$ that this sheet is realised at.

In Chakrabarty et al. (2015), the matrix of velocities was vectorised, so that the observable was then a vector. In this application, the observable $\boldsymbol{V}$ is a matrix. The process of vectorisation, causes Chakrabarty et al. (2015) to face loss of correlation information since vectorisation of the stellar velocity matrix implies the same correlation assigned to all velocity components of all stars, which may be a misrepresentation. However, the high-dimensional learning strategy that we discuss above, allows for clear quantification of such covariances. More importantly, such a strategy provides a clear template for implementing learning given high-dimensional data that comprise measurements of a tensor-valued observable. As mentioned above, the empirical estimate of the mean tensor is obtained, and used as the mean of the Tensor Normal density that represents the likelihood.

To learn $\boldsymbol{\xi}(\cdot)$, we model it as a realisation from a high-dimensional GP, s.t. joint of n values of $\boldsymbol{\xi}(\cdot)$–computed at $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$–is 3rd-order Tensor Normal, with 3 covariance matrices: that inform on:
–amongst-observer-location covariance ($\boldsymbol{\Sigma}_3^{(216\times 216)}$);
–amongst-stars-at-different-relative-position-w.r.t.-observer covariance ($\boldsymbol{\Sigma}_2^{(50\times 50)}$);
–amongst-velocity-component covariance ($\boldsymbol{\Sigma}_1^{(2\times 2)}$).

The elements of $\boldsymbol{\Sigma}_2$ are not learnt by MCMC.
–Firstly, there is no input space variable that can be identified, at which the ij-th element of $\boldsymbol{\Sigma}_2$ can be considered to be realised; $i, j = 1, \ldots, 50$, where this ij-th element gives the covariance amongst the i-th and j-th, 216×2-dimensional matrices within the 3-rd ordered tensor $\boldsymbol{D_V}$. Effectively, the 41st star could have been referred to as the 3rd star in this stellar sample, and the vice versa, i.e. there is no meaningful ordering in the labelling of the sampled stars with these indices. So we cannot use these labels as values of an input space variable, in terms of which, covariance between i-th and j-th 216×2-dimensional velocity matrices can be kernel-parametrised.
–Secondly, direct learning of the $50(51)/2$ distinct elements of $\boldsymbol{\Sigma}_2$, using MCMC, is ruled out, given that this is a large number.
–Given this, we will perform the plugin, i.e. the empirical estimation of $\boldsymbol{\Sigma}_2$.

Definition 30 *Covariance between the 216×2-dimensional stellar velocity matrix $\boldsymbol{W}_i := [v_{pq}^{(i)}]$ of the sampled star labelled by index i, and the matrix $\boldsymbol{W}_j := [v_{pq}^{(j)}]$ of the star labelled as j, ($p = 1, \ldots, 216; q = 1, 2$), is estimated as $\widehat{\sigma_{ij}^{(2)}}$, where:*

$$\widehat{\sigma_{ij}^{(2)}} = \frac{1}{2-1} \times \sum_{q=1}^{2} \left[\frac{1}{216} \times \left(\sum_{p=1}^{216} (v_{pq}^{(i)} - \bar{v}_q^{(i)}) \times (v_{pq}^{(j)} - \bar{v}_q^{(j)}) \right) \right],$$

where $\bar{v}_q^{(i)} = \dfrac{\left(\sum_{p=1}^{216} v_{pq}^{(i)}\right)}{216}$ is the sample mean of the q-th column of the matrix $\boldsymbol{V}_i = [v_{pq}^{(i)}]$.

The 3 distinct elements of the 2×2-dimensional covariance matrix $\boldsymbol{\Sigma}_1$ are learnt directly from MCMC. These include the 2 diagonal elements $\sigma_{11}^{(1)}$, $\sigma_{22}^{(1)}$ and $\rho := \dfrac{\sigma_{12}^{(1)}}{\sqrt{\sigma_{11}^{(1)}\sigma_{22}^{(1)}}}$

$\boldsymbol{\Sigma}_3$ is kernel-parametrised, using the SQE kernel such that the jp-th element of $\boldsymbol{\Sigma}_3$ is kernel-parametrised as $[\sigma_{jp}] = \exp\left(-(\boldsymbol{s}_j - \boldsymbol{s}_p)^T \boldsymbol{Q}^{-1} (\boldsymbol{s}_j - \boldsymbol{s}_p)\right)$ $j, p = 1, \ldots, 216$. Since $\boldsymbol{S}$

is a 2-dimensional vector, $\boldsymbol{Q}$ is a 2× 2 square diagonal matrix, elements ℓ_1, ℓ_2 of which, represent the correlation length scales.

Then in the "*nonnested−GP*" model, we learn the (modelled as stationary) ℓ_1, ℓ_2, along with $\sigma^{(1)}_{11}$, $\sigma^{(1)}_{22}$ and ρ.

Under the $nested-GP$ model, ℓ_c is modelled as $\ell_c = g_{\boldsymbol{x}_c}(t)$, where at iteration number $T = t$ $g_{\boldsymbol{x}_c}(t)$ is sampled from the c-th zero-mean, scalar variate GP, amplitude a_c and correlation length scale δ_c of which we learn, for $c = 1, 2$, in addition to the parameters $\sigma^{(1)}_{11}$, $\sigma^{(1)}_{22}$ and ρ.

The likelihood of the training data given the covariance matrices of the tensor-variate GP, is then given as per Equation 1:

$$\begin{aligned}\mathcal{L}(\mathbf{D}|\ell_1, \ell_2, \sigma^{(1)}_{11}, \sigma^{(1)}_{22}, \rho) = (2\pi)^{-m/2}(\prod_{i=1}^{3} |\boldsymbol{\Sigma}_i|^{-m/2m_i})\\ \times \exp(-\|(\boldsymbol{D}_{\boldsymbol{V}} - \hat{\boldsymbol{M}}) \times_1 {\boldsymbol{A}_1}^{-1} \times_2 {\hat{\boldsymbol{A}}_2}^{-1} \times_3 \boldsymbol{A}_3^{-1}\|^2/2).\end{aligned} \tag{14}$$

where $\boldsymbol{\Sigma}_p = \boldsymbol{A}_p\boldsymbol{A}_p^T$, $p = 1, 2, 3$ and $\hat{\boldsymbol{M}}$ is the empirical estimate of the mean tensor and $\hat{\boldsymbol{\Sigma}}_2$ is the empirical estimate of the covariance matrix $\boldsymbol{\Sigma}_2$ such that $\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{A}}_2{\hat{\boldsymbol{A}}_2}^T$. Here $m_3 = 216$, $m_2 = 50$, $m_1 = 2$, and $m = m_1m_2m_3$. One or more of the covariance matrices is kernel parametrised, where the kernel is a function of pairs of values of the input variable $\boldsymbol{S}$–this explains the dependence of the RHS of this equation on the whole of $\mathbf{D}$, with the data tensor $\boldsymbol{D}_{\boldsymbol{V}}$ contributing partly to training data $\mathbf{D}$.

This allows us to write the joint posterior probability density of the unknown parameters given training data $\mathbf{D}$. To write this posterior, we impose non-informative priors $\pi_0(\cdot)$ on each unknown (Gaussian with wide, experimentally chosen variances, and mean that is the arbitrarily chosen seed value of $\ell_\cdot$; Jeffry's priors on $\boldsymbol{\Sigma}_1$). The posterior probability density of the unknown GP parameters, given the training data is then

$$\pi(\ell_1, \ell_2, \sigma^{(1)}_{11}, \sigma^{(1)}_{22}, \rho|\mathbf{D}) \propto \mathcal{L}(\boldsymbol{D}_{\boldsymbol{V}}|\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_3) \times \pi_0(\ell_1)\pi_0(\ell_2)\pi_0(\boldsymbol{\Sigma}_1). \tag{15}$$

The results of this learning and estimation of the covariance and mean of the GP invoked in the modelling, are discussed below in Section 9.

Definition 31 *The joint posterior probability density of the unknown parameters given the training data* $\mathbf{D}$ *that comprises the velocity tensor* $\boldsymbol{D}_{\boldsymbol{V}}$*, under the* $nested - GP$ *model is given by*

$$\begin{aligned}\pi(\delta_1, \delta_2, a_1, a_2.\ell_1, \ell_2, \sigma^{(1)}_{11}, \sigma^{(1)}_{22}, \rho|\mathbf{D}) \propto (2\pi)^{-m/2}\left(\prod_{i=1}^{3} |\boldsymbol{\Sigma}_i|^{-m/2m_i}\right)\\ \times \exp(-\|(\boldsymbol{D}_{\boldsymbol{V}} - \hat{\boldsymbol{M}}) \times_1 {\boldsymbol{A}_1}^{-1} \times_2 {\hat{\boldsymbol{A}}_2}^{-1} \times_3 \boldsymbol{A}_3^{-1}\|^2/2)\times\\ \prod_{c=1}^{2} \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Psi}_{\boldsymbol{x}_c})}} \exp\left[-\frac{1}{2}(\boldsymbol{\ell}_c^{(t_0)})^T\,(\boldsymbol{\Psi}_{\boldsymbol{x}_c})^{-1}\,(\boldsymbol{\ell}_c^{(t_0)})\right] \times \pi_0(\boldsymbol{\Sigma}_1),\end{aligned} \tag{16}$$

where $\boldsymbol{\ell}_c^{(t_0)} := (\ell_c^{(t-t_0)}, \ldots, \ell_c^{(t-1)})^T$, *and* ij*-th element of the covariance matrix* $\boldsymbol{\Psi}_{\boldsymbol{x}_c}$ *is* $\left[a_c \exp\left[-\dfrac{(t_i - t_j)^2}{2(\delta_c)^2}\right]\right]$, $i, j = 1, \ldots, t_0$. *N.B. the* t*-dependence of the covariance matrix* $\boldsymbol{\Psi}_{\boldsymbol{x}_c}$ *is effectively suppressed, given that this dependence comes in the form* $t - t_i - (t - t_j)$.

We generate posterior samples using Metropolis-within-Gibbs, to identify the marginal posterior probability distribution of each unknown. The marginal then allows for the computation of the 95% HPD.

8. Inverse Prediction–2 Ways

We aim to predict the location vector $\boldsymbol{s}^{(test)}$ of the Sun in the Milky Way disk, at which real (test) data $\boldsymbol{v}^{(test)}$ on the 2-dimensional velocity vectors of 50 identified stellar neighbours of the Sun, measured by the *Hipparcos* satellite. We undertake this, subsequent to learning of relation $\boldsymbol{\xi}(\cdot)$ between solar location variable $\boldsymbol{S}$ and stellar velocity matrix-valued variable $\boldsymbol{V}$, using astronomically-simulated (training data).

Definition 32 *The tensor that includes both test and training data has dimensions of* $217 \times 50 \times 2$. *We call this augmented data* $\boldsymbol{D}^* = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_{50}, \boldsymbol{v}^{(test)}\}$, *to distinguish it from the tensor* $\boldsymbol{D}_{\boldsymbol{V}}$ *that comprises the training data.* $\boldsymbol{v}_i$ *is realised at design point* $\boldsymbol{s}_i$, *but* $\boldsymbol{s}^{(test)}$, *at which* $\boldsymbol{v}^{(test)}$ *is realised, is unknown.*

Remark 33 *This 217-th sheet of (test) data is realised at the unknown value* $\boldsymbol{s}^{(test)}$ *of* $\boldsymbol{S}$, *and upon its inclusion, the updated covariance amongst the sheets generated at the different values of* $\boldsymbol{S}$, *is renamed* $\boldsymbol{\Sigma}_1^*$, *which is now rendered* 217×217*-dimensional. Then* $\boldsymbol{\Sigma}_1^*$ *includes information about* $\boldsymbol{s}^{(test)}$ *via the kernel-parametrised covariance matrix* $\boldsymbol{\Sigma}_3$. *The effect of inclusion of the test data on the other covariance matrices is less; we refer to them as (empirically estimated)* $\hat{\boldsymbol{\Sigma}}_2^*$ *and* $\boldsymbol{\Sigma}_3^*$. *The updated (empirically estimated) mean tensor is* $\hat{\boldsymbol{M}}^*$.

The likelihood for the augmented data is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{D}^* | \boldsymbol{s}^{(test)}, \boldsymbol{\Sigma}_1^*, \boldsymbol{\Sigma}_3^*) =& (2\pi)^{-m/2} \left(\prod_{i=1}^{3} |\boldsymbol{\Sigma}_i^*|^{-m/2m_i}\right) \times \\ & \exp\left[-\|(\boldsymbol{D}^* - \hat{\boldsymbol{M}}^*) \times_1 (\boldsymbol{A}_1^*)^{-1} \times_2 (\hat{\boldsymbol{A}}_2^*)^{-1} \times_3 (\boldsymbol{A}_3^*)^{-1}\|^2/2\right]\end{aligned} \tag{17}$$

where $\hat{\boldsymbol{A}}_2^*$ is the square root of $\hat{\boldsymbol{\Sigma}}_2^*$. Here $m_1 = 217$, $m_2 = 50$, $m_3 = 2$, and $m = m_1 m_2 m_3$. Here $\boldsymbol{A}_1^*$ is the square root of $\boldsymbol{\Sigma}_1^*$ and depends on $\boldsymbol{s}^{(test)}$.

The posterior of the unknowns given the test+training data is:

$$\begin{aligned}\pi(s_1^{(test)}, s_2^{(test)}, \boldsymbol{\Sigma}_1^*, \boldsymbol{\Sigma}_3^* | \boldsymbol{D}^*) \propto & \mathcal{L}(\boldsymbol{D}^* | s_1^{(test)}, s_2^{(test)}, \boldsymbol{\Sigma}_1^*, \boldsymbol{\Sigma}_3^*) \times \\ & \pi_0(s_1^{(test)}) \pi_0(s_2^{(test)}) \pi_0(q_2^{(*)}) \pi_0(q_1^{(*)}) \pi_0(\boldsymbol{\Sigma}_3^*).\end{aligned} \tag{18}$$

Remark 34 *In this application, we use the prior probability density* $\pi_0(s_p^{(test)}) = \mathcal{U}(l_p, u_p), p = 1, 2$*, where* l_p *and* u_p *are chosen depending on the spatial boundaries of the fixed area of the Milky Way disk that was used in the astronomical simulations by Chakrabarty (2007). Recalling that the observer is located in a quadrant of a two-dimensional polar grid, Chakrabarty (2007) set the lower boundary on the angular position of the observer to 0, and upper to* $\pi/2$ *radians, where the observer's angular coordinate is the angle made by the observer-Galactic centre line to a chosen line in the MW disk. The observer's radial location is maintained within the interval [1.7, 2.3] in model units, where model units for length are related to galactic unit for length, as discussed in Section 9.4.*

In the second method for prediction, we infer $\boldsymbol{s}^{(test)}$ by sampling from the posterior of $\boldsymbol{s}^{(test)}$ given the test data and the modal values of the parameters $q_1, q_2, \sigma_{11}^{(1)}, \rho, \sigma_{22}^{(1)}$ that were learnt using the training data. Let modal value of $\boldsymbol{\Sigma}_3$, learnt using $\mathbf{D}$ be $[(\sigma_3^{(M)})_{jp}]_{j=1;p=1}^{217,217}$, Similarly, the modal value $\boldsymbol{\Sigma}_1^{(M)}$ that was learnt using the training data, is used. The posterior of $\boldsymbol{s}^{(test)}$, at learnt (modal) values is then

$$\begin{aligned}&\pi(s_1^{(test)}, s_2^{(test)} | \boldsymbol{D}^*, \boldsymbol{\Sigma}_1^{(M)}, \boldsymbol{\Sigma}_3^{\star}) \propto \\ &\mathcal{L}(\boldsymbol{D}^* | s_1^{(test)}, s_2^{(test)}, \boldsymbol{\Sigma}_1^{(M)}, \boldsymbol{\Sigma}_3^{\star}) \times \pi_0(s_1^{(test)})\pi_0(s_2^{(test)}) \times \\ &\pi_0(q_2^{(M)})\pi_0(q_1^{(M)})\pi_0(\boldsymbol{\Sigma}_3)|\boldsymbol{V}^*).\end{aligned} \tag{19}$$

where $\mathcal{L}(\boldsymbol{D}^* | s_1^{(test)}, s_2^{(test)}, \boldsymbol{\Sigma}_1^*, \boldsymbol{\Sigma}_3^{(M)})$ is as given in Equation 14, with $\boldsymbol{\Sigma}_3$ replaced by $\boldsymbol{\Sigma}_3^*$, and $\boldsymbol{\Sigma}_1$ replaced by its modal value $\boldsymbol{\sigma}_1^{(M)}$. The priors on $s_1^{(test)}$ and $s_2^{(test)}$ are as discussed above. For all parameters, we use Normal proposal densities that have experimentally chosen variances.

9. Results

In this section, we present results of the learning methodology described above, implemented within the considered application. Thus, results of learning the unknown parameters of the 3rd-order tensor-normal likelihood, given training as well as training+test data, are discussed here.

While Figure 1 in the in Appendix A, and Figure 1 depict results obtained from using the $nonnested-GP$, in the following figures, results of the learning of all relevant unknown parameters, using the $nested-GP$ model, are included. Figures that depict results from the $nested-GP$ model include results of the learning of amplitude a_c and smoothing parameters $d_c := 1/\delta_c$ parameters. Also, our modelling under the $nested - GP$ paradigm relies on a lookback-time t_0 which gives the number of iterations over which we gather generated ℓ_c values.

9.1 Effect of *discontinuity in the data*, manifest in our results

One difference between the learning of parameters from the $nested - GP$, as distinguished from the $nonnested - GP$ models is the quality of the inference, in the sense that the uncertainty of parameters (i.e. the 95% HPDs) learnt using the $nested-GP$ models, is less

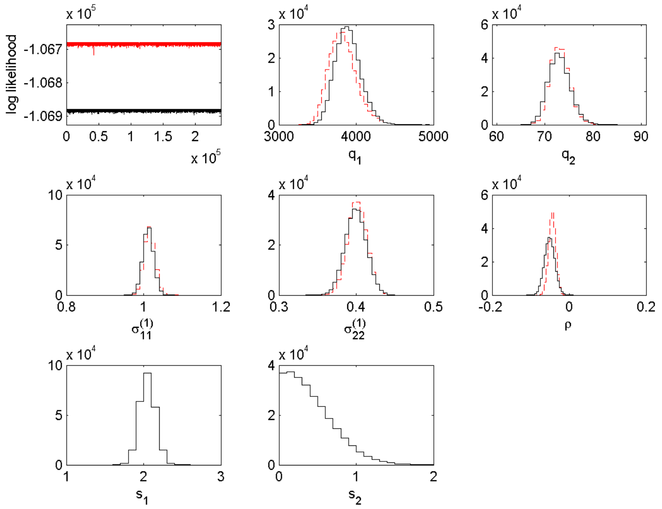

Figure 1: Results from chain run with training data $\mathbf{D}$ with the $nonnested - GP$ model, are shown in grey (or red in the electronic version), while results from chain undertaken with training and test data, $\boldsymbol{D}^{\star}$, in this $nonnested - GP$ model, are depicted in black. Traces of the logarithm of the likelihood are displayed from the two chains in the top left panel. Reciprocals of length scale parameters are the shown in the top middle and right panels; here $q_c = \ell_c^{-1}$, $c = 1, 2$. Histograms representing marginal posterior probability density of the learnt diagonal elements $\sigma_{11}^{(1)}$ and $\sigma_{22}^{(1)}$, of the covariance matrix $\boldsymbol{\Sigma}_1$, are shown in the mid-row, left and middle panels (given respective data). Histograms representing marginals of parameter $\rho = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}^{(1)}\sigma_{22}^{(1)}}}$ are displayed in the mid-row right panel. Prediction of the values of the input parameter $\boldsymbol{S} = (S_1, S_2)^T$ is possible only in the run performed with both training and test data. Marginals of S_1 and S_2 values learnt via MCMC-based sampling from the joint of all unknown parameters given $\boldsymbol{D}^{\star}$, are shown in the lower panel, as approximated by histograms.

than that learnt using the $nonnested - GP$ models. This difference in the learnt HPDs is most marked for the learning of values of Q_1 and S_1, and S_2 to a lesser extent.

To analyse possible discontinuities in the training data used in the application, we refer to Figure 8 of Chakrabarty (2007), page 152 of the figure that is available at `https://www.aanda.org/articles/aa/pdf/2007/19/aa6677-06.pdf`. This figure informs on the following. Compatibility of the value $\boldsymbol{v}$ of the stellar velocity matrix variable, realised in these astronomical simulations at a given $\boldsymbol{s}$, to the test velocity matrix $\boldsymbol{v}^{(test)}$ (recorded by the *Hipparcos* satellite), is parametrised. This compatibility parameter is plotted as a function of the two components s_1 and s_2 of the vector $\boldsymbol{s}$, and plotted as a contour plot in the space $\mathcal{D}$, where

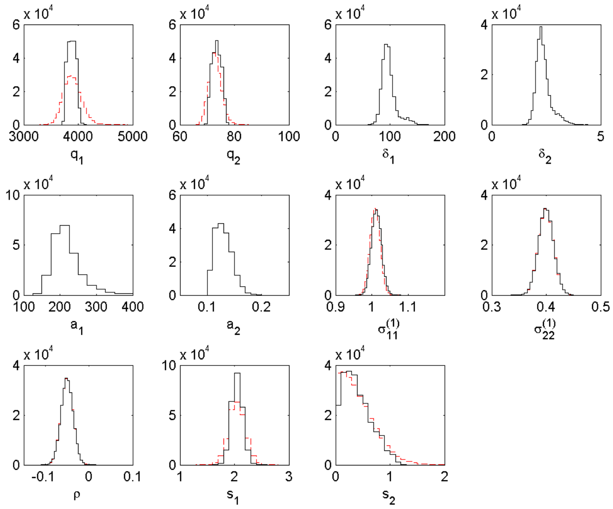

Figure 2: Results from chain run with test+training data $\boldsymbol{D}^{\star}$ within the $nested{-}GP$ model, are shown in black, as distinguished from the results of learning given the same data, using the $nonnested - GP$ model, that are depicted in grey (or red in the electronic version). Here the used value of T_0 is 200 iterations. Histograms approximating the marginal posterior probability densities of each sought unknown is depicted. Here, sought hyperparameter values a_c and δ_c are relevant only to the $nested - GP$ model ($c = 1, 2$). Here, we have undertaken sampling from the joint posterior of all parameters, including the input parameter values $s_1^{(test)}$ and $s_2^{(test)}$, at which the test data are realised. Histograms approximating marginal posterior of each learnt unknown are presented.

$\boldsymbol{S} \in \mathcal{D} \subset \mathbb{R}^2$. Here, the 2 components of $\boldsymbol{S}$ are represented in polar coordinates, with S_1 the radial and S_2 the angular component. We see clearly from this figure, that the distribution across S_1 is highly discontinuous, at given values of S_2 (i.e. at fixed angular bins). In fact, this distribution is visually more discontinuous than the distribution across S_2, at given values of S_1, i.e. at fixed radial bins (each of which is represented by the space between two bounding arcs). In other words, the velocity matrices that are astronomically simulated at different $\boldsymbol{S}$ values, are differently compatible with a given reference velocity matrix, (such as $\boldsymbol{v}^{(test)}$). Distribution of the velocity matrix variable $\boldsymbol{V}$, is discontinuous across values of $\boldsymbol{S}$, and in fact, less smoothly distributed at fixed s_2, than at fixed s_1. Thus, this figure brings forth the discontinuity with the input-space variable $\boldsymbol{S}$, in the data tensor $\boldsymbol{D}_{\boldsymbol{V}}$ that is part of the training data.

Then, it is incorrect to use a stationary kernel to parametrise the covariance $\boldsymbol{\Sigma}_3$, that informs on the covariance between velocity matrices generated at different values of $\boldsymbol{S}$. Our implementation of the $nested - GP$ model tackles this shortcoming of the model. Thus, when we implement the $nonnested{-}GP$ model, Metropolis needs to explore a wider volume of the state space to accommodate parameter values, given the data at hand–and even then,

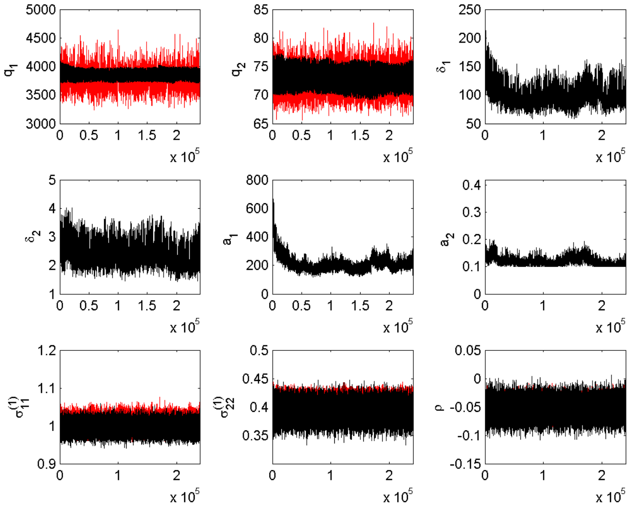

Figure 3: Traces of parameters learnt using the training data **D**, in the run performed with the $nonnested - GP$ model, are compared to traces of the corresponding parameter obtained in the run performed with the $nested - GP$ model. Traces of parameters learnt within the $nonnested - GP$ model are in grey (or red in the e-version) while the traces obtained using the $nested - GP$ model are shown in black.

.

there is a possibility for incorrect inference under the stationary kernel model. This explains the noted trend of higher 95% HPDs on most parameters learnt using the $nonnested - GP$ model, compared to the $nested - GP$ model, as observed in comparison of results from runs done with training data alone, or both training and test data; compare Figure 2 to Figure 3, and note the comparison in the traces as displayed in Figure 3. Indeed, this also explains the bigger difference noted in these figures when we compare the learning of q_1 over q_2, in runs that use the stationary model, as distinguished from the non-stationary model. After all, the discontinuity across S_1 is discussed above, to be higher than across S_2.

9.2 Effect of varying lookback times, i.e. length of historical data

To check for the effect of the lookback time t_0, we present traces of the covariance parameters and kernel hyperparameters learnt from runs undertaken within the $nested - GP$ model, with different t_0 values of 50 and 100, in Figure 4, which we can compare to the traces obtained in runs performed under the $nested - GP$ model, with $t_0 = 200$, as displayed in Figure 3.

It is indeed interesting to note the trends in traces of the the smoothness parameters $q = 1/\ell$ parameters; values of the amplitude (a_1, a_2) parameters; and values of the length scale hyperparameters (δ_1, δ_2), evidenced in Figure 4 and in black in Figure 3). A zeroth-order model for these parameters that are realisations from a non-stationary process, is

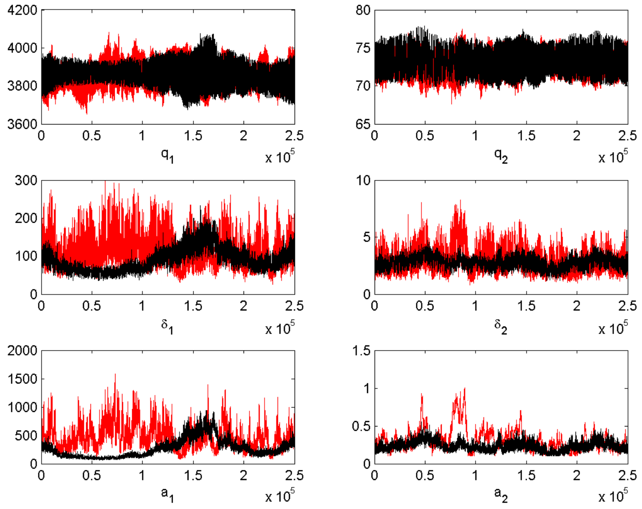

Figure 4: Comparison of traces of unknown smoothness parameters of $\boldsymbol{\Sigma}_3$ and hyperparameters of GPs invoked to model these parameters, obtained in runs performed with training data $\mathbf{D}$ and $t_0 = 50$ (in grey, or red in the e-version) and $t_0 = 100$ (in black).

a moving averages time-series model – $MA(t_0)$ to be precise. We note the increase in fluctuation amplitude of the traces, with decreasing t_0. For smaller values of lookback time t_0, the average covariance between $g_{\boldsymbol{x}_c}(t_1)$ and $g_{\boldsymbol{x}_c}(t_2)$ is higher, than when t_0 is higher, where the averaging is performed over a t_0-iteration long interval that has its right edge on the current iteration; here $\boldsymbol{x}_c = (a_c, \delta_c)^T$, $c = 1, 2$ and as introduced above, we model the length scale parameter of the kernel that parametrises $\boldsymbol{\Sigma}_3$, as $\ell_c = g_{\boldsymbol{x}_c}(t)$. Here $g_{\boldsymbol{x}_c}(\cdot)$ is modelled as a realisation from a scalar-variate GP with covariance kernel that is itself kernel-parametrised using an SQE kernel with amplitude a_c and correlation-length δ_c. Then higher covariances between values of $g_{\boldsymbol{x}_c}(\cdot)$ at different t-values in general would suggest higher values of global amplitude of this parametrised kernel, and higher values of length-scales of this SQE kernel.

Indeed an important question is, what is the "best" t_0, given our data. Such question is itself of relevance, and discussed intensively under distributed lag models, often within Econometrics (Shirley, 1965). An interesting trend noted in the parameter traces presented in Figure 4 for $t_0 = 50, 100$, and to a lesser extent for $t_0 = 200$, in the results in black in Figure 3, is the global near-periodic existence of crests and troughs in these traces. This periodic fluctuation is more marked for smoothness q_1 $(=1/\ell_1)$ and the hyperparameters of the scalar-variate GP used to model $g_{\boldsymbol{x}_1}(\cdot)$, than for q_2 (and a_2 and δ_2).

From the point of view of a polynomial (of order t_0) model for the lag operator – that transfers information from the past t_0 realisations from a stochastic process to the current iteration – the shape of the trace will be dictated by parameters of this model. If this polynomial admits complex roots, then coefficients of the relevant lag terms will behave like a damped sine function with iterations. For a different value of t_0, such a pronounced oscillatory trend might not be equally apparent. Loosely speaking, the value of ℓ_c in any iteration, represented by a moving average, will manifest the result of superposition of the different (discontinuous) modal neighbourhoods present in the data. The more multimodal the data, i.e. larger the number of "classes" (by correlation-length scales) of functional form $\boldsymbol{\xi}(\cdot)$ sampled from the tensor-variate GP, s.t. superposition of the sample paths will cause a washing-out of the effect of the different modes, and a less prominent global trend will be manifest in the traces. However, for data that is globally bimodal, the superposition of the two "classes" of sampled functions $\boldsymbol{\xi}(\cdot)$ will create a periodicity in the global trend of the generated ℓ_c values (and thereby of the smoothness parameter values q_c, where $q = \ell_c^{-1}$). Again, the larger the value t_0 of the lookback-time parameter, the moving average is over a larger number of samples, and hence greater is the washing-out effect. Thus, depending on the *discontinuity in the data*, it is anticipated that there is a range of optimal lookback-time values, for which, the global periodicity is most marked. This is what we might be noticing in the trace of q_1 at $t_0 = 100$ displaying the global periodicity more strongly than that at $t_0 = 200$ (see Figure 4 and Figure 3).

Another point is that the strength of this global periodicity will be stronger for the correlation-length scale along that direction in input-space, the discontinuity along which is stronger. Indeed, as we have discussed above, the *discontinuity in the data* with varying S_1 is anticipated to be higher than with S_2. So we would expect a more prominent periodic trend in the trace of q_1 than q_2. This is indeed what to note in Figure 4. A simulation study can be undertaken to explore the effects of empirical discontinuities.

The arguments above qualitatively explain the observed trends in the traces of the hyperparameters, obtained from runs using different t_0. That in spite of discrepancies in a_c and δ_c, with t_0, values of the length scale parameter ℓ_c (and therefore its reciprocal q_c) are concurrent within the 95% HPDs, is testament to the robustness of inference. Stationarity of the traces betrays the achievement of convergence of the chain.

Table 1 below, and Table 1 in Appendix A include results on the learning of each parameter, under every considered model, and parameter prediction undertaken, tabulated as 95% HPD credible regions. In addition, Table 2 in Appendix A presents the effect of inference made with 3 different values of the lookback time.

We notice that the reciprocal correlation length scale q_1 is a couple of orders of magnitude higher than q_2; correlation between values of the sampled function $\boldsymbol{\xi}(\cdot)$, at 2 different S_1 values (at the same s_2), then wanes more quickly than correlation between sampled functions computed at same s_1 and different S_2 values. Here $\boldsymbol{s} = (s_1, s_2)^T$ and given that $\boldsymbol{S}$ is the location of the observer who observes the velocities of her neighbouring stars on a two-dimensional polar grid, S_1 is interpreted as the radial coordinate of the observer's location in the Galaxy and S_2 is the observer's angular coordinate. Then it appears that the velocities measured by observers at different radial coordinates, but at the same angle, are correlated over shorter radial-length scales than velocities measured by observers at the same radial coordinate, but different angles. This is understood to be due to the astro-dynamical influences of the Galactic features included by Chakrabarty (2007) in the simulation that generates the training data that we use here. This simulation incorporates the joint dynamical effect of the Galactic spiral arms and the elongated Galactic bar (made of stars) that rotate at different frequencies (as per the astronomical model responsible for the generation of our training data), pivoted at the centre of the Galaxy. An effect of this joint handiwork of the bar and the spiral arms is to generate distinctive stellar velocity distributions at different radial (i.e. along the S_1 direction) coordinates, at the same angle (s_2). On the other hand, the stellar velocity distributions are more similar at different S_2 values, at the same s_1. This pattern is borne by the work by Chakrabarty (2004), in which the radial and angular variation of the standard deviations of these bivariate velocity distributions are plotted. Then it is understandable why correlation length scales are shorter along the S_1 direction, than along S_2.

Furthermore, for the correlation parameter ρ, physics suggests that the correlation will be zero among the two components of a velocity vector. These two components are after all, components of the velocity vector in a 2-dimensional orthogonal basis. However, our results indicate a small (negative) correlation between the two components of the stellar velocity vector.

9.3 Predicting $\boldsymbol{s}^{(test)}$

Figure 1, displays histogram-representations of marginal posterior probability densities of the solar location coordinates $s_1^{(test)}$, $s_2^{(test)}$; q_1^* and q_2^* that get updated once the test data is added to augment the training data, and parameters σ_{11}^{1*}, σ_{22}^{1*} and ρ^*. 95% HPD credible regions computed on each parameter in this inference scheme, are displayed in Table 1 in Appendix A. These figures display these parameters in the $nonnested-GP$ model. When the $nested-GP$ model is used, histogram-representations of the marginals of the aforementioned parameters, are displayed in Figure 2.

Prediction of $\boldsymbol{s}^{(test)}$ using the $nested-GP$ models gives rise to similar results as when the $nonnested-GP$ models are used, (see Figure 2 that compares the marginals of the solar location parameters sampled from the joint of all unknowns, given all data, in $nested-GP$ models, against those obtained when $nonnested-GP$ models are used).

The marginal distributions of $s_1^{(test)}$ indicates that the marginal is unimodal and converges well, with modes at about 2 in model units. Distribution of $s_2^{(test)}$ on the other hand is quite strongly skewed towards values of $s_2^{(test)} \lesssim 1$ radians, i.e. $s_2^{(test)} \lesssim 57$ degrees, though the probability mass in this marginal density falls sharply after about 0.4 radians, i.e. about 23 degrees. These values tally quite well with previous work (Chakrabarty et al., 2015). In that work, using the training data that we use in this work, (constructed using the the astronomical model $sp3bar3\_18$ discussed by Chakrabarty et al. (2015)), the marginal distribution of $s_1^{(test)}$ was learnt to be bimodal, with modes at about 1.85 and 2, in model units. The distribution of $s_2^{(test)}$ found by Chakrabarty et al. (2015) is however more constricted, with a sharp mode at about 0.32 radians (i.e. about 20 degrees). We do notice a mode at about this value in our inference, but unlike in the results of Chakrabarty et al. (2015), we do not find the probability mass declining to low values beyond about 15 degrees. One possible reason for this lack of compatibility could be that in Chakrabarty et al. (2015), the matrix $\boldsymbol{V}$ of velocities was vectorised, rendering the data a matrix, rather than the 3-tensor as we know it to be. Such vectorisation triggers loss of correlation information, possibly explaining their results. Model checking of our models and results is undertaken in Appendix B.

9.4 Astronomical implications

The radial coordinate of the observer in the Milky Way, i.e. the solar radial location, is dealt with in model units, but will need to be scaled to real galactic unit of distance, which is kilo parsec (kpc). Now, from independent astronomical work, the radial location of the Sun is set as 8 kpc. Then our learnt value of $S_1^{(test)}$ is to be scaled to 8 kpc, which gives 1 model unit of length to be $m := \left(\dfrac{8\text{kpc}}{\text{learnt value of } S_1^{(test)}}\right)$. Our main interest in learning the solar location, is to learn frequency Ω_{bar} with which the Galactic bar is rotating, pivoted at the galactic centre. Here $\Omega_{bar} = \dfrac{v_0}{1 \text{ model unit of length}} = \dfrac{v_0}{m}$, where $v_0 = 220$ km/s (see Chakrabarty (2007) for details). Here, S_2 is the angular distance between the Sun-Galactic centre line, and long axis of the bar, and informs on the angular location of the Galactic bar (see Table 1).

Table 1: 95% HPD on each Galactic feature parameter learnt from the solar location coordinates learnt using the two predictive inference schemes listed above and as reported in a past paper for the same training and test data.

	95% HPD for Ω_{bar} (km/s/kpc)	for angular distance of bar to Sun (degrees)
from posterior predictive	$[48.11, 57.73]$	$[4.53, 43.62]$
from joint posterior	$[48.25, 57.244]$	$[2.25, 46.80]$
from Chakrabarty et. al (2015)	$[46.75, 62.98]$	$[17.60, 79.90]$

Table 1 displays the Galactic feature parameters derived from the learnt solar location parameters, under the $nonnested - GP$ model, using sampling from the joint posterior probability of all parameters given all data, and from the posterior predictive of the solar location coordinates given test data and GP parameters already learnt from training data alone. Derived Galactic feature parameters are: bar rotational frequency Ω_{bar} in the real astronomical units (km/s/kpc), and angular distance between the bar and the Sun, in degrees. The table includes results from Chakrabarty et al. (2015).

10. Conclusions

We discuss a supervised learning method for learning tensor-valued functional relations between a sytem parameter vector, and a tensor-valued observable, multiple measurements of which build up a hypercuboidally-shaped data, that is in general not continuous, thus demanding a non-stationary covariance structure of the invoked tensor-variate GP that this sought high-dimensional function is modelled with. We prove the need for generalising a stationary, kernel-parametrised covariance function of this high-dimensional GP, into one, in which each of the hyperparameters of this covariance kernel is treated as dependent on the sample function of the invoked tensor-variate GP. This model of each sample-path dependence of each kernel hyperparameter is rephrased to model each kernel hyperparameter as a random function of the time-step at which a sample-path of the tensor-variate GP is generated. Each such an unknown scalar-valued random function is treated as a realisation from a distinct scalar-variate GP, that we learn. There are as many such scalar-variate GPs involved, as there are hyperparameters of the covariance kernel of the tensor-variate GP. We prove stationarity of each such scalar-variate GP. Thus our learning strategy comprises two layers, namely an outer layer made of a non-stationary (in general) tensor-variate GP, that lies compounded with multiple stationary scalar-variate GPs that build the inner layer. We prove sufficiency of this dual-layered learning, even when the data n the tensor-valued observable is discontinuously distributed. In our learning, we undertake Metropolis-within-Gibbs-based inference, that allows comprehensive and objective uncertainties on all learnt unknowns. Ultimately, we make an inverse Bayesian prediction of system parameter values at which test data on the observable is realised. The Generalised Wishart nature of the generative process underlining temporally-evolving covariance matrices is proved. While in this work we focussed on the learning given *discontinuities in the data*, the inclusion of non-stationarity in the covariance is a generic cure for data that bears discontinuities; applications to temporally varying, datasets are posible using such a learning strategy.

Appendix A: Results

Figure 5 display traces of the sought parameters learnt using the *nonnested*$-$*GP* model. In Figure 6, histogram representation of marginal posterior probability density of the sought parameters, given training data, are obtained using the $nonnested - GP$ model. These results are compared to the corresponding result obtained from the $nested - GP$ model. In this $nested - GP$ model, the covariance matrix $\boldsymbol{\Sigma}_3$ (that bears information about the covariance structure between sheets of data generated at different values of the input variable $\boldsymbol{S} = (S_1, S_2)^T$), is parameterised using a kernel, each length-scale hyperparameter of which, is itself modelled as a dynamically-varying function that is considered sampled from a GP. For each such scalar-variate GP that generates the length-scale ℓ_c, $c = 1, \ldots, d = 2$ the covariance matrix is itself kernel-parametrised using a stationary kernel, with an amplitude parameter value a_c and length-scale parameter δ_c.

95% HPD credible regions computed on each learnt parameter given the $nonnested{-}GP$ model, are displayed in Table 2. Again, a similar set of results from the chains run with the $nested - GP$ models are displayed in Table 3. The results on prediction of $\boldsymbol{s}^{(test)}$ are also presented in Table 2 and Table 3.

Table 2: 95% HPD credible regions on each learnt parameter, from the $nonnested - GP$ model

Parameters	using only training data	sampling from posterior predictive	sampling from joint
q_1	[3492.1,4198.1]		[3573.2,4220.8]
q_2	[68.92,76.88]		[68.37,77.33]
$\sigma_{11}^{(1)}$	[0.9837,1.0380]		[0.9797,1.0338]
ρ	[-0.0653,-0.0275]		[-0.0798,-0.0261]
$\sigma_{22}^{(1)}$	[0.3747,0.4234]		[0.3703,0.4237]
s_1	-	[1.8212,2.1532]	[1.8038,2.1960]
s_2	-	[0.0421,1.2052]	[0.0157,1.2172]

Appendix B: Model Checking

One way to check for the model and results, given the data at hand, is to generate data from the learnt model, and then compare this generated data with the observed data. Now, the model that we learn, is essentially the tensor-variate GP that is used to model the functional relationship $\boldsymbol{\xi}(\cdot)$ between the observable $\boldsymbol{V}$ and the input-space parameter $\boldsymbol{S}$. By, saying that we want to generate new data, we imply the prediction of a new value of $\boldsymbol{V}$, given the learnt model of this GP.

This prediction of new datum on $\boldsymbol{V}$, is fundamentally different from the inverse prediction of the value $\boldsymbol{s}^{(test)}$ of the input-space parameter $\boldsymbol{S}$ that we have undertaken – as discussed above – where the sought $\boldsymbol{s}^{(test)}$ is the value of $\boldsymbol{S}$ at which test data $\boldsymbol{v}^{(test)}$ on $\boldsymbol{V}$ is recorded. There is no closed-form solution to the posterior predictive of $\boldsymbol{s}^{(test)}$ given the test data and the learnt GP parameters.

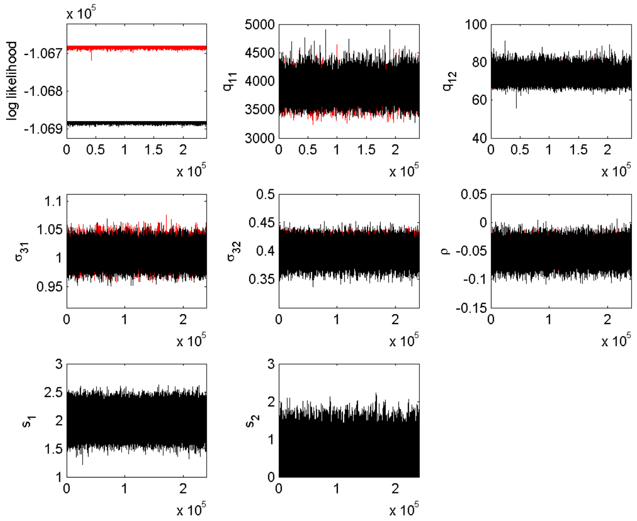

Figure 5: Results from run done with training data $\mathbf{D}$ with the $nonnested-GP$ model, are shown in grey (or red in the electronic copy of the paper) while results from run undertaken with training and test data, $\boldsymbol{D}^\star$, in this $nonnested-GP$ model, are depicted in black. Traces of logarithm of the likelihood are displayed from the two runs in the top left panel. Reciprocal of the length scale parameters are shown in the top middle and right panels; here $q_c = \ell_c^{-1}$, $c = 1, 2$. Traces of the learnt diagonal elements $\sigma_{11}^{(1)}$ and $\sigma_{22}^{(1)}$, of the covariance matrix $\boldsymbol{\Sigma}_1$, are shown in the mid-row, left and middle panels. Trace of the correlation $\rho = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}^{(1)}\sigma_{22}^{(1)}}}$ is displayed in the mid-row right panel. Prediction of the values of the input parameter $\boldsymbol{S} = (S_1, S_2)^T$ is possible only in the run performed with both training and test data. Traces of S_1 and S_2 values learnt via MCMC-based sampling from the joint of all unknown parameters given $\boldsymbol{D}^\star$, are shown in the lower panel.

.

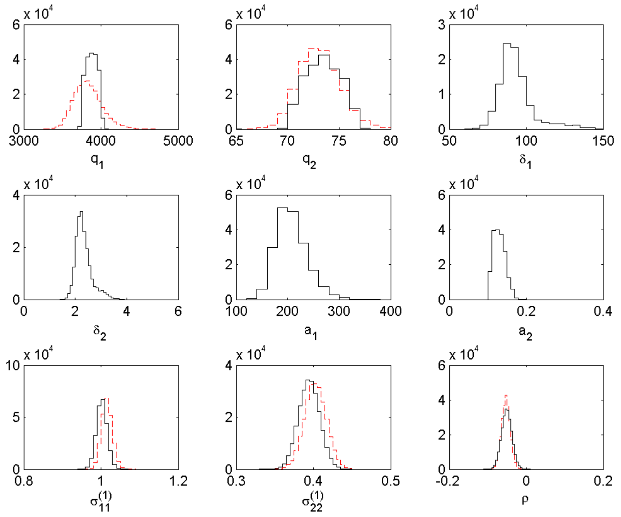

Figure 6: Marginal posterior probability densities of unknown parameters, given training data $\mathbf{D}$, are depicted as histograms. Histograms obtained from the run done with only within the $nested - GP$ model, shown in black, as distinguished from the results of learning given the same data, and the $nonnested - GP$ model depicted in grey (or red in the electronic copy of the thesis). Given the data used here, $s_1^{(test)}$ and $s_2^{(test)}$, are not learnt.

.

Table 3: 95% HPD credible regions on each learnt parameter, from the $nested - GP$ model

Parameters	$t_0 = 200$	$t_0 = 100$	$t_0 = 50$
q_1	[3740.96, 3917.32]	[3710.4, 4011.66]	[3650.92, 4033.51]
q_2	[70.34, 75.70]	[70.42, 76.43]	[68.94, 76.22]
a_1	[78.67, 124.02]	[43.82, 167.35]	[48.27, 219.37]
a_2	[1.88, 3.03]	[2.12, 3.57]	[1.64, 6.16]
d_1	[155.64, 301.65]	[78.47, 521.67]	[123.42, 828.37]
d_2	[0.10, 0.15]	[0.12, 0.46]	[0.10, 0.52]
σ_{31}	[0.97, 1.02]	[0.97, 1.03]	[0.98, 1.02]
σ_{32}	[0.37, 0.41]	[0.37, 0.41]	[0.38, 0.41]
ρ	[-0.076, -0.031]	[-0.073, -0.03]	[-0.075, -0.032]
s_1	[1.83, 2.16]	[1.77, 2.22]	[1.76, 2.24]
s_2	[0.138, 1.15]	[0.112, 1.16]	[0.071, 1.15]

Here we discuss our undertaking of the checking of the model and the results that were presented in the last section, in the context of the Galactic application discussed above in Section 7 of main paper, using the training data used in this application. At chosen values of $\boldsymbol{S}$ – chosen to be the design points in this training data, for convenience – the kernel-parametrised covariance function $\boldsymbol{\Sigma}_3$ of the 3rd-order Tensor Normal likelihood GP, is known, given the learnt values of the parameters of the kernel used to parametrise $\boldsymbol{\Sigma}_3$. However, in our Bayesian inference, we learn the marginal posterior of each unknown parameter, given the data. Thus, in order to pin the value of each element of $\boldsymbol{\Sigma}_3$, we identify the parameter value corresponding to a selected summary of this posterior distribution. For example, we could choose to define $\boldsymbol{\Sigma}_3$ at pairs of known design points $\boldsymbol{s}_i$, $\boldsymbol{s}_j$, and the modal value of ℓ_c – identified from the marginal posterior of ℓ_c inferred upon, given the data. Here $i, j \in \{1, \ldots, n = 216\}$. The resulting value of the ij-th element of $\boldsymbol{\Sigma}_3$ will then provide one summary, of the covariance between the 50×2 stellar velocity matrix $\boldsymbol{v}_i$ realised at $\boldsymbol{S} = \boldsymbol{s}_i$, and $\boldsymbol{v}_j$ realised at $\boldsymbol{S} = \boldsymbol{s}_j$. Similarly, the learnt modal values of the parameters $\sigma_{11}^{(1)}$, $\sigma_{22}^{(1)}$ and ρ define one summary of the covariance matrix $\boldsymbol{\Sigma}_1$ that informs on the covariance between the 2 216×50-dimensional sheets of data on each component of the 2-dimensional stellar velocity vector. Again, other summaries of the parameter values could be used as well, for example, the parameter value identified at the mean of the marginal posterior density of this parameter, as learnt given the training data, is also used.

In this model checking exercise, the unknowns are certain elements of the cuboidally-shaped data comprising the 216 number of 50×2-dimensional stellar velocity matrices generated by astronomical simulation, at chosen design points $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{216}$, i.e. the 3rd-order tensor $\mathbf{D}_V := \{\boldsymbol{v}_1, \vdots \boldsymbol{v}_2, \vdots \ldots, \vdots \boldsymbol{v}_{216}\}$. In the first attempt to model checking, we generate all elements of the q-th such simulated stellar velocity matrix $\boldsymbol{v}_q$, (that is generated at the known design point $\boldsymbol{s}_q$), i.e. generate values of $50 \times 2 = 100$ unknown elements of matrix $\boldsymbol{v}_q$. We refer to these unknown elements of $\boldsymbol{v}_q$ as $v_{11}^{(q)}, v_{12}^{(q)}, v_{21}^{(q)}, \ldots, v_{50,2}^{(q)}$. The 3rd-ordered tensor without the q-th slice, is referred to as $\mathbf{D}_V^{(-q)} := \{\boldsymbol{v}_1, \vdots \boldsymbol{v}_2, \vdots \ldots, \vdots \boldsymbol{v}_{q-1}, \vdots \boldsymbol{v}_{q+1}, \vdots \boldsymbol{v}_{216}\}$. The joint posterior probability density of the 100 unknowns, at the learnt modal values $q_1^{(mode)}, q_2^{(mode)}, \sigma_{11}^{(1,mode)}, \sigma_{22}^{(1,mode)}, \rho^{(mode)}$ is

$$\pi\left(v_{11}^{(q)}, v_{12}^{(q)}, v_{21}^{(q)}, \ldots, v_{50,2}^{(q)} | \mathbf{D}_V^{(-q)}\right) \propto \mathcal{TN}_{2\times 50\times 216}(\hat{\boldsymbol{M}}, \boldsymbol{\Sigma}_1^{(mode)}, \hat{\boldsymbol{\Sigma}}_2, \boldsymbol{\Sigma}_3^{(mode)}),$$

where,
–the 3rd-ordered tensor-valued data that enters the parametric form of the 3rd-ordered tensor-normal density on the RHS, has elements of its q-th slice, (out of a total of 216 slices), unknown. All other elements of this $2 \times 50 \times 216$-dimensional tensor are known;
–uniform priors are used on the unknowns; –$\boldsymbol{\Sigma}_1^{(mode)}$ is the learnt modal value of the 2×2-dimensional covariance matrix $\boldsymbol{\Sigma}_1$ s.t. its $1,1$-th element is $\sigma_{11}^{(1,mode)}$, $2,2$-th element is $\sigma_{22}^{(1,mode)}$, $1,2$-th element is $\rho^{(mode)}\sqrt{\sigma_{22}^{(1,mode)}\sigma_{11}^{(1,mode)}}$, and the $2,1$-th element is equal to the $1,2$-th element (as this is a covariance matrix);
–$\boldsymbol{\Sigma}_3^{(mode)}$ is the learnt modal value of the 216×216-dimensional covariance matrix $\boldsymbol{\Sigma}_3$, s.t. its ij-th element is $\exp\left[-(\boldsymbol{s}_i - \boldsymbol{s}_j)^T \boldsymbol{Q}^{(mode)}(\boldsymbol{s}_i - \boldsymbol{s}_j)\right]$, with the non-zero elements of the

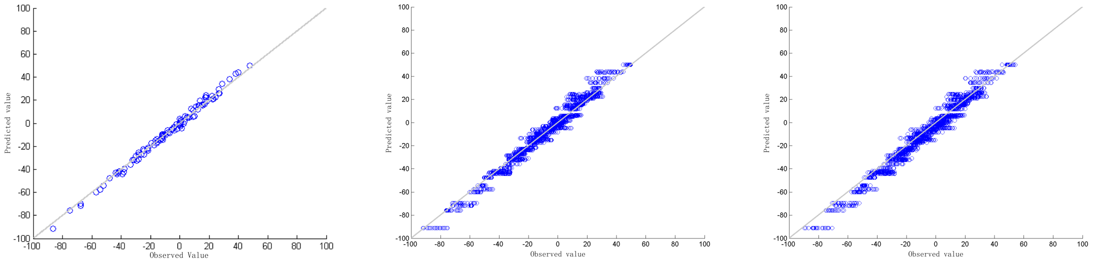

Figure 7: *Left:* Comparison of the observed and predicted values of elements of the q-th 50×2-dimensional stellar velocity matrix $\boldsymbol{v}_q$, where 216 such matrices constitute the training data $\mathbf{D}_V$ (on velocities of 50 stellar neighbours of the Sun) that is generated by astronomical simulations. The predicted or learnt values are obtained from a RW-MCMC chain undertaken with the all elements of the 3rd-order tensor $\mathbf{D}_V$ known, except for the elements of its q-th slice, and the learnt values of the parameters of the GP used to model the data at hand, at a chosen summary, namely the mode, of the marginal posterior density of each such learnt GP parameter. Here q=200. Equality of the observed and predicted values of he elements of $\boldsymbol{v}_q$ is indicated by the point lying on the drawn straight line with unit slope; the predicted values are found to lie close to this line. *Middle:* Depicts a similar comparison, as displayed in the left panel, but for 20 distinct values of q, namely for $q = 190, 191, \ldots, 210$. *Right:* Depicts the same comparison of observed and predicted values of elements of 20 slices $\boldsymbol{v}_{190}, \ldots, \boldsymbol{v}_{210}$, but this time, the employed GP parameters are the means of their respective marginals. Thus, this model-checking exercise checks for the used models and results obtained (given the data at hand) at the mean of the respective posterior.

.

diagonal 2×2-dimensional $\boldsymbol{Q}^{(mode)}$-matrix given by $q_1^{(mode)}$ and $q_2^{(mode)}$. $\boldsymbol{s}_i$ being the i-th design point, is known $\forall i, j = 1, \ldots, 216$.

To learn the 100 unknowns $v_{11}^{(q)}, v_{12}^{(q)}, v_{21}^{(q)}, \ldots, v_{50,2}^{(q)}$, we run a RW Metropolis-Hastings chain, with the data defined as above, the known 216 number of design points, and all the learnt, modal parameter values. The joint posterior of the unknowns that defines the acceptance ratio in this chain, is given as in the last equation. The chain is run for 20,000 iterations, for q=200, and the mean of the last 1000 samples of $v_{ij}^{(200)}$ is recorded, where $i = 1, \ldots, 50$, $j = 1, 2$. These sample means $\bar{v}_{ij}^{(200)}$ then constitute the learnt value of the 100 elements of the 200-th stellar velocity matrix $\boldsymbol{v}_{200}$. We plot the pairs of learnt value $\bar{v}_{ij}^{(200)}$ of elements of the $\boldsymbol{v}_{200}$ matrix, against the empirically observed value of this element, $\forall i = 1, \ldots, 50$, $\forall j = 1, 2$. The plot is presented in the left panel of Figure 7. Thus, each point on this plot is a pair (empirically observed value of$v_{ij}^{(200)}$, $\bar{v}_{ij}^{(200)}$), and there are $50 \times 2 = 100$ points in this plot. The points are found to lie around the straight line with slope 1. In other words, the values of the elements in the q-th (=200-th) slice of the training data that we learn using our model, are approximately equal to the empirically observed values of these elements. This is corroboration of our models and results.

We attempt a similar prediction of elements of the training data for other values of q, namely for $q = 190, \ldots, 210$. The learnt values of elements of $\boldsymbol{v}_q$, for each q, is plotted against the empirically observed elements of $\boldsymbol{v}_q$. We have superimposed results for all 20 values of q in the same plot, resulting in the middle panel of Figure 7. Again, the values predicted for all 20 slices, are found to be close to the empirical observations, as betrayed by the points lying close to the straight line of unit slope.

Lastly, we wanted to ensure that the encouraging results from our model checking exercise is robust to changes in the posterior summary of the learnt GP parameters. Thus, we switch to using the mean of the parameter marginal posterior from the posterior mode, and carry out the same exercise of predicting elements of slices $\boldsymbol{v}_{190}, \ldots, \boldsymbol{v}_{210}$. Results are displayed in the right panel of Figure 7. Again, very encouraging corroboration of our used models and results (of learning the GP parameters) is noted. Indeed, in such model checking exercises, encouraging match between the predictions and the empirical observations lends confidence in the used models and results obtained therefrom, given the data at hand–such models and results are the inputs to this exercise. However, if lack of compatibility is noted in such a model checking exercise, between empirical observations and predictions, then it implies that either the used modelling is wrong, and/or the results obtained therefrom given the data are wrong. However, the model checking exercise that we undertake, vindicates our models and results, given the data at hand.

References

J. A. D. Aston and Claudia Kirch. Evaluating stationarity via change-point alternatives with applications to fmri data. *Annals of Applied Statistics*, 6(4):1906–1948, 2012.

Timothy A Barton and Daniel R Fuhrmann. Covariance structures for multidimensional data. *Multidimensional Systems and Signal Processing*, 4(2):111–123, 1993.

Fetsje Bijma, Jan C De Munck, and Rob M Heethaar. The spatiotemporal meg covariance matrix modeled as a sum of kronecker products. *NeuroImage*, 27(2):402–415, 2005.

D. Chakrabarty. Patterns in the outer parts of galactic disks. *Monthly Notices of the Royal Astronomical Society*, 352:427, 2004.

Dalia Chakrabarty. Phase space structure in the solar neighbourhood. *Astronomy & Astrophysics*, 467(1):145–162, 2007.

Dalia Chakrabarty, Munmun Biswas, Sourabh Bhattacharya, et al. Bayesian nonparametric estimation of milky way parameters using matrix-variate data, in a new gaussian process based method. *Electronic Journal of Statistics*, 9(1):1378–1403, 2015.

Raj Chari, Bradley P Coe, Emily A Vucic, William W Lockwood, and Wan L Lam. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC systems biology*, 4(1):67, 2010a.

Raj Chari, Kelsie L Thu, Ian M Wilson, William W Lockwood, Kim M Lonergan, Bradley P Coe, Chad A Malloff, Adi F Gazdar, Stephen Lam, Cathie Garnis, et al. Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. *Cancer and Metastasis Reviews*, 29(1):73–93, 2010b. doi: 10.1007/s10555-010-9199-2.

Robert Clarke, Habtom W Ressom, Antai Wang, Jianhua Xuan, Minetta C Liu, Edmund A Gehan, and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37, 2008.

Ian L Dryden, Li Bai, Christopher J Brignell, and Linlin Shen. Factored principal components analysis, with applications to face recognition. *Statistics and Computing*, 19(3): 229–238, 2009.

Piers K Dunstan, Scott D Foster, Francis KC Hui, and David I Warton. Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of agricultural, biological, and environmental statistics*, 18(3):357–375, 2013.

M. L. Eaton. Chapter 8: The wishart distribution. In *Multi-variate Statistics: A Vector Space Approach*, pages 302–333. OH: Institute of Mathematical Statistics, 1990.

Yimei Fan. *Statistical Learning with Applications in High Dimensional Data in Health Care Analytics*. PhD thesis, University of Maryland, 2017.

Xiping Fu. *Exploring geometrical structures in high-dimensional computer vision data*. PhD thesis, University of Otago, 2016.

R. Gramacy. *Bayesian treed Gaussian process models*. PhD thesis, University of California, SC, 2005.

P. Hajlasz. Geometric analysis. *http://www.pitt.edu/~hajlasz/Notatki/Analysis4.pdf linked from http://www.pitt.edu/~hajlasz/Teaching/Math2304Spring2014/m2304Spring2014.htm* 2014.

M. Heinonen, H. Mannerstrm, J. Rousu, S. Kaski, and H. Lhdesmki. Non-stationary gaussian process regression with hamiltonian monte carlo. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 732–740, Cadiz, Spain, 2016. PMLR.

Karla Hoff. Bayesian learning in an infant industry model. *Journal of International Economics*, 43(3-4):409–436, 1997.

Peter D Hoff et al. Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196, 2011.

B. Kerkheim. Rectifiable metric spaces: local structure and regularity of the hausdorff measure. *Proceedings of American Mathematical Society*, 121:113–124, 1994.

D. Knuth. *The Art of Computer Programming: Seminumerical Algorithms*. Addison-Wesley Longman Publishing Co., Boston, MA, USA, 1997.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. doi: 10.1137/07070111X.

Pedro J. Leitao, Marcel Schwieder, Stefan Suess, Ins Catry, Edward J. Milton, Francisco Moreira, Patrick E. Osborne, Manuel J. Pinto, Sebastian van der Linden, and Patrick Hostert. Mapping beta diversity from space: Sparse generalised dissimilarity modelling (sgdm) for analysing high-dimensional data. *Methods in Ecology and Evolution*, 6(7): 764–771, 2015. ISSN 2041-210X. doi: 10.1111/2041-210X.12378.

Ameur M. Manceur and Pierre Dutilleul. Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37 – 49, 2013. ISSN 0377-0427. doi: 10.1016/j.cam.2012.09.017.

Kanti V Mardia and Colin R Goodall. *Spatial-temporal analysis of multivariate environmental monitoring data*, volume 6. Elsevier New York, 1993.

P. McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, 1987.

Ann L Oberg, Brett A McKinney, Daniel J Schaid, V Shane Pankratz, Richard B Kennedy, and Gregory A Poland. Lessons learned in the analysis of high-dimensional data in vaccinomics. *Vaccine*, 33(40):5262–5270, 2015. doi: 10.1016/j.vaccine.2015.04.088.

Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33 (5):2295–2317, 2011.

C. Paciorek and M. J. Schervish. Nonstationary covariance functions for gaussian process regression. In *In NIPS*, pages 273–280, 2004.

Ying Han Pang, Ean Yee Khor, and Shih Yin Ooi. Biometric access control with high dimensional facial features. In *Australasian Conference on Information Security and Privacy*, pages 437–445. Springer, 2016.

Q. Qiang and Z. Fei. Generation of facial gesture and expression in high-dimensional space. In *2011 International Conference on Internet Technology and Applications*, pages 1–5, Aug 2011. doi: 10.1109/ITAP.2011.6006383.

Andreas Richter, Jussi Salmi, and Visa Koivunen. Ml estimation of covariance matrix for tensor valued signals in noise. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2349–2352. IEEE, 2008.

P. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87:108–119, 1992.

Chandrima Sarkar. *Improving Predictive Modeling in High Dimensional, Heterogeneous and Sparse Health Care Data*. PhD thesis, University of Minnesota, 2015.

A. Schmidt and A. OHagan. Bayesian inference for non-stationary spatial covariance structures via spatial deformations. *Journal of the Royal Statistical Society Series B*, 65: 743–758, 2003.

A. Shirley. The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1):178–196, 1965.

J. Snoek, K. Swersky, R. Zemel, and R. Adams. Input warping for bayesian optimization of non-stationary functions. In *In ICML*, pages 1674–1682, 2014.

Douglas L Theobald and Deborah S Wuttke. Accurate structural correlations from maximum likelihood superpositions. *PLoS computational biology*, 4(2):e43, 2008.

V. Tolvanen, P. Jylnki, and A. Vehtari. Expectation propagation for nonstationary heteroscedastic gaussian process regression. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2014.

Jianzhong Wang. *Geometric structure of high-dimensional data and dimensionality reduction*. Springer, 2011.

Wei Wang, Lei Chen, and Qian Zhang. Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation. *Computer Networks*, 88:136–148, 2015.

David I. Warton. Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics*, 67(1):116–123, 2011. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2010.01438.x.

Karl Werner, Magnus Jansson, and Petre Stoica. On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491, 2008.

Andrew Wilson and Zoubin Ghahramani. Generalised wishart processes. In *In Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 736–744. Corvallis, OregonAUAI Press, 2011.