arXiv: math.PR/0000000

Bayesian Nonparametric Estimation of Milky Way Parameters Using Matrix-Variate Data, in a New Gaussian Process Based Method

Dalia Chakrabarty[*,§]**,, Munmun Biswas**[†,¶]**, Sourabh Bhattacharya**[‡,¶] **,**

[§] *Department of Statistics*
University of Warwick
Coventry CV4 7AL, U.K.
d.chakrabarty@warwick.ac.uk
and
Department of Mathematics
University of Leicester
Leicester LE1 7RH, U.K.
dc252@le.ac.uk

[¶] *Indian Statistical Institute*
203, B. T. Road
Kolkata 700108, India
munmun.biswas08@gmail.com sourabh@isical.ac.in

Abstract: In this paper we develop an inverse Bayesian approach to find the value of the unknown model parameter vector that supports the real (or test) data, where the data comprises measurements of a matrix-variate variable. The method is illustrated via the estimation of the unknown Milky Way feature parameter vector, using available test and simulated (training) stellar velocity data matrices. The data is represented as an unknown function of the model parameters, where this high-dimensional function is modelled using a high-dimensional Gaussian Process ($\mathcal{GP}$). The model for this function is trained using available training data and inverted by Bayesian means, to estimate the sought value of the model parameter vector at which the test data is realised. We achieve a closed-form expression for the posterior of the unknown parameter vector and the parameters of the invoked $\mathcal{GP}$, given test and training data. We perform model fitting by comparing the observed data with predictions made at different summaries of the posterior probability of the model parameter vector. As a supplement, we undertake a leave-one-out cross validation of our method.

Keywords and phrases: Supervised learning, Inverse problems, Gaussian Process, Matrix-variate Normal, Transformation-based MCMC.

1. Introduction

Curiosity about the nature of the parameter space of the Milky Way that we earthlings live in, is only natural. In this paper, we discuss the learning of the parameters characterising those Milky Way features that bear influence upon the motion of individual stars that lie in the neighbourhood of the Sun. Astrophysical modelling indicates that in the solar neighbourhood, effects of different features

[*] Associate Research fellow at Department of Statistics, University of Warwick and Lecturer of Statistics at Department of Mathematics, University of Leicester

[†] PhD student in Statistics and Mathematics Unit, Indian Statistical Institute

[‡] Assistant Professor in Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute

of the Milky Way are relevant (Minchev et al., 2009; Chakrabarty, 2007; Antoja et al., 2009). Such features include an elongated bar-like structure made of stars (the stellar bar) that rotates, pivoted at the centre of the Galaxy. In addition, the spiral arms of the Galaxy are also relevant. Thus, the motions of stars in the solar neighbourhood are affected by the parameters that define these Galactic features. Included in these feature parameters are the locations of the observer of such motions–we from Earth observe such motions, so that the stellar velocities are recorded to attain the observed values, given where in the Galaxy we are measuring these velocities from. On astronomical scales, the Earth's location in the Milky Way is equivalent to the location of the Sun inside the Galaxy. Our location in the two-dimensional (by assumption) Galactic disk, is given by the angular separation of the Sun from a chosen line (an identified axis of the aforementioned stellar bar) and the distance from the Sun to the centre of the Galaxy. These two location parameters are the components of the two-dimensional location $\boldsymbol{S}$ of the observer. As motivated above, parameters of the bar, spiral pattern and other Milky Way features, can also affect the motions of stars that are observed. (See section **S-1** of the supplementary material for details). Given that these galactic feature parameters affect the solar neighbourhood, if motions of a sample of stars in this neighbourhood are measured, such data will harbour information about these feature parameters. Then, the inversion of such measured motions will in principle, allow for the learning of the unknown feature parameters. This approach has been adopted in the modelling of our galaxy, to result in the estimation of the angular separation of the Sun from a chosen axis of the bar, and the distance of the Sun from the Galactic centre (Minchev et al., 2010; Fux, 2001; Dehnen, 2000; Simone et al., 2004). The other relevant feature parameters are typically held constant in such modelling.

The above inverse problem is then an example application of the method of science that is typified by attempts at learning the unknown model parameter vector given observed data, where the causal relationship between the observable and the model parameter vector $\boldsymbol{S}$, is not necessarily known. This unknown relationship or function, can itself be learnt using available "training data". Once this function is learnt, it can in principle be inverted to predict the unknown value of $\boldsymbol{S}$ at which the measured data–i.e. "test data"–is realised. Such test data is contrasted with "training data", which is data generated at known or chosen values of $\boldsymbol{S}$ (for example, via simulations or obtained as archival data).

The learning of a high-dimensional function from available training data, using standard non-parametric methods (such as spline fitting or wavelet based learning) is expected to be unsatisfactory since modelling high-dimensional functions using splines/wavelets may fail to adequately take into account the correlation structure between the component functions. Also, the complexity of the computational task of learning the unknown function from the data–and in particular of inverting it–only increases with dimensionality. Furthermore, the additional worry in the classical approach is that parameter uncertainty is ignored, though the same can be addressed in a Bayesian framework. An added advantage of the Bayesian approach is that priors on the unknown parameters can bring in extra information into the model, allowing for a training data set of comparatively smaller size (than that required in the classical approach), to be adequate.

Solving for the value of $\boldsymbol{S}$ that supports the real or test data requires operating the inverse of the learnt function on the test data. The existence and uniqueness of such solution can be questioned given that the problem may not even be well-posed in a Hadamard sense (Kabanikhin, 2008; B.Hofmann, 2011; Tarantola, 2005). The problem may even be ill-conditioned since errors in the measurement may exist. Such worries about ill-posedness and ill-conditioning are mitigated in the Bayesian framework (Carreira-Perpin, 2001; Stuart, 2013). In this approach, the solution entails computation of the posterior probability of the unknown $\boldsymbol{S}$ (at which the test data is realised), given all data. Given the inherent inadequacies of learning using splines/wavelets discussed above, we opt to model the unknown functional relationship between data and model parameter $\boldsymbol{S}$ with a high-dimensional $\mathcal{GP}$. Similarly, in our application of interest, the unknown functional relation between the high-dimensional observations on stellar motions and the unknown observer location vector $\boldsymbol{S}$ is modelled as a high-dimensional $\mathcal{GP}$. In this exercise, Galactic feature parameters other than the observer location are maintained as constants.

Chakrabarty (2007) constructed four different base-astronomical models of the solar neighbourhood, each at a chosen value of the ratio of the rate of rotation of the spiral pattern (Ω_s) to that of the bar (Ω_b). Non-linear dynamical evolution of each of these four base-astronomical models were carried out by Chakrabarty (2007), resulting in four independent data sets, each consisting of n blocks of j number of k-dimensional stellar velocity vectors, where each block is generated at a chosen value of $\boldsymbol{S}$ (aka, a "design point"). At each possible chosen location $\boldsymbol{s}$ of the Sun, the dynamical evolution of a given base-astronomical model of the Galaxy generates a block representing the k-dimensional velocity of each of j stars, where these stars are chosen as neighbours of the Sun. Thus, there are n design points and each training data set consists of n number of $j \times k$-matrices, with a matrix generated at the corresponding design point. There are four such training data sets generated, by performing the evolution of each base-astronomical model. In addition, there is a measured, stellar velocity data matrix–of dimensionality $j \times k$ again–available, but this time, we do not know what is the value of $\boldsymbol{S}$ at which this measured/test data has been realised. It is this unknown value of $\boldsymbol{S}$ that we seek to Bayesianly learn, given the test data and one training data set at a time.

It maybe asked that if a stellar velocity matrix can be generated at a chosen $\boldsymbol{s}$, via the evolution of a base-astronomical model, does this not amount to stating that the causal relationship between the observable (velocity matrix) and model parameter ($\boldsymbol{S}$) is already known? Indeed this knowledge must be embedded within the evolutionary scheme implemented on any base-astronomical model. Thus, the forward evolution of a base-astronomical model is possible (via Newton's equations of motion), in order to generate a velocity matrix at a chosen $\boldsymbol{s}$. However the inversion of this evolution–aimed at recovering the sought $\boldsymbol{s}$ at which the measured velocity matrix is generated–is not possible in general, owing to non-linear dynamical effects, or chaos, that impede reversibility in evolution; see Sengupta (2003), Section 6.6 of Chakrabarty (2007), Section 7 of Fux (2001). The strength of such chaos is different in the different base-astronomical models, caused by the different values of Ω_s/Ω_b, (discussed below in Section 3). This difficulty of inversion triggers the need to learn the inverse of the function that expresses the observable as a function of $\boldsymbol{S}$, independently from each of the four available training data sets. This learnt inverse function is then to be operated upon the measured (test) data to predict the value of $\boldsymbol{S}$ in the Milky Way, in each of the four cases that represent four possible astronomical models of the Milky Way. We of course, predict this value of $\boldsymbol{S}$ Bayesianly, by using a high-dimensional $\mathcal{GP}$ to model the velocity data. We then achieve a closed-form posterior probability density of the sought $\boldsymbol{s}$ and relevant parameters of this $\mathcal{GP}$, given the test and training data. Marginal posterior distribution of the components of the sought $\boldsymbol{s}$ vector are inferred using MCMC, for each base-astronomical model (i.e. each training data set) used. Our focus in this work is to make inference on all values of $\boldsymbol{S}$ at which the test data is realised, in each of the four astronomical models of the Galaxy–selection of the base-astronomical model is beyond the scope of this paper (see Section 4).

In the astronomical literature, Milky Way feature parameters in the solar neighbourhood have been explored via simulation based studies (Englmaier and Gerhard, 1999; Fux, 1997) while similar estimation is performed using other (astronomical) model-based studies (Aumer and Binney, 2009; Perryman, 2012; Golubov, 2012). Chakrabarty (2007) attempted estimation of the sought Galactic parameters via a test of hypothesis exercise: a non-parametric frequentist test was designed to test for the null that the observed stellar velocity data matrix is sampled from the estimated density of a synthetic velocity data matrix generated at the corresponding chosen value of the Milky Way feature parameter vector $\boldsymbol{S}$. The p-value of the used test statistic was recorded for each choice of $\boldsymbol{s}$. The choices of $\boldsymbol{s}$ at which the highest p-values were obtained, were considered better supported by the observed data. Hence the empirical distribution of these p-values in the space of $\boldsymbol{S}$, was used to provide interval estimates of the Milky Way feature vector. However, this method required computational effort and is highly data intensive since the best match is sought over a very large collection of training data points. This shortcoming had compelled Chakrabarty (2007) to resort to an unsatisfactory coarse gridding of the space of $\boldsymbol{S}$. This problem gets acute enough for the method to be rendered useless when the dimensionality of the vector $\boldsymbol{S}$ that we hope to learn, increases. Moreover, the method of quantification of uncertainty of the estimate of the location is also unsatisfactory, dependent crucially on the binning details, which in turn is bounded by cost and memory considerations.

In the method we develop here, we demonstrate the effectiveness of our Gaussian Process based method with much smaller data sets than were used in the past. The other major advantage of this presented method is that it readily allows for the expansion of dimensionality of the model parameter vector and is capable of taking measurement errors into account.

The rest of the paper is structured as follows. In Section 2, we present the details of the modelling strategy that we adopt. The treatment of measurement errors within the modelling is discussed in Section 2.6. In Section 3 we discuss the application via which the new method is illustrated while details of the inference are discussed in Section 3.1. Section 4 contains results obtained from using available real and training data. We compare the obtained results with the estimates available in the astronomical literature in Section 4.1. Section 5 presents results of model fitting by comparing test data with predictions made at different summaries of the posterior of the model parameter vector $\boldsymbol{S}$. The paper is rounded up with Section 6.

2. Model

In this section we discuss the generic methodology that we use to learn the unknown location vector of the observer in the Milky Way disk, given the matrix-variate test and training stellar velocity data. Once the method is motivated, we implement it in the following section, to perform the learning relevant to the application at hand.

If a matrix-variate observable is expressed as an unknown matrix-variate function of the model parameter $\boldsymbol{S}$, and this unknown causal relationship between observable and $\boldsymbol{S}$ is modelled by a matrix-variate Gaussian Process ($\mathcal{GP}$), it would imply that one realisation from such a matrix-variate $\mathcal{GP}$ would be a set of the observed matrices that will be jointly distributed as 3-tensor normal, parametrised by a mean matrix and 3 covariance matrices (Hoff, 2011). While applications of the same are being developed (Wang & Chakrabarty), here we undertake an alternative and equivalent modelling strategy. We vectorise our intrinsically matrix-variate data sets to achieve a close-form expression for the joint posterior probability of the unknown parameters that we are interested in learning from the data. This leads to the functional relationship between the data and model parameter vector being rendered vector-valued, modelled by a vector-variate $\mathcal{GP}$, a set of realisations from which is jointly matrix normal, parametrised by matrix-variate parameters that we intend to learn from the data, along with the unknown $\boldsymbol{s}$ at which the measured data is realised.

Let j number of measurements of a k-dimensional variable be available; this vector variable is referred to below as the "observable". Thus the measurements of this observable constitute a $j \times k$-dimensional matrix. We refer to the measured data as test data and seek the unknown value $\boldsymbol{s}^{(new)}$ of model parameter $\boldsymbol{S}$ at which it is realised. Let data be generated at n known values of $\boldsymbol{S}$: $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star$. Then $\{\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star\}$ is the design set and $\boldsymbol{s}_i^\star$ is the i-th design vector at which the i-th synthetic data matrix is generated, $i = 1, 2, \ldots, n$. Then these n synthetic data matrices comprise a training data set. Here a data matrix is $j \times k$-dimensional. As motivated in the introductory section, we express the relation between the observable $\mathbf{V}$ and unknown model parameter vector $\boldsymbol{S}$ as $\mathbf{V} = \boldsymbol{\xi}(\boldsymbol{S})$, where $\boldsymbol{\xi}(\cdot)$ is an unknown function. We train the model for $\boldsymbol{\xi}(\cdot)$ using the training data and invert the function using Bayesian means to estimate the unknown $\boldsymbol{s}^{(new)}$ at which the test data is realised.

As discussed above, we vectorise the intrinsically $j \times k$-dimensional matrix-variate data sets as jk-dimensional vectors. In this treatment, as a measurement is rendered vector-valued, $\boldsymbol{\xi}(\cdot)$ is vector-valued and $\boldsymbol{\xi}(\cdot)$ can be modelled by a vector-variate $\mathcal{GP}$ so that realisations from this $\mathcal{GP}$ are jointly matrix normal. Thus, we consider the j number of measurements of the k-dimensional observable, as a jk-dimensional observed vector $\mathbf{v}^{(test)}$. This test data is realised at the unknown value $\boldsymbol{s}^{(new)}$ of $\boldsymbol{S}$. Again, a $j \times k$-dimensional synthetic data matrix is treated as a jk-dimensional synthetic data vector $\mathbf{v}_i$, $i = 1, 2, \ldots, n$, along the lines of the observed data. Then all the n synthetic data vectors together comprise the training data $\mathcal{D}_s = (\mathbf{v}_1 \vdots \mathbf{v}_2 \vdots \ldots \vdots \mathbf{v}_n)^T$ where $\mathbf{v}_i$ is generated at the chosen value $\boldsymbol{s}_i^\star$ of $\boldsymbol{S}$, $i = 1, \ldots, n$. Given our treatment of $\mathbf{v}_i$ as a jk-dimensional vector, the training data set $\mathcal{D}_s$ is a matrix with n rows and jk columns.

Thus in this treatment, we have n jk-dimensional synthetic data vectors (inputs), each generated at a chosen value of the model parameter vector (target), i.e. we have the n observations $(\mathbf{v}_1, \boldsymbol{s}_1^\star), \ldots, (\mathbf{v}_n, \boldsymbol{s}_n^\star)$, and the aim is to predict the unknown model parameter vector $\boldsymbol{s}^{(new)}$ at which the input is the test data, i.e. the data vector $\mathbf{v}^{(test)}$. In this paradigm of supervised learning akin to the discussion in Neal (1998), a predictive distribution of $\boldsymbol{s}^{(new)}$ is sought, conditioned on the test data $\mathbf{v}^{(test)}$ and the training data $\mathcal{D}_s = (\mathbf{v}_1 \vdots \mathbf{v}_2, \vdots \ldots \vdots \mathbf{v}_n)^T$.

We begin the discussion on the model by elaborating on the detailed structure of the used $\mathcal{GP}$. In this section we ignore measurement errors and present our model of these n vector-variate functions. Later in Section 2.6, we delineate the method used to take measurement uncertainties on board.

As the data are vectorised as jk-dimensional vectors, $\boldsymbol{\xi}(\cdot)$ is also rendered a jk-variate vector function whose ℓ-th component function is $\xi_\ell(\cdot)$. Then we can write $\mathbf{v}_i = \boldsymbol{\xi}(\boldsymbol{s}_i) := (\xi_1(\boldsymbol{s}_i), \ldots, \xi_{jk}(\boldsymbol{s}_i))^T$, $\forall\, i = 1, \ldots, n$. We model the jk-dimensional function $\boldsymbol{\xi}(\cdot)$ with a jk-dimensional $\mathcal{GP}$, so that one realisation $\{\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2), \ldots, \boldsymbol{\xi}(\boldsymbol{s}_n)\}$, from this $\mathcal{GP}$, is jointly matrix normal, with adequate parametrisation. We represent this as

$$\{\boldsymbol{\xi}(\boldsymbol{s}_1), \boldsymbol{\xi}(\boldsymbol{s}_2), \ldots, \boldsymbol{\xi}(\boldsymbol{s}_n)\} \sim \mathcal{MN}_{n,jk}(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\Omega}), \tag{2.1}$$

where the mean matrix of this matrix normal distribution is the $n \times jk$-dimensional matrix $\boldsymbol{\mu}$, the left covariance matrix is the $n \times n$-dimensional $\boldsymbol{A}$ and the right covariance matrix is the $jk \times jk$-dimensional matrix $\boldsymbol{\Omega}$. These individual matrix-variate parameters of this distribution stem from the parametrisation of the high-dimensional $\mathcal{GP}$ that is used to model $\boldsymbol{\xi}(\cdot)$; we discuss such parametrisation below. Before proceeding to that, we note that Equation 2.1 is the same as saying that the likelihood is matrix normal.

2.1. Parameters of the matrix-normal distribution

Assuming $\boldsymbol{\xi}(\cdot)$ to be continuous, the applicability of a stationary covariance function is expected to suffice. We choose to implement the popularly used square exponential covariance function (Rasmussen and Williams, 2006; Scholkopf and Smola, 2002; Santner et al., 2003). This covariance function is easy to implement and renders the sampled functions smooth and infinitely differentiable. Also, we relax the choice of a zero mean function though that is another popular choice. Instead we choose to define the mean function in a way that is equivalent to the suggestion that the data is viewed as centred around a linear model with the residuals characterised by a vector-variate $\mathcal{GP}$ (A. O'Hagan, 1978; Cressie, 1993). We then integrate over all such possible global intercepts to arrive at a result that is more general than if the mean is fixed at zero. An advantage of the non-zero mean function is that in the limit of the smoothness parameters (characterising the smoothness of the functions sampled from this $\mathcal{GP}$) approaching large values, the random function reduces to a linear regression model. This appears plausible, as distinguished from the result that in this limit of very large smoothness, the random function will concur with the errors, as in models with a zero mean function.

The non-zero mean function $\boldsymbol{\mu}(\cdot)$ of the $\mathcal{GP}$ is represented as factored into a matrix $\boldsymbol{H}$ that bears information about its shape and another ($\boldsymbol{B}$) that tells us about its amplitude, or the extent to which this chosen mean function deviates from being zero. Thus, $\boldsymbol{\mu}(\cdot) := \boldsymbol{HB}$, where

$$\begin{aligned} \boldsymbol{H}^T &:= [\boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_1), \ldots, \boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_n)], \quad \text{with} \\ m &:= d+1 \\ \boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_i) &= (1, s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(d)})^T \end{aligned} \tag{2.2}$$

where $\boldsymbol{s}_i = (s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(d)})^T$ for $i = 1, \ldots, n$ and we have recalled the suggestion that such a non-zero mean function be expressed in terms of a few basis functions (Rasmussen and Williams, 2006), (prompting us to choose to fix this functional form such that $\boldsymbol{h}(\boldsymbol{s}) := (1, \boldsymbol{s})^T$ for all values of $\boldsymbol{S}$). A similar construct was used by Blight and Ott (1975) who performed a $\mathcal{GP}$-based polynomial regression analysis. Thus, in our treatment, $\boldsymbol{h}(\cdot)$ is a $(d+1)$-dimensional vector. The coefficient matrix $\boldsymbol{B}$ is

$$\boldsymbol{B} = (\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{\beta}_{1k}, \ldots, \boldsymbol{\beta}_{jk}) \tag{2.3}$$

where for $p = 1, \ldots, j$, $p' = 1, \ldots, k$, $\boldsymbol{\beta}_{pp'}$ is an m-dimensional column vector. As we choose to set $m = d+1$, $\boldsymbol{B}$ is a matrix with $d+1$ rows and jk columns.

The covariance function of the $\mathcal{GP}$ is again represented as factored into a matrix $\boldsymbol{\Omega}$ that tells us about the amplitude of the covariance and another $\boldsymbol{A}$ that bears information about its shape. The amplitude matrix $\boldsymbol{\Omega}$ is $jk \times jk$-dimensional and is defined as

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{C} \tag{2.4}$$

where $\boldsymbol{\Sigma}$ is the $k \times k$ matrix telling us the amplitude of the covariance amongst the j different observations, for each of the k components of the data vector, at a fixed value of $\boldsymbol{S}$. On the other hand, $\boldsymbol{C}$ is the $j \times j$ matrix giving the amplitude of covariance amongst the k different components of the vector-valued observable, at each of the j observations, at a given value of $\boldsymbol{S}$. Thus in our application, an element of $\boldsymbol{\Sigma}$ is the matrix is the amplitude of the covariance of a given component of the velocity vectors of the different stars that are observed. This matrix can then tell us about how a given component of the velocity vectors of the different stars in the observed sample, correlate with each other. On the other hand, the matrix $\boldsymbol{C}$ informs us about the amplitude of covariance amongst the different components of the velocity vectors of a given star in the sample.

We realise that under the assumption of Gaussian errors in the measurements, the error variance matrix will be added to $\boldsymbol{\Omega}$. We discuss this in detail later in Section 2.6.

The shape of the covariance function is borne by the matrix $\boldsymbol{A}$ which is $n \times n$-dimensional. Given our choice of square exponential covariance function, it is defined as

$$\begin{aligned} \boldsymbol{A}^{(n\times n)} &:= [a(\cdot,\cdot)], \quad \text{where} \\ a(\boldsymbol{s},\boldsymbol{s}') &\equiv \exp\{-(\boldsymbol{s}-\boldsymbol{s}')^T\boldsymbol{Q}(\boldsymbol{s}-\boldsymbol{s}')\}, \end{aligned} \tag{2.5}$$

for any 2 values $\boldsymbol{s}$ and $\boldsymbol{s}'$ of $\boldsymbol{S}$. Here, $\boldsymbol{Q}^{(d\times d)}$ represents the inverse of the scale length that underlies correlation between functions at any two values of the function variable. In other words, $\boldsymbol{Q}$ is the matrix of the smoothness parameters. Thus, $\boldsymbol{Q}$ is a matrix that bears information about the smoothness of the sampled functions; it is a diagonal matrix consisting of d non-negative smoothness parameters denoted by $b_1, \ldots, b_d$. In other words, we assume the same smoothness for each component function of $\boldsymbol{\xi}(\cdot)$. This smoothness is determined by the parameters $b_1, \ldots, b_d$. We will learn these smoothness parameters in our work from the data. Of course, though we say that the smoothness is learnt in the data, the underlying effect of the choice of the square exponential covariance function on the smoothness of the sampled functions is acknowledged. Indeed, as Snelson (2007) states, one concern about the square exponential function is that it renders the functions sampled from it as artificially smooth. An alternative covariance function, such as the Matern class of covariances (Matern, 1986; Tilmann Gneiting and William Kleiber and Martin Schlather, 2010; Snelson, 2007), could give rise to sampled functions that are much rougher than those obtained using the square exponential covariance function, for the same values of the hyper-parameters of amplitude and scale that characterise these covariance functions(see Chapter 1 of Snelson's thesis).

Let $\omega_{r\ell}$ denote the (r,ℓ)-th element of $\boldsymbol{\Omega}$, $c_{r\ell}$ the (r,ℓ)-th element of $\boldsymbol{C}$ and let $\sigma_{r\ell}$ denote the (r,ℓ)-th element of $\boldsymbol{\Sigma}$. Let the ℓ-th component function of $\boldsymbol{\xi}(\cdot)$ be $\xi_\ell(\cdot)$ with $\ell = m_1 k + m_2$, where $\ell = 1, \ldots, jk$ and $m_2 = 1, 2, \ldots, k$, $m_1 = 0, 1, \ldots, j-1$. Then the correlation between the components of $\boldsymbol{\xi}(\cdot)$ yields the following correlation structures:

$$corr\left(\xi_{m_1k+m_2}(\boldsymbol{s}_i), \xi_{m_1'k+m_2}(\boldsymbol{s}_i)\right) = \frac{\sigma_{m_1m_1'}}{\sqrt{\sigma_{m_1m_1}\sigma_{m_1'm_1'}}} \;\forall\; m_2, i \;\text{ and }\; m_1 \neq m_1' \tag{2.6}$$

$$corr\left(\xi_{m_1k+m_2}(\boldsymbol{s}_i), \xi_{m_1k+m_2'}(\boldsymbol{s}_i)\right) = \frac{c_{m_2m_2'}}{\sqrt{c_{m_2m_2}c_{m_2'm_2'}}} \;\forall\, m_1, i \;\text{ and }\; m_2 \neq m_2' \tag{2.7}$$

$$corr\left(\xi_{m_1k+m_2}(\boldsymbol{s}_i), \xi_{m_1'k+m_2'}(\boldsymbol{s}_i)\right) = \frac{c_{m_2m_2'}\sigma_{m_1m_1'}}{\sqrt{c_{m_2m_2}\sigma_{m_1m_1}c_{m_2'm_2'}\sigma_{m_1'm_1'}}} \forall i, m_1 \neq m_1', m_2 \neq m_2' \tag{2.8}$$

$$corr\left(\xi_\ell(\boldsymbol{s}_1), \xi_\ell(\boldsymbol{s}_2)\right) = a(\boldsymbol{s}_1, \boldsymbol{s}_2) \forall\; \ell \;\text{ and }\; \boldsymbol{s}_1 \neq \boldsymbol{s}_2 \tag{2.9}$$

The 1st of the above 4 equations shows the correlation between the component functions for the same component of the vector-valued observable at 2 (of the j) different measurements, taken at a given value of the $\boldsymbol{S}$. For a given measurement, the correlation between 2 different components of (the k components of) the observable is given by the 2nd equation above. For a given value of $\boldsymbol{S}$, if we seek the correlation between the component functions for 2 different measurements of 2 different components of the observables, this is provided in the 3rd equation. The correlation between component functions for 2 different values of $\boldsymbol{S}$ is given in the last of the above 4 equation. Then these 4 correlations give the full correlation structure amongst components of $\boldsymbol{\xi}(\cdot)$.

2.2. *Likelihood*

The training data is the $n \times jk$-dimensional matrix $\mathcal{D}_s = (\mathbf{v}_1 \vdots \mathbf{v}_2 \vdots \ldots \vdots \mathbf{v}_n)^T$ where $\mathbf{v}_i$ is the jk-dimensional synthetic motion vector generated at design vector $\boldsymbol{s}_i^\star$, $i = 1, 2, \ldots, n$. To express the likelihood, we recall that the distribution of the training data $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, i.e. the joint distribution of $\{\boldsymbol{\xi}(\boldsymbol{s}_1^\star), \boldsymbol{\xi}(\boldsymbol{s}_2^\star), \ldots, \boldsymbol{\xi}(\boldsymbol{s}_n^\star)\}$ is matrix normal (Equation 2.1). In order to achieve this likelihood, we rewrite the $\boldsymbol{S}$-dependent parameters of this matrix normal distribution at the values of $\boldsymbol{S}$ at which the training data $\mathcal{D}_s$ is realised, i.e. in terms of the design vectors. Thus, we define

- the $n \times jk$-dimensional mean function $\boldsymbol{H}_D\boldsymbol{B}$, where the linear form of the mean structure is contained in $\boldsymbol{H}_D^{(n\times m)} := [\boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_1^\star), \ldots, \boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_n^\star)]$ (and the coefficient matrix $\boldsymbol{B}$ is defined in Equation 2.3).
- the square exponential factor in the covariance matrix $\boldsymbol{A}_D^{(n\times n)} := [\exp\{-(\boldsymbol{s}^\star - \boldsymbol{s}'^\star)^T \boldsymbol{Q}(\boldsymbol{s}^\star - \boldsymbol{s}'^\star)\}]$ (see Equation 2.5).

Then it follows from the matrix normal distribution of Equation 2.1–with mean function defined in Equation 2.2 and Equation 2.3, and covariance matrix defined using Equation 2.5 and Equation 2.4–that $\mathcal{D}_s$ is distributed as matrix normal with mean matrix $\boldsymbol{H}_D\boldsymbol{B}$, left covariance matrix $\boldsymbol{A}_D$ and right covariance matrix $\boldsymbol{\Omega}$, i.e.

$$[\mathcal{D}_s \mid \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\Sigma}, \boldsymbol{Q}] \sim \mathcal{MN}_{n,jk}(\boldsymbol{H}_D\boldsymbol{B}, \boldsymbol{A}_D, \boldsymbol{\Omega}) \tag{2.10}$$

Thus, using known ideas about the matrix normal distribution - see Dawid (1981), Carvalho and West (2007) - we write

$$[\mathcal{D}_s \mid \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\Sigma}, \boldsymbol{Q}] = \frac{1}{(2\pi)^{\frac{njk}{2}} |\boldsymbol{A}_D|^{\frac{jk}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} tr \left[\boldsymbol{\Omega}^{-1}(\mathcal{D}_s - \boldsymbol{H}_D\boldsymbol{B})^T \boldsymbol{A}_D^{-1}(\mathcal{D}_s - \boldsymbol{H}_D\boldsymbol{B})\right]\right\} \tag{2.11}$$

The interpretation of the above is that the r-th row of $[\mathcal{D}_s|\mathbf{B}, \boldsymbol{\Sigma}, \mathbf{C}, \boldsymbol{Q}]$ is multivariate normal with mean corresponding to row of the mean matrix $\boldsymbol{H}_D\boldsymbol{B}$ and with covariance matrix $\boldsymbol{\Omega}$. Rows r and ℓ of $[\mathcal{D}_s|\mathbf{B}, \boldsymbol{\Sigma}, \mathbf{C}, \boldsymbol{Q}]$ has covariance matrix $a(\boldsymbol{s}_r, \boldsymbol{s}_\ell)\boldsymbol{\Omega}$. Similarly, the ℓ-th column of it is distributed as multivariate normal with mean being the ℓ-th column of $\boldsymbol{H}_D\boldsymbol{B}$ and with covariance matrix $\omega_{\ell,\ell}\boldsymbol{A}_D$, where $\omega_{r,\ell}$ denotes the (r, ℓ)-th element of $\boldsymbol{\Omega}$. The covariance between columns r and ℓ is given by the matrix $\omega_{r,\ell}\boldsymbol{A}_D$.

2.3. *Estimating* $\boldsymbol{s}^{(new)}$

In order to predict the unknown model parameter vector $\boldsymbol{s}^{(new)}$ when the input is the measured real data vector $\mathbf{v}^{(test)}$, we would need the posterior predictive distribution of $\boldsymbol{s}^{(new)}$, given $\mathbf{v}^{(test)}$ and the training data $\mathcal{D}_s$. This posterior predictive is usually computed by integrating over all the matrix-variate $\mathcal{GP}$ parameters realised at the chosen design vectors $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star$.

While it is possible to analytically integrate over $\boldsymbol{B}$ and $\boldsymbol{C}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{Q}$ cannot be analytically integrated out. In fact, we find it useful to learn the d smoothing parameters i.e. the d diagonal elements of $\boldsymbol{Q}$, given the data. Thus, one useful advantage of our method is that the smoothness of the process does not need to be imposed by hand, but can be learnt from the data, if desired.

Given that we are then learning of $\boldsymbol{s}^{(new)}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{Q}$, we rephrase our motivation as seeking to compute the joint posterior probability of $\boldsymbol{s}^{(new)}$, $\boldsymbol{Q}$ and $\boldsymbol{\Sigma}$, conditional on the real data and the training data, for a choice of the design set. In fact, we achieve a closed form expression of this joint posterior of $\boldsymbol{s}^{(new)}$, $\boldsymbol{Q}$ and $\boldsymbol{\Sigma}$, by integrating over the other hyper-parameters, namely, the amplitude of the mean function ($\boldsymbol{B}$) and the matrix $\boldsymbol{C}$ that bears information about covariance between different components of the data vector for each of the j observations, at a fixed value of $\boldsymbol{S}$. From this closed form expression, the marginal posterior probability densities of $\boldsymbol{Q}$, $\boldsymbol{\Sigma}$ and any of the d components of the $\boldsymbol{s}^{(new)}$ vector can be obtained, using the transformation based MCMC sampling method (Dutta and Bhattacharya, 2013) that we adopt.

Thus, for a given choice $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star$ of the design vectors, the posterior distribution $[\mathbf{s}^{(new)}, \boldsymbol{\Sigma}, \boldsymbol{Q} | \mathbf{v}^{(test)}, \mathcal{D}_s]$ is sought, by marginalising $[\mathbf{s}^{(test)}, \boldsymbol{\Sigma}, \boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{C} | \mathbf{v}^{(test)}, \mathcal{D}_s]$ over the process matrices $\boldsymbol{B}$ and $\boldsymbol{C}$.

2.4. *Priors used*

We use uniform prior on $\boldsymbol{B}$ and a simple non-informative prior on $\boldsymbol{C}$, namely, $\pi(\boldsymbol{C}) \propto \mid \boldsymbol{C} \mid^{-(j+1)/2}$. As for the priors on the other parameters, we assume uniform prior on $\boldsymbol{Q}$ and use the non-informative prior $\pi(\boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-(k+1)/2}$. The prior information available in the literature will be considered to select the prior on $\boldsymbol{s}^{(new)}$; below we use uniform priors on all components of the $\boldsymbol{s}^{(new)}$ vector (see Section 4 for greater details in regard to the application that we discuss later).

2.5. *Posterior of $\boldsymbol{s}^{(new)}$ given training and test data*

Since our interest lies in estimating $\boldsymbol{s}^{(new)}$, given the real (test) data and the simulated (training) data, as well as in learning the smoothness parameter matrix $\boldsymbol{Q}$ and the matrix $\boldsymbol{\Sigma}$ that bears the covariance amongst the j observables, we compute the joint posterior probability density $[\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathbf{v}^{(test)}, \mathcal{D}_s]$. As expressed above, we achieve this by writing $[\boldsymbol{s}^{(new)}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathbf{v}^{(test)}, \mathcal{D}_s]$ and marginalise over $\boldsymbol{B}$ and $\boldsymbol{C}$.

To construct an expression for this posterior distribution, we first collate the training and test data to construct the augmented data set $\mathcal{D}_{aug}^T = (\mathbf{v}_1^T \vdots \ldots \vdots \mathbf{v}_n^T \vdots (\mathbf{v}^{(test)})^T)$. Then the set of values of the model parameter vector $\boldsymbol{S}$ that supports $\mathcal{D}_{aug}$ is $\{\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star, \boldsymbol{s}^{(new)}\}$ of which only $\boldsymbol{s}^{(new)}$ is unknown.

We next write the $\boldsymbol{S}$-dependent matrix-variate parameters at those values of $\boldsymbol{S}$ at which the augmented data set is realised. Thus we define

- $\boldsymbol{H}_{\mathcal{D}_{aug}}^{((n+1)\times m)} := [\boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_1^\star), \ldots, \boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_n^\star), \boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}^{(new)})]$, where our choice of the functional form of $\boldsymbol{h}(\cdot)$ has been given in Section 2 and we also set $m = d + 1$,
- $\boldsymbol{A}_{\mathcal{D}_{aug}}^{((n+1)\times(n+1))} := [\exp\{-(\boldsymbol{s}_i' - \boldsymbol{s}_{i'}')^T \boldsymbol{Q} (\boldsymbol{s}_i' - \boldsymbol{s}_{i'}')\}]$ where $\boldsymbol{s}_i'$ and $\boldsymbol{s}_{i'}'$ are members of the set $\{\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star, \boldsymbol{s}^{(new)}\}$,
- $\boldsymbol{M}_{aug} := \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1} - \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1} \boldsymbol{H}_{\mathcal{D}_{aug}} [\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1} \boldsymbol{H}_{\mathcal{D}_{aug}}]^{-1} \boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}$.
- $(\mathcal{D}_{aug}^T \boldsymbol{M}_{aug} \mathcal{D}_{aug})^{(jk\times jk)} := [\boldsymbol{M}_{tu}^*; t, u = 1, \ldots, k]$, where $\boldsymbol{M}_{tu}^*$ is a matrix with j rows and j columns. Given $\boldsymbol{\Sigma}$, we define $m = d + 1$ and ψ_{tu}^{-1} as the (t, u)-th element of $\boldsymbol{\Sigma}^{-1}$, so that $(n + 1 - m)k\hat{\boldsymbol{C}}_{GLS,aug} := \sum_{t=1}^k \sum_{u=1}^k \psi_{tu}^{-1} \boldsymbol{M}_{tu}^*$, where $(n + 1 - m)k\hat{\boldsymbol{C}}_{GLS,aug}$ is used in the closed-form expression for $[\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathbf{v}^{(test)}, \mathcal{D}_s]$ that we seek.

The priors used on $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{Q}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{s}^{(new)}$ are listed in Section 2.4. Using these, and recalling Equation 2.11, we get the joint posterior probability density of all unknown parameters given all data, i.e.

$$[\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C} \mid \mathbf{v}^{(test)}, \mathcal{D}_s] \propto [\mathcal{D}_{aug} \mid \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{s}^{(new)}][\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{s}^{(new)}],$$

which we then marginalise over $\boldsymbol{B}$ and $\boldsymbol{C}$ to get the joint posterior $[\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathbf{v}^{(test)}, \mathcal{D}_s]$, as

$$\begin{aligned}
&[\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathbf{v}^{(test)}, \mathcal{D}_s] \\
= & \int\int [\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C} \mid \mathbf{v}^{(test)}, \mathcal{D}_s] d\boldsymbol{B} d\boldsymbol{C} \\
\propto & \ |\boldsymbol{A}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} |\{\boldsymbol{H}_{\mathcal{D}_{aug}}\}^T \{\boldsymbol{A}_{\mathcal{D}_{aug}}\}^{-1} \{\boldsymbol{H}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} \times |\boldsymbol{\Sigma}|^{-\frac{j(n+1-m)+k+1}{2}} |(n+1-m)k\hat{\boldsymbol{C}}_{GLS,aug}|^{-\frac{(n+1-m)k}{2}}
\end{aligned} \tag{2.12}$$

Thus, we obtain a closed-form expression of the joint posterior of $\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{\Sigma}$, given training and test data, for a given choice of the design matrix (Equation 2.12), up to a normalising constant. The $\mathcal{GP}$ prior is strengthened by the n number of samples taken from it at the training stage. We sample from the achieved posterior using MCMC techniques to achieve the marginal posterior probabilities of $\boldsymbol{Q}$, $\boldsymbol{\Sigma}$ or any component of $\boldsymbol{s}^{(new)}$, given all data. We conduct posterior inference using the TMCMC methodology (Dutta and Bhattacharya, 2013) that works by constructing proposals that are deterministic bijective transformations of a random vector drawn from a chosen distribution.

2.6. Errors in measurement

In our application, the errors in the measurements are small and will be ignored for the rest of the analysis. In general, when errors in the measurements that comprise the training data and the test data are not negligible, we assume Gaussian measurement errors ε_t, in $\mathbf{v}_t$, with $t = 1, 2, \ldots$, such that $\varepsilon_t \sim \mathcal{N}_{jk}(\mathbf{0}, \varsigma)$, where $\varsigma = \boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_2$; $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ being positive definite matrices. If both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are chosen to be diagonal matrices, then ς is a diagonal matrix; assuming same diagonal elements would simplify ς to be of the form $\varphi \times \boldsymbol{I}$, where $\boldsymbol{I}$ is the $jk \times jk$-th order identity matrix. This error variance matrix ς must be added to $\boldsymbol{\Omega}$ before proceeding to the subsequent calculations. TMCMC can be then be used to update ς.

3. Case study

Using the methodology discussed above we attempt an estimate of the unknown Milky Way feature parameter vector $\boldsymbol{S} \in \mathbb{R}^d$ using the available stellar velocity data. In our application, the dimensionality of $\boldsymbol{S}$ is 2 as we estimate the coordinates of the radial location $r_\odot$ of the Sun with respect to the Galactic centre and the angular separation $\phi_\odot$ of the Sun-Galactic centre line from a pre-set line in the Milky Way disk (see Figure 1 in supplementary section **S-1**). Then for the Sun, $R = r_\odot$ and $\Phi = \phi_\odot$ where the variable R gives radial distance from the Galactic centre of any point on the disk of the Milky Way and the variable Φ gives the angular separation of this point from this chosen pre-set line. The reason for restricting our application to the case of d=2 is the existence of simulated stellar velocity data (aka training data) generated by scanning over chosen guesses for $r_\odot$ and $\phi_\odot$, with all other feature parameters held constant. If simulated data distinguished by choices of other Milky Way feature parameters become available, then the implementation of such data as training data will be possible, allowing then for the learning of Milky way parameters in addition to $r_\odot$ and $\phi_\odot$. In this method, computational costs are the only concern in extending to cases of $d > 2$; extending to a higher dimensional $\boldsymbol{S}$ only linearly scales computational costs (Section 6).

Also, the stellar velocity vector is 2-dimensional, i.e. k=2 in this application. Then the measured data in this application is a $j \times 2$-dimensional matrix. In our Bayesian approach, a much smaller j (=50) allows for inference on the unknown value $\boldsymbol{s}^{(new)}$ of the Milky Way feature parameter vector, than j ~3000 that is demanded by the aforementioned calibration approach used by Chakrabarty (2007).

In our application, the available data include the measured or test data and 4 sets of synthetic (or training) data sets obtained via dynamical simulations of each of 4 distinct base-astronomical models of our galaxy, advanced by Chakrabarty (2007). As the analysis is performed with each training data set at a time, we do not include reference to the corresponding base model in the used notation. The simulated data presented in Chakrabarty (2007) that we use here, is generated at 216 distinct values of $\boldsymbol{S}$, i.e. n=216. Thus, our design set comprises the 216 chosen values of $\boldsymbol{S}$: $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_{216}^\star$. For each of the 4 base astrophysical models, at each chosen $\boldsymbol{s}_i^\star$, 50 2-dimensional stellar velocity vectors are generated from dynamical simulations of that astrophysical model (of the Milky Way), performed at that value of $\boldsymbol{S}$. These 50 2-dimensional velocity vectors are treated in our work as a 50×2=100-dimensional motion vector $\mathbf{v}_i$; $i = 1, \ldots, 216$. Then at the 216 design vectors, $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_{216}^\star$, 216 motion vectors are generated: $\mathbf{v}_1, \ldots, \mathbf{v}_{216}$. Then the training data in our work comprises all such motion vectors and is represented as $\mathcal{D}_s^{(216\times 100)}$. The real or test data is treated in our work as the 100-dimensional motion vector $\mathbf{v}^{(test)}$.

As said above, there are 4 distinct training data sets available from using the 4 base astronomical models of the Milky Way, as considered by Chakrabarty (2007). The choice of the base astrophysical model is distinguished by the ratio of the rates of rotation of the spiral to the bar, Ω_s/Ω_b. That this ratio is relevant to stellar motions in the Galaxy is due to the fact that Ω_s/Ω_b can crucially control the degree of chaos in the Galactic model[1]. Thus, the 4 base astrophysical models are differently chaotic. This results in 4 distinct simulated velocity data sets $\mathcal{D}_s^{(1)}$, $\mathcal{D}_s^{(2)}$, $\mathcal{D}_s^{(3)}$, $\mathcal{D}_s^{(4)}$ that bear the effects of such varying degrees of chaos, each generated at the chosen design set $\{\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_n^\star\}$. Details of the dynamical simulations performed on the 4 astrophysical models are given in the supplementary section **S-2**.

3.1. Details of our implementation of TMCMC

As indicated above, we use the Transformation-based MCMC (TMCMC) advanced by Dutta and Bhattacharya (2013) to conduct posterior inference. In TMCMC, high-dimensional pa-

[1] For example, it is well known in chaos theory that when Ω_s/Ω_b is such that one of the radii at which the bar and the stellar disk resonate, concurs with a radius at which the spiral and the stellar disk resonate, global chaos is set up in the system (G. Walker and J. Ford, 1969). Chakrabarty and Sideris (2008) have corroborated that the degree of chaos is maximal in the astrophysical Galactic model marked by such a ratio (Ω_s/Ω_b=22/55). They report that in models marked by slightly lower (Ω_s/Ω_b=18/55) or higher (Ω_s/Ω_b = 25/55) values of this ratio, chaos is still substantial. In the Galactic model that precludes the spiral however, chaos was quantified to be minimal. It is these 4 states of chaos - driven by the 4 values of Ω_s/Ω_b - that mark the 4 astrophysical models as distinct.

rameter spaces are explored by constructing bijective deterministic transformations of a low-dimensional random vector. The random vector of which a proposal density is a transformation of, can be chosen to be of dimensionality between 1 and the dimensionality of the parameters under the target posterior. The acceptance ratio in TMCMC does not depend upon the distribution of the chosen random vector. In our application we use TMCMC to update the entire block $(\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{\Sigma})$ at the same time using additive transformations of a one-dimensional random variable $\epsilon \sim \mathcal{N}(0,1)I_{\{\epsilon>0\}}$. In the t-th iteration, the state of the unknown parameters is $(\boldsymbol{s}^{(new,t)}, \boldsymbol{Q}^{(t)}, \boldsymbol{\Sigma}^{(t)}) := \boldsymbol{\varphi}^{(t)}$. We update $\boldsymbol{\varphi}^{(t)}$ by setting, with probabilities π_j and $(1-\pi_j)$, $\varphi_j^{(t+1)} = \varphi_j^{(t)} \pm c_j\epsilon$ (forward transformation) and $\boldsymbol{\varphi}_j^{(t+1)} = \boldsymbol{\varphi}_j^{(t)} - c_j\epsilon$ (backward transformation), respectively, where, for $j = 1, \ldots, d$, π_j are appropriately chosen probabilities and c_j are appropriately chosen scaling factors. Assume that for $j_1 \in \mathcal{U}$, $\boldsymbol{\varphi}_{j_1}^{(t)}$ gets the positive transformation, while for $j_2 \in \mathcal{U}^c$, $\boldsymbol{\varphi}_{j_2}^{(t)}$ gets the backward transformation. Here $\mathcal{U} \cup \mathcal{U}^c = \{1, \ldots, d^*\}$, where $d^* = 2d + \frac{k(k+1)}{2}$. The proposal $\boldsymbol{\varphi}^{(t+1)}$ is accepted with acceptance probability given in Supplementary Section **S-3**. Once the proposal mechanism and the initial values are decided, we discard the first 100,000 iterations of our final TMCMC run as burn-in and stored the next 1,000,000 iterations for inference. For each model it took approximately 6 hours on a laptop to generate 1,100,000 TMCMC iterations.

4. Results using real data

The training data that we use was obtained by Chakrabarty (2007), by choosing the solar radial location from the interval $[1.7, 2.3]$ in model units. This explains the motivation for selecting the bounds on $r_\odot$ to be the edges of this interval. Here, values of distances are expressed in the units implemented in the base astrophysical models of the Milky Way. However, to make sense of the results we have obtained, these model units will need to be scaled to provide values in real astronomical units of distances inside galaxies, such as the "kiloparsec" (abbreviated as "kpc"). A distance of 1 in model unit scales to $\dfrac{\mathcal{R}}{\hat{r}_\odot}$kpc, where $\mathcal{R}$ is the solar radius obtained in independent astronomical studies (Binney and Merrifield, 1998, $\mathcal{R}$=8kpc) and $\hat{r}_\odot$ is the estimate of the solar radius in our work. The ulterior aim in estimating the solar radius is in estimating the rotational frequency Ω_b of the bar, where $\Omega_b = \dfrac{v_0}{\frac{\mathcal{R}}{\hat{r}_\odot}}$, with v_0=220kms^{-1} and $\mathcal{R}$=8kpc. Then, we get $\Omega_b = \dfrac{220}{\frac{8}{\hat{r}_\odot}}$kms^{-1}/kpc. See Section **S-1** of the attached supplementary material to see a schematic representation of the central bar in the Galaxy and Section **S-2** for details of the scaling between the model units and real astronomical units.

Our other estimate is of the angular separation between the long axis of the bar and the line that joins the Sun to the Galactic centre. It is suggested in past astronomical modelling work to be an acute angle (Chakrabarty, 2007; Englmaier and Gerhard, 1999; Fux, 2001). Indeed, the training data used here was generated in simulations performed by Chakrabarty (2007), in which $\phi_\odot$ is chosen from the interval $[0, 90^\circ]$. This motivates the consideration of the interval of $[0, 90^\circ]$ for the angular location of the Sun.

Given the bounds on $r_\odot$ and $\phi_\odot$ presented above, in our TMCMC algorithm, we reject those moves that suggest $r_\odot$ and $\phi_\odot$ values that fall outside these presented intervals.

The 4 astrophysical models of the Galaxy that were used to generate the 4 training data sets, are marked by the same choice of the value of Ω_b and the background Galactic model parameters, while

TABLE 1

Summary of the posterior distributions of the radial component $r_\odot$ and azimuthal component $\phi_\odot$ of the unknown observer location vector for the 4 base astrophysical models and the unknown bar rotational frequency Ω_b computed using the 95% HPDs on the learnt radial location $r_\odot$ in these models.

Model	$r_\odot$ (in units of r_{CR})		Ω_b (in kms^{-1}/kpc)	$\phi_\odot$	
	Mode	95% HPD	95% HPD	Mode	95% HPD
$bar6$	2.20	$[2.04, 2.30]$	$[56.1, 63.25]$	23.50	$[21.20, 25.80]$
$sp3bar3$	1.73	$[1.70, 2.26] \cup [2.27, 2.28]$	$[46.75, 62.15] \cup [62.45, 62.7]$	18.8	$[9.6, 61.5]$
$sp3bar3\_18$	1.76	$[1.70, 2.29]$	$[46.75, 62.98]$	32.5	$[17.60, 79.90]$
$sp3bar3\_25$	1.95	$[1.70, 2.15]$	$[46.75, 59.12]$	37.6	$[28.80, 40.40]$

they are distinguished by the varying choices of the ratio $\Omega_s : \Omega_b$, where the Galactic spiral pattern rotates with rate Ω_s. In fact, the astrophysical model bar_6 is the only one that does not include the influence of the spiral pattern while the other three astrophysical models include the influence of both the bar and the spiral. For the astrophysical models $sp3bar3\_18$, $sp3bar3$ and $sp3bar3\_25$, $\Omega_s : \Omega_b$ is respectively set to $18\Omega_b/55,\ 22\Omega_b/55,\ 25\Omega_b/55$. The physical effect of this choice is to induce varying levels of chaoticity in the 4 astrophysical models. Thus, Chakrabarty and Sideris (2008) confirmed that of the 4 models, bar_6 manifests very low chaoticity while $sp3bar3$ manifests maximal chaos, though both $sp3bar3\_18$, $sp3bar3\_25$ are comparably chaotic.

Ancillary real data needs to be brought in to judge the relative fit amongst the astrophysical base models. In fact, Chakrabarty (2007) brought in extra information to perform model selection. Such information was about the observed variance of the components of stellar velocities and this was used to rule out the model bar_6 as physically viable, though the other three models were all acceptable from the point of view of such ancillary observations that are available. This led to the inference that $\Omega_s \in [18\Omega_b/55, 25\Omega_b/55]$.

It is to be noted that if there was 1 data set and we were trying to fit 4 different models to that same data, then it is very much possible that for this 1 data set, the average of 4 models could have been achieved. However, here we are dealing with 4 base models, each of which is giving rise to a distinct training data set, in fact under mutually contradicting physics. Therefore, such model averaging is not relevant for this work. Cross-validation of these 4 models is indeed possible and we present this in Section **S-5** of the attached Supplementary Materials.

The marginal posterior densities of $(r_\odot, \phi_\odot)$ corresponding to the 4 base astrophysical models of the Milky Way, are shown in Figures 1, 2, 3 and 4. It merits mention that the multi-modality manifest in the marginal posterior distributions in 3 of the 4 base models is not an artifact of inadequate convergence but is a direct fallout of the marked amount of chaoticity in all 3 base models except in the model bar_6, (Chakrabarty and Sideris, 2008). In Section **S-6**, we discuss the connection between chaos and consistency of multiple observer locations with available stellar velocity data.

Table 1 presents the posterior mode, the 95% highest posterior density (HPD) credible region of $r_\odot$ and $\phi_\odot$ respectively, associated with the four base models. Here $r_\odot$ is expressed in the model units of length, i.e. in units of r_{CR}. $\phi_\odot$ is expressed in degrees. The HPDs are computed using the methodology discussed in Carlin and Louis (1996). Disjoint HPD regions, characterise the highly multi-modal posterior distributions of the unknown location. Using the 95% HPDs of the estimate $\hat{r_\odot}$ expressed in model units, and using the independently known astronomical measurement of the solar radial location as 8kpc, the bar rotational frequency Ω_b is computed (see third enumerated point discussed above) in Table 1.

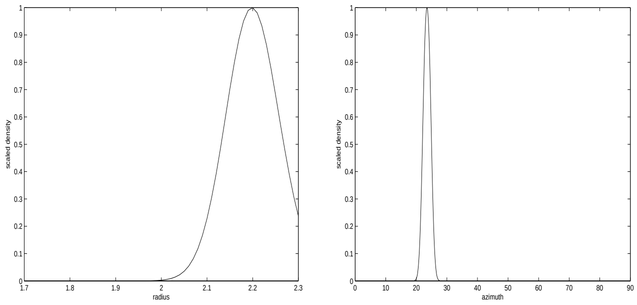


FIG 1. *Posteriors of $r_\odot$ in model units of r_{CR} and $\phi_\odot$ (in degrees) for the model bar_6.*

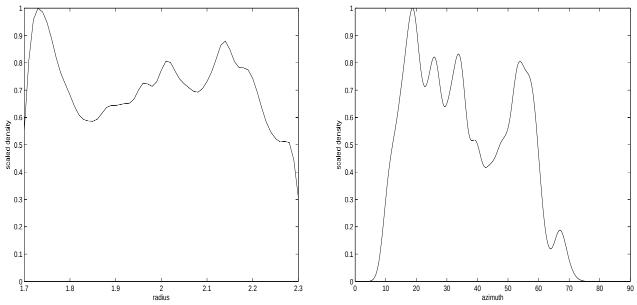


FIG 2. *Posteriors of $r_\odot$ in units of r_{CR} and $\phi_\odot$ (in degrees) for the model $sp3bar3$.*

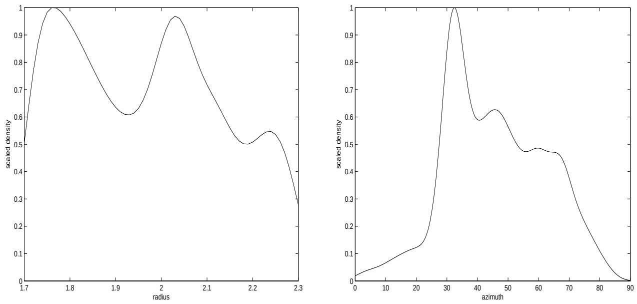


FIG 3. *Posteriors of $r_\odot$ in model units of r_{CR} and $\phi_\odot$ (in degrees) for the model $sp3bar3\_18$.*

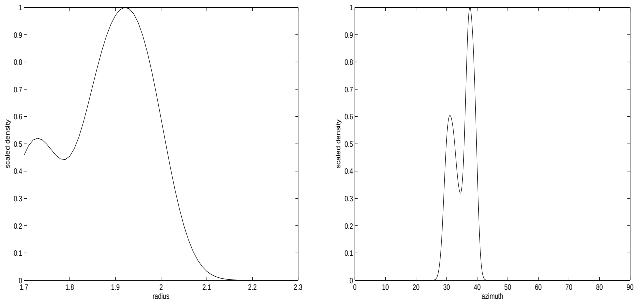


FIG 4. *Posteriors of $r_\odot$ in units of r_{CR} and $\phi_\odot$ (in degrees) for the model $sp3bar3\_25$.*

Summaries of the posteriors (mean, variance and 95% credible interval) of the smoothness parameters b_1, b_2 and $\mathbf{\Sigma}$ are presented in Tables 2, 3. Notable in all these tables are the small posterior

TABLE 2
Summary of the posterior distributions of the smoothness parameters b_1, b_2 for the 4 models.

Model	b_1			b_2		
	Mean	Var	95% CI	Mean	Var	95% CI
bar_6	0.9598155	3.15×10^{-9}	$[0.959703, 0.959879]$	1.005078	2.85×10^{-9}	$[1.004985, 1.005142]$
$sp3bar3$	0.8739616	6.72×10^{-7}	$[0.872347, 0.875052]$	1.003729	8.98×10^{-7}	$[1.002500, 1.005500]$
$sp3bar3\_18$	0.9410686	1.46×10^{-5}	$[0.938852, 0.955264]$	0.999010	4.08×10^{-6}	$[0.997219, 1.004945]$
$sp3bar3\_25$	0.7597931	5.64×10^{-10}	$[0.759743, 0.759833]$	0.992174	2.89×10^{-9}	$[0.992067, 0.992246]$

TABLE 3
Summary of the posterior distribution of the diagonal and one non-diagonal element of $\mathbf{\Sigma}$, from the 4 base astrophysical models.

Model	σ_{11}	σ_{22}	σ_{12}
	95% CI	95% CI	95% CI
bar_6	$[5.40 \times 10^{-5}, 4.0 \times 10^{-4}]$	$[6.20 \times 10^{-5}, 4.76 \times 10^{-4}]$	$[0, 1.30 \times 10^{-5}]$
$sp3bar3$	$[3.66 \times 10^{-3}, 1.03 \times 10^{-2}]$	$[6.53 \times 10^{-3}, 1.83 \times 10^{-2}]$	$[-6.40 \times 10^{-5}, 2.68 \times 10^{-4}]$
$sp3bar3\_18$	$[1.45 \times 10^{-3}, 1.68 \times 10^{-1}]$	$[1.29 \times 10^{-3}, 1.50 \times 10^{-1}]$	$[-1.19 \times 10^{-4}, 2.16 \times 10^{-3}]$
$sp3bar3\_25$	$[1.21 \times 10^{-4}, 5.69 \times 10^{-4}]$	$[1.13 \times 10^{-4}, 5.21 \times 10^{-4}]$	$[-1.00 \times 10^{-6}, 1.50 \times 10^{-5}]$

variances of the quantities in question; this is indicative of the fact that the data sets we used, in spite of the relatively smaller size compared to the astronomically large data sets used in the previous approaches in the literature, are very much informative, given our vector-variate $\mathcal{GP}$-based Bayesian approach. Owing to our Gaussian Process approach, the posterior of $\mathbf{\Sigma}$ should be close to the null matrix *a posteriori* if the choice of the design set and the number of design points are adequate. Quite encouragingly, Table 3, shows that indeed $\mathbf{\Sigma}$ is close to the null matrix *a posteriori*, for all the four models, signifying that the unknown velocity function has been learned well in all the cases.

4.1. Comparison with results in astrophysical literature

The estimates of the anglar separation of the long axis of the bar from the Sun-Galactic centre line and the rotation rate of the bar compare favourably with results obtained by Chakrabarty (2007), Englmaier and Gerhard (1999), Debattista et al. (2002), Benjamin et al. (2005),Antoja et al. (2011). A salient feature of our implementation is the vastly smaller data set that we needed to invoke than any of the methods reported in the astronomical literature, in order to achieve the learning of the two-dimensional vector $\boldsymbol{S}$ - in fact while in the calibration approach of Chakrabarty (2007), the required sample size is of the order of 3,500, in our work, this number is 50. Thus, data sufficiency issues, when a concern, are well tackled by our method.

Upon the analyses of the viable astrophysical models of the Galaxy, Chakrabarty (2007) reported the result that $r_\odot \in [1.9375, 2.21]$ in model units while $\phi_\odot \in [0^\circ, 30^\circ]$, where these ranges correspond to the presented uncertainties on the estimates, which were however, rather unsatisfactorally achieved (see Section 2). The values of the components of $\boldsymbol{S}$, learnt in our work, overlap well with these results. As mentioned above, the models $sp3bar3\_18$, $sp3bar3$ and $sp3bar3\_25$ are distinguished by distinct values of the ratios of the rotational rates of the spiral pattern Ω_s to that of the bar (Ω_b) in the Galaxy. Then the derived estimate for Ω_b (Table 1) suggests values of Ω_s of the Milky Way spiral.

Another point that merits mentions is that the estimates of $r_\odot$ and $\phi_\odot$ presented by Chakrabarty

(2007) exclude the model $sp3bar3$ which could not be used to yield estimates given the highly scattered nature of the corresponding p-value distribution. Likewise, in our work, the same model manifests maximal multi-modality amongst the others, but importantly, our approach allows for the representation of the full posterior density using which, the computation of the 95$\%$ HPDs is performed.

That the new method is able to work with smaller velocity data sets, is an important benefit, particularly in extending the application to galaxies other than our own, in which small numbers of individual stars are going to be tracked in the very near future for their velocities, under observational programmes such as PANStarrs (Johnston et al., 2009) and GAIA (Lindegren et al., 2007; Kucinskas et al., 2005, http://www.rssd.esa.int/index.php?project=GAIA&page=index); the sample sizes of measured stellar velocity vectors in these programmes will be much smaller in external galaxies than what has been possible in our own. At the same time, our method is advanced as a template for the analysis of the stellar velocity data that is available for the Milky Way, with the aim of learning a high-dimensional Galactic parameter vector; by extending the scope of the dynamical simulations of the Galaxy, performed on different astrophysical models of the Milky Way, the Milky Way models will be better constrained. The mission GAIA - a mission of the European Space Agency - is set to provide large sets of stellar velocity data all over the Milky Way. Our method, in conjunction with astrophysical models, can allow for fast learning of local and global model parameters of the Galaxy.

5. Model fitting

In this section we compare the test data with predictions for the observable that we make at a summary $\tilde{\boldsymbol{s}}$ of the posterior of the model parameter vector $\boldsymbol{S}$. To achieve this, we first need to provide a suitable estimator of the function $\boldsymbol{\xi}(\cdot)$ that defines the relatioship between the observable and the model parameter $\boldsymbol{S}$. We attempt to write the conditional distribution of $\boldsymbol{\xi}(\tilde{\boldsymbol{s}})$ given the augmented data $\mathcal{D}_a$ that comprises training data $\mathcal{D}_s$, augmented by test data $v^{(test)}$. Here we consider the test data $v^{(test)}$ realised at $\boldsymbol{S} = \tilde{\boldsymbol{s}}$, where we use different candidates for $\tilde{\boldsymbol{s}}$. In particular, we choose $\tilde{\boldsymbol{s}}$ to be (1) the median $\boldsymbol{s}^{(median)}$ of the posterior of $\boldsymbol{S}$ given $\mathcal{D}_a$, (2) the mode $\boldsymbol{s}^{(mode)}$ of this posterior, (3) or $\boldsymbol{s}^{(u)}$, u=1,2,3,4–the end points of the disjoint 95% HPD region of the posterior of $\boldsymbol{S}$ (see Table 1).

Since $\{\boldsymbol{\xi}(\boldsymbol{s}_1), \ldots, \boldsymbol{\xi}(\boldsymbol{s}_n), \boldsymbol{\xi}(\tilde{\boldsymbol{s}})\}$ is jointly matrix-normal, $[\boldsymbol{\xi}(\tilde{\boldsymbol{s}})|\boldsymbol{\xi}(\boldsymbol{s}_1), \ldots, \boldsymbol{\xi}(\boldsymbol{s}_n))] \equiv [\boldsymbol{\xi}(\tilde{\boldsymbol{s}})|\mathcal{D}_s]$, is jk-variate normal. The mean function of this multivariate normal, at different $\tilde{\boldsymbol{s}}$, is then compared to the test data. Thus, the estimate of the function that we seek is $\mathbb{E}[\boldsymbol{\xi}(\boldsymbol{S})|\mathcal{D}_s, \boldsymbol{S}, \boldsymbol{Q}]$, given the dependence of $\boldsymbol{\xi}(\cdot)$ on the smoothness parameters (elements of $\boldsymbol{Q}$) that we anticipate.

However, we only know the conditional of $\boldsymbol{\xi}(\cdot)$ on all the $\mathcal{GP}$ parameters, including the ones that we do not learn from the data, namely $\boldsymbol{B}$ and $\boldsymbol{C}$. So we need to marginalise $[\boldsymbol{\xi}(\cdot) \mid \boldsymbol{\Sigma}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s]$ over $\boldsymbol{B}$ and $\boldsymbol{C}$. To achieve this, we need to invoke the conditional distribution of $\boldsymbol{B}$ and $\boldsymbol{C}$ with respect to the other $\mathcal{GP}$ parameters and $\mathcal{D}_s$. We recall the priors on the $\mathcal{GP}$ parameters $\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}$ (from Section 2.4) to write $\pi(\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}) \propto \mid \boldsymbol{\Sigma} \mid^{-(k+1)/2} \mid \boldsymbol{C} \mid^{-(j+1)/2}$. It then follows that

$$[\boldsymbol{B} \mid \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s] \sim \mathcal{N}_{m,jk}(\hat{\boldsymbol{B}}_{GLS}, (\boldsymbol{H}_D^T \boldsymbol{A}_D^{-1} \boldsymbol{H}_D)^{-1}, \boldsymbol{\Omega}), \tag{5.1}$$

where, we recall from Section 2.1 that we had set $m = d + 1$, with $\boldsymbol{S} \in \mathbb{R}^d$. Here, $\hat{\boldsymbol{B}}_{GLS} = (\boldsymbol{H}_D^T \boldsymbol{A}_D^{-1} \boldsymbol{H}_D)^{-1}(\boldsymbol{H}_D^T \boldsymbol{A}_D^{-1} \mathcal{D}_s)$. Marginalising the jk-variate normal that is the conditional $[\boldsymbol{\xi}(\cdot) \mid \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s]$ over $\boldsymbol{B}$ (using Equation 5.1), it can be shown that

$$[\boldsymbol{\xi}(\cdot) \mid \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s] \sim \mathcal{N}_{jk}(\boldsymbol{\mu}_2(\cdot), a_2(\cdot, \cdot)\boldsymbol{\Omega}), \tag{5.2}$$

where

$$\begin{aligned}\boldsymbol{\mu}_2(\cdot) &= \hat{\boldsymbol{B}}_{GLS}^T\boldsymbol{h}(\cdot) + (\mathcal{D}_s - \boldsymbol{H}_D\hat{\boldsymbol{B}}_{GLS})^T\boldsymbol{A}_D^{-1}\boldsymbol{\sigma}_D(\cdot); \quad (5.3)\\ a_2(\boldsymbol{s}_1,\boldsymbol{s}_2) &= a_1(\boldsymbol{s}_1,\boldsymbol{s}_2) + [\boldsymbol{h}(\boldsymbol{s}_1) - \boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{s}_D(\boldsymbol{s}_1)]^T(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1}\\ &\quad [\boldsymbol{h}(\boldsymbol{s}_2) - \boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{s}_D(\boldsymbol{s}_2)]. \quad (5.4)\end{aligned}$$

We define $(n-m)\hat{\boldsymbol{\Omega}}_{GLS} = (\mathcal{D}_s - \boldsymbol{H}_D\hat{\boldsymbol{B}}_{GLS})^T\boldsymbol{A}_D^{-1}(\mathcal{D}_s - \boldsymbol{H}_D\hat{\boldsymbol{B}}_{GLS})$, i.e. $(n-m)\hat{\boldsymbol{\Omega}}_{GLS} = \mathcal{D}_s^T\boldsymbol{M}\mathcal{D}_s$, with $\boldsymbol{M} = \boldsymbol{A}_D^{-1} - \boldsymbol{A}_D^{-1}\boldsymbol{H}_D(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1}\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}$).

We consider the mean $\boldsymbol{\mu}_2(\cdot)$ of the conditional posterior given by (5.3) as a suitable estimator of the velocity function in our case. Note that $\boldsymbol{\mu}_2$ involves the unknown smoothness parameters; we plug-in the corresponding posterior medians $0.874254, 1.003545$ for these.

It is important to mention that though the mean and variance in Equations 5.3 and Equation 5.4 were developed using $\mathcal{D}_s$, in our construction of the velocity function estimator $\boldsymbol{\mu}_2$, $\mathcal{D}_a$ is implemented, where $\mathcal{D}_a$ is obtained by augmenting $\mathcal{D}_s$ with $\boldsymbol{v}^{(test)}$ that is realised at $\boldsymbol{S} = \tilde{\boldsymbol{s}}$. The underlying theory remains the same as above.

It is important to note that $\boldsymbol{\mu}_2(\boldsymbol{S})$, where $\boldsymbol{S}$ is the unknown location, is a random variable, and even though the posterior of $\boldsymbol{\Sigma}$ is concentrated around the null matrix, the variance of $\boldsymbol{\mu}_2(\boldsymbol{S})$ is not $\boldsymbol{0}$, thanks to the fact that $\boldsymbol{S}$ does not have $\boldsymbol{0}$ variance. Consequently, the posterior variance of $\boldsymbol{\xi}(\boldsymbol{S})$ does not have $\boldsymbol{0}$ variance. To see this formally, note that

$$\begin{aligned}Var\left[\boldsymbol{\xi}(\boldsymbol{S})|\mathcal{D}_a\right] &= Var\left[\mathbb{E}\left\{\boldsymbol{\xi}(\boldsymbol{S})|\boldsymbol{\Sigma},\boldsymbol{C},\boldsymbol{Q},\boldsymbol{S},\mathcal{D}_a\right\}\right] + \mathbb{E}\left[Var\left\{\boldsymbol{\xi}(\boldsymbol{S})|\boldsymbol{\Sigma},\boldsymbol{C},\boldsymbol{Q},\boldsymbol{S},\mathcal{D}_a\right\}\right]\\ &= Var\left[\boldsymbol{\mu}_2(\boldsymbol{S})|\mathcal{D}_a\right] + \mathbb{E}\left[a_2(\boldsymbol{S},\boldsymbol{S})\boldsymbol{\Omega}|\mathcal{D}_a\right]. \quad (5.5)\end{aligned}$$

Since the posterior $[\boldsymbol{\Sigma}|\mathcal{D}_a]$ is concentrated around the $k \times k$-dimensional null matrix, it follows that the posterior $[\boldsymbol{\Omega}|\mathcal{D}_a]$ is also concentrated around the $jk \times jk$-dimensional null matrix. Hence, in (5.5), $\mathbb{E}\left[a_2(\boldsymbol{S},\boldsymbol{S})\boldsymbol{\Omega}|\mathcal{D}_a\right] \approx \boldsymbol{0}^{(jk\times jk)}$. However, the first part of (5.5), $Var\left[\boldsymbol{\mu}_2(\boldsymbol{S})|\mathcal{D}_a\right]$, is strictly (and significantly) positive, showing that the variance of the posterior of $\boldsymbol{\xi}(\boldsymbol{S})$ is significantly positive.

The above result shows that it should not be expected that the observed test velocity data $\boldsymbol{v}^{(test)}$ will be predicted accurately by $\boldsymbol{\mu}_2(\boldsymbol{s})$, for any given $\boldsymbol{s}$. This is in contrast with the usual Gaussian process emulators, where the argument of the unknown function is non-random, so that if the posterior of the function variance is concentrated around $\boldsymbol{0}$, then the posterior variance of the emulator would be close to $\boldsymbol{0}$.

In Figure 5 we illustrate, in the case of $sp3bar3$ (the most chaotic model), the degree of agreement of $\boldsymbol{\mu}_2(\boldsymbol{s})$ with $\boldsymbol{v}^{(test)}$ for different choices of $\boldsymbol{s}$. We compare with $\boldsymbol{v}^{(test)}$ the predictions $\boldsymbol{\mu}_2(\boldsymbol{s}^{(mode)})$, $\boldsymbol{\mu}_2(\tilde{\boldsymbol{s}})$ and $\boldsymbol{\mu}_2(\boldsymbol{s}^{(u)})$; $u = 1,2,3,4$, Here , $\boldsymbol{s}^{(mode)} = (1.73, 18.8°)$ is the (component-wise) posterior mode and $\tilde{\boldsymbol{s}} = (2.2, 35°)$ is a point somewhat close to the (component-wise) posterior median $\boldsymbol{s}^{(median)} = (1.994478, 33.59429°)$ (grid-point closest to $\boldsymbol{s}^{(median)}$.

As observed in Figure 5 the best fit of $\boldsymbol{v}^{(test)}$ has been provided by $\boldsymbol{\mu}_2(\tilde{\boldsymbol{s}})$ where $\tilde{\boldsymbol{s}}$ is close to the median $\boldsymbol{s}^{(median)}$; as the point $(\boldsymbol{s}^{(median)}, \boldsymbol{v}^{(test)})$ is in the training data constituting $\boldsymbol{\mu}_2$, this is to be expected. The estimators $\boldsymbol{\mu}_2(\boldsymbol{s}^{(mode)})$ and $\boldsymbol{\mu}_2(\boldsymbol{s}^{(1)})$ perform somewhat reasonably, but the remaining estimators $\boldsymbol{\mu}_2(\boldsymbol{s}^{(u)})$; $u = 2,3,4$ do not perform adequately, signifying the effect of variablity of our estimator due the posterior of $\boldsymbol{S}$.

While it is the randomness of the argument $\boldsymbol{S}$ of the unknown function $\boldsymbol{\xi}(\cdot)$ that causes the variability of our estimator, such variability is highest in the most chaotic of the 4 base astrophysical models ($sp3bar3$), and least in the only non-chaotic base astrophysical model (bar_6). A similar exercise of predicting $v^{(test)}$ using the training data simulated from this non-chaotic base model gives excellent fits at all the aforementioned used values of $\boldsymbol{S}$; see Figure 6.

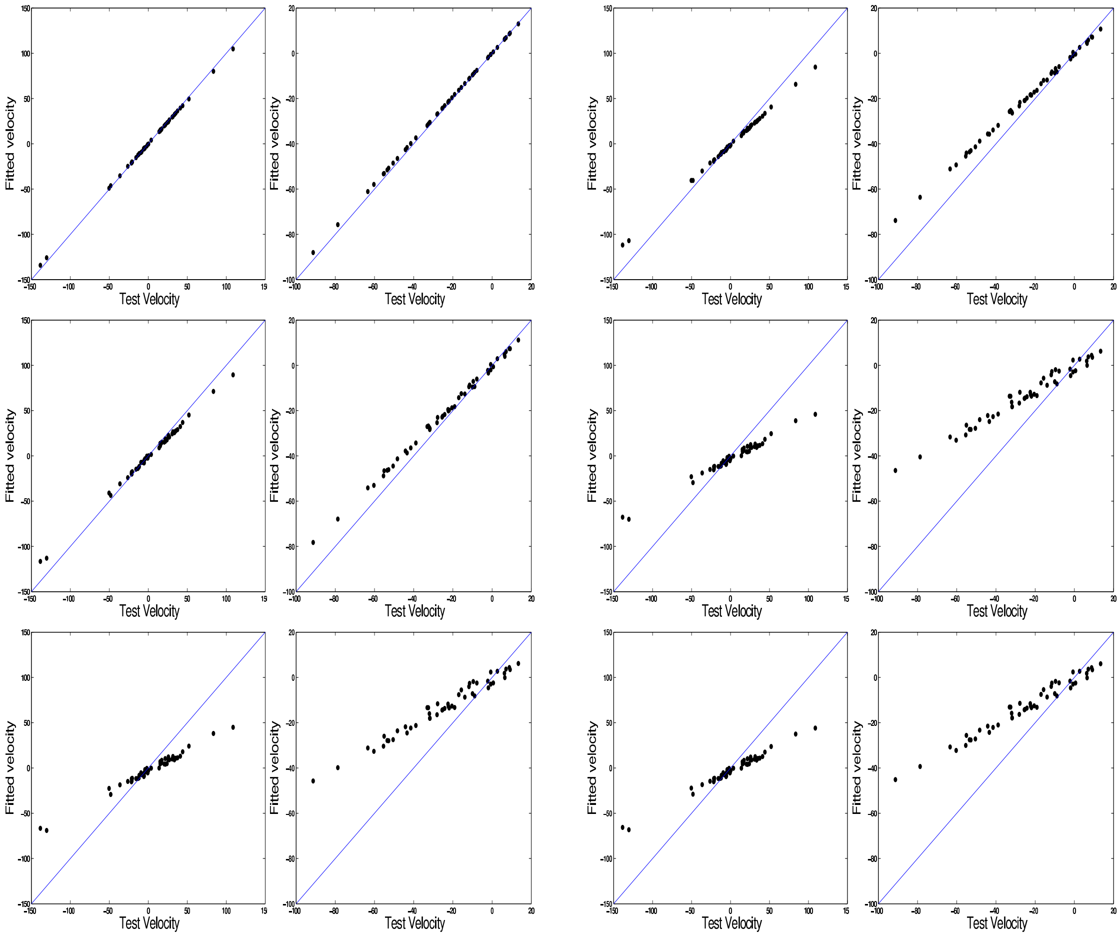


FIG 5. *Prediction of $\boldsymbol{v}^{(test)}$ for model $sp3bar3$: plots of 2 components of $\boldsymbol{\mu}_2(\boldsymbol{s})$ against $\boldsymbol{v}^{(test)}$ for $\boldsymbol{s} = \tilde{\boldsymbol{s}}$ (2 left hand sided panels on the top row), $bs = \boldsymbol{s}^{(mode)}$ (2 right panels on the top), $\boldsymbol{s} = \boldsymbol{s}^{(1)}$ (2 left panels in the middle row), $\boldsymbol{s} = \boldsymbol{s}^{(2)}$ (2 left panels in the middle row), $\boldsymbol{s} = \boldsymbol{s}^{(3)}$ (2 left panels in the lowest row), $\boldsymbol{s} = \boldsymbol{s}^{(4)}$ (2 right panels in the lowest row).*

6. Discussions

Computational complexity scales only linearly with the dimensionality of the unknown model parameter $\boldsymbol{S}$. Thus, porting a training data comprised of n independent values of $\boldsymbol{S}$, $\boldsymbol{s}_i$, $i = 1, \ldots, n$, where $\boldsymbol{s}_i$ is a d-dimensional vector, $d > 2$, is not going to render the computational times infeasible. This allows for the learning of high-dimensional model parameter vectors in our method.

In contrast to the situation with increasing the dimensionality of the unknown model parameter, increasing the dimensionality of the measurable will but imply substantial increase in the run time, since the relevant computational complexity then scales non-linearly, as about $O(k^3)$, (in addition to the cost of k square roots), where k is the dimensionality of the observed variable. This is because of the dimensionality of the aforementioned $\boldsymbol{\Sigma}$ matrix is $k \times k$, and the inverse of this enters the computation of the posterior via the definition $\hat{\boldsymbol{C}}_{GLS,aug}$. Thus, for example, increasing the dimensions of the measurable from 2 to 4 increases the run time 8-fold, which is a large jump in the required run time. However, for most applications, we envisage the expansion of the dimensionality of the unknown model parameter, i.e. d, rather than that of the measurable, i.e. k. Thus, the method is expected to yield results within acceptable time frames, for most practical applications.

The other major benefit of our work is that it allows for organic learning of the smoothness parameters, rather than results being subject to *ad hoc* choices of the same.

As more Galactic simulations spanning a greater range of model parameters become available, the rigorous learning of such Milky Way parameters using our method will become possible, given the available stellar velocity data. This will enhance the quality of our knowledge about our own galaxy. That our method allows for such learning even for under-abundant systems, is encouraging for application of a similar analysis to galaxies other than our own, in which system parameters may be learnt using the much smaller available velocity data sets, compared to the situation in our galaxy.

Supplementary material

Some background details on the application to the Milky Way are discussed in Section **S-1** of the attached supplementary material. Section **S-2** discusses the details of the dynamical simulations that lead to the training data set used in our supervised learning of the Milky Way feature parameters. In Section **S-3** we present details of the TMCMC methodology that we use here. **S-4** discusses the cross-validation of our model and methodology, on simulated as well the real stellar velocity data. The effect of chaos on the modality of the posterior distributions of our unknowns is discussed in Section **S-5**.

References

A. O'Hagan (1978), "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society B*, 40, 1–42.

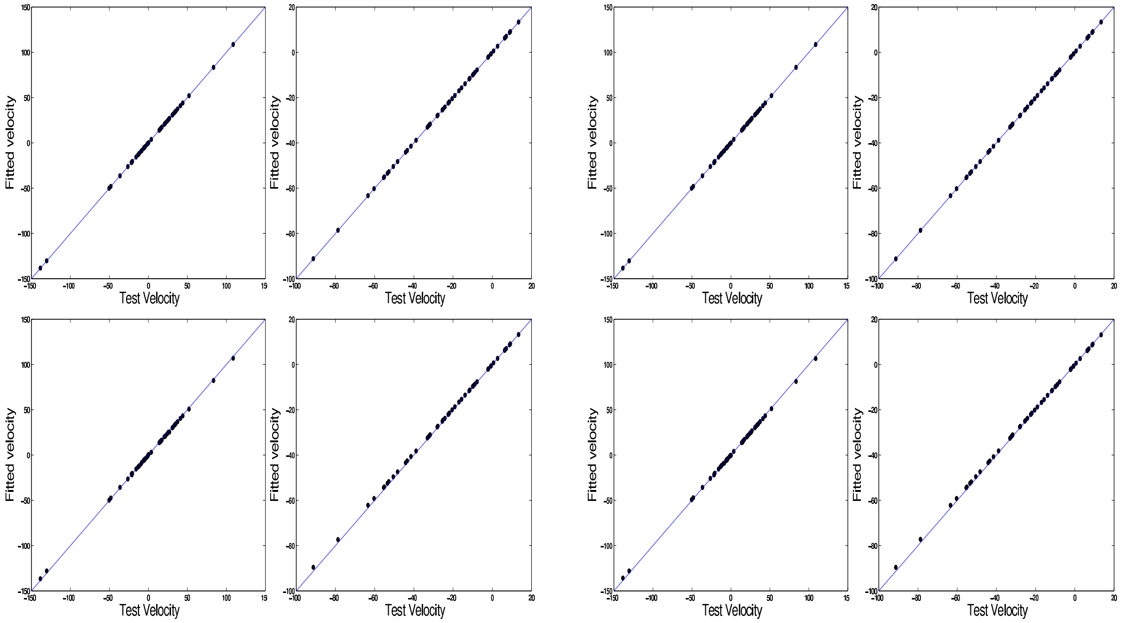


FIG 6. *Prediction of $\boldsymbol{v}^{(test)}$ for model $bar6$: plots of 2 components of $\boldsymbol{\mu}_2(\boldsymbol{s})$ against $\boldsymbol{v}^{(test)}$ for $\boldsymbol{s} = \tilde{\boldsymbol{s}}$ (2 adjacent panels on the left hand side of the top row), $\boldsymbol{s}^{(mode)}$ (2 panels on the right of top row), $\boldsymbol{s}^{(1)}$ (2 panels on the left in the lower row), $\boldsymbol{s}^{(2)}$ (2 panels on the left in the lower row).*

Antoja, T., Figueras, F., Romero-Gómez, M., Pichardo, B., Valenzuela, O., and Moreno, E. (2011), "Understanding the spiral structure of the Milky Way using the local kinematic groups," *Monthly Notices of the Royal Astronomical Society*, 418, 1423–1440.

Antoja, T., Valenzuela, O., Pichardo, B., Moreno, E., Figueras, F., and Fernández, D. (2009), "Stellar Kinematic Constraints on Galactic Structure Models Revisited: Bar and Spiral Arm Resonances," *Astrophysical Jl. Letters*, 700, L78–L82.

Aumer, M., and Binney, J. J. (2009), "Kinematics and history of the solar neighbourhood revisited," *Monthly Notices of the Royal Astronomical Society*, 397, 1286–1301.

Benjamin, R. A., Churchwell, E., Babler, B. L., Indebetouw, R., Meade, M. R., Whitney, B. A., Watson, C., Wolfire, M. G., Wolff, M. J., Ignace, R., Bania, T. M., Bracker, S., Clemens, D. P., Chomiuk, L., Cohen, M., Dickey, J. M., Jackson, J. M., Kobulnickyand E. P. Mercer, H. A., Mathis, J. S., Stolovy, S. R., and Uzpen, B. (2005), "First GLIMPSE Results on the Stellar Structure of the Galaxy," *Astrophysical Jl. Letters*, 630, L149–L152.

B.Hofmann (2011), Ill-posedness and regularization of inverse problemsa review on mathematical methods,, in *The Inverse Problem. Symposium ad Memoriam H. v. Helmholtz, H. Lubbig (Ed). Akademie-Verlag, Berlin; VCH, Weinheim*, pp. 45–66.

Binney, J., and Merrifield, M. (1998), *Galactic Astronomy*, Princeton: Princeton University Press.

Blight, B. J. N., and Ott, L. (1975), "A Bayesian Approach to Model Inadequacy for Polynomial Regression," *Biometrika*, 62(1), 79–88.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall. Second Edition.

Carreira-Perpin, M. A. (2001), Continuous latent variable models for dimensionality reduction and sequential data reconstruction, Doctoral thesis, University of Sheffield.

Carvalho, C. M., and West, M. (2007), "Dynamic Matrix-Variate Graphical Models," *Bayesian Analysis*, 2, 69–98.

Chakrabarty, D. (2007), "Phase Space around the Solar Neighbourhood," *Astronomy & Astrophysics*, 467, 145.

Chakrabarty, D., and Sideris, I. (2008), "Chaos in Models of the Solar Neighbourhood," *Astronomy & Astrophysics*, 488, 161.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, New York: Wiley.

Dawid, A. P. (1981), "Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 265–274.

Debattista, V. P., Gerhard, O., and Sevenster, M. N. (2002), "The pattern speed of the OH/IR stars in the Milky Way," *Monthly Notice of Royal Astronomical Society*, 334, 355.

Dehnen, W. (2000), "The Effect of the Outer Lindblad Resonance of the Galactic Bar on the Local Stellar Velocity Distribution," *Astronomical Journal*, 11, 800.

Dutta, S., and Bhattacharya, S. (2013), "Markov Chain Monte Carlo Based on Deterministic Transformations,". Accepted in Statistical Methods; available at arxiv:1106.5850v3 with supplementary section in arxiv.org/pdf/1306.6684.

Englmaier, P., and Gerhard, O. (1999), "Gas dynamics and large-scale morphology of the Milky Way galaxy," *Monthly Notices of the Royal Astronomical Society*, 304, 512–534.

Fux, R. (1997), "3D self-consistent N-body barred models of the Milky Way. I. Stellar dynamics," *Astronomy & Astrophysics*, 327, 983–1003.

Fux, R. (2001), "Order and chaos in the local disc stellar kinematics induced by the Galactic bar," *Astronomy & Astrophysics*, 373, 511–535.

G. Walker and J. Ford (1969), “Amplitude Instability and Ergodic Behavior for Conservative Nonlinear Oscillator Systems,” *Physical Review*, 188, 416–432.

Golubov, O. (2012), Modelling the Milky Way Disk, Doctoral thesis, University of Heidelberg.

Hoff, P. D. (2011), “Hierarchical multilinear models for multiway data,” *Computational Statistics & Data Analysis*, 55, 530–543.

Johnston, K., Bullock, J. S., and Strauss, M. (2009), The Milky Way and Local Volume as Rosetta Stones in Galaxy Formation,, in *astro2010: The Astronomy and Astrophysics Decadal Survey*, Vol. 2010, p. 142.

Kabanikhin, S. I. (2008), “Definitions and examples of inverse and ill-posed problems,” *J. Inv. Ill-Posed Problems*, 16, 317–357.

Kucinskas, A., Lindegren, L., and Vansevicius, V. (2005), Beyond the Galaxy with Gaia: Evolutionary Histories of Galaxies in the Local Group,, in *The Three-Dimensional Universe with Gaia*, eds. C. Turon, K. S. O’Flaherty, and M. A. C. Perryman, Vol. 576 of *ESA Special Publication*.

Lindegren, L., Babusiaux, C., Bailer-Jones, C., Bastian, U., Brown, A., Cropper, M., Hg, E., Jordi, C., Katz, D., van Leeuwen, F., Luri, X., Mignard, F., de Bruijne, J., and Prusti, T. (2007), The Gaia mission: science, organization and present status,, in *A Giant Step: from Milli- to Micro-arcsecond Astrometry, Proceedings IAU Symposium No. 248*, eds. W. Jin, I. Platais, and M. Perryman, pp. 217–223.

Matern, B. (1986), *Spatial Variation (2nd ed.)*, Springer: Springer-Verlag.

Minchev, I., Boily, C., Siebert, A., and Bienayme, O. (2010), “Low-velocity streams in the solar neighbourhood caused by the Galactic bar,” *Monthly Notices of the Royal Astronomical Society*, 407, 2122–2130.

Minchev, I., Quillen, A. C., Williams, M., Freeman, K. C., Nordhaus, J., Siebert, A., and Bienaymé, O. (2009), “Is the Milky Way ringing? The hunt for high-velocity streams,” *Monthly Notices of the Royal Astronomical Society*, 396, L56–L60.

Neal, R. M. (1998), Regression and classification using Gaussian process priors (with discussion),, in *Bayesian Statistics 6*, ed. J. M. B. et. al, Oxford University Press, pp. 475–501.

Perryman, M. (2012), *Astronomical Applications of Astrometry: Ten Years of Exploitation of the Hipparcos Satellite Data*, Cambridge: Cambridge University Press.

Rasmussen, C. E., and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, MIT: The MIT Press.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The design and analysis of computer experiments*, Springer Series in Statistics, New York, Inc.: Springer-Verlag.

Scholkopf, B., and Smola, A. J. (2002), *Learning with Kernels*, MIT: MIT Press.

Sengupta, A. (2003), “Toward a Theory of Chaos,” *International Journal of Bifurcation and Chaos*, 13, 3147–3233.

Simone, R. D., Wu, X., and Tremaine, S. (2004), “The stellar velocity distribution in the solar neighbourhood,” *Monthly Notices of the Royal Astronomical Society*, 350, 627.

Snelson, E. L. (2007), Flexible and efficient Gaussian process models for machine learning, Doctoral thesis, University of London.

Stuart, A. (2013), “Bayesian Approach to Inverse Problems,”. provide an introduction to the forthcoming book Bayesian Inverse Problems in Differential Equations by M. Dashti, M. Hairer and A.M. Stuart; available at arXiv:math/1302.6989.

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Model Parameter Estimation*, Philadelphia: SIAM.

Tilmann Gneiting and William Kleiber and Martin Schlather (2010), “Matern Cross-Covariance Functions for Multivariate Random Fields,” *Journal of te American Statistical Association*, 105(491), 1167–1177.

arXiv: math.PR/0000000

Supplementary section for "Bayesian Nonparametric Estimation of Milky Way Parameters Using Matrix-Variate Data, in a New Gaussian Process Based Method"

Dalia Chakrabarty[*,§]**,, Munmun Biswas**[†,¶]**, Sourabh Bhattacharya**[‡,¶] **,**

[§] *Department of Statistics*
University of Warwick
Coventry CV4 7AL, U.K.
d.chakrabarty@warwick.ac.uk
and
Department of Mathematics
University of Leicester
Leicester LE1 7RH, U.K.
dc252@le.ac.uk

[¶] *Indian Statistical Institute*
203, B. T. Road
Kolkata 700108, India
munmun.biswas08@gmail.com sourabh@isical.ac.in

Throughout, we refer to our main manuscript as CBB.

1. Background of the Application

As indicated in Figure 1, we approximate the geometry of the Milky Way disc as a 2-dimensional disc. Thus, we confine analysis to such a two-dimensional spatial geometry, rendering the location

[*] Associate Research fellow at Department of Statistics, University of Warwick and Lecturer of Statistics at Department of Mathematics, University of Leicester

[†] PhD student in Statistics and Mathematics Unit, Indian Statistical Institute

[‡] Assistant Professor in Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute

vector of the a-th star as $(X_1^{(a)}, X_2^{(a)})^T$, and the velocity vector as $(U^{(a)}, V^{(a)})^T$. Let such locations and velocities of j stars be measured. It is to be emphasised that $(X_1^{(a)}, X_2^{(a)})^T$ is the measurement taken from the Sun (i.e. from or location in he Galaxy), i.e. $(X_1^{(a)}, X_2^{(a)})^T$ is the heliocentric location of the a-th star. Thus, if the Solar location with respect to the Galactic centre is $(r_\odot, \phi_\odot)^T$, then the value of the Galactocentric location of the a-th star is $(r_\odot, \phi_\odot)^T + (x_1^{(a)}, x_2^{(a)})^T$. To put this in the context of the lower panel in Figure 1, the general location vector to a star at point C inside the sampled region centred at S, is along line-segment OC and is given by the sum of the location vectors along OS and SC. However, the Galactocentric stellar location is unknown as is the Solar location $(r_\odot, \phi_\odot)^T$. Thus, using the measured value $(x_1^{(a)}, x_2^{(a)})^T$, the unknown $(r_\odot, \phi_\odot)^T$ cannot be estimated. Furthermore, the spatial locations of the sampled j stars lie inside a circle with radius ϵ centred at the Sun (Fux, 2001), and are assumed to be distributed uniformly within this circle. Then the summary of the distribution of the measured locations $\{(x_1^{(a)}, x_2^{(a)})^T\}_{a=1}^{j}$ will always coincide with the centre of this circle - which is the Sun - irrespective of what the galactocentric location of this centre is. Thus, these measured spatial locations cannot constrain the sought galactocentric location of the Sun.

Similarly, the recorded values of stellar velocities, $(u^{(a)}, v^{(a)})^T$, $a = 1, \ldots, j$, are as measured from the Sun and are therefore with respect to the solar velocity. These measured heliocentric stellar velocities are however affected by the choice of the location of the observer, i.e. the location of the Sun, i.e. $(r_\odot, \phi_\odot)^T$. This is because, a given star, if observed from different spatial locations in the Galaxy, would appear to move in different ways. For example, as indicated in Figure 1, if a star appears to have a velocity vector directed along the line that joins itself to the observer at point A on the Milky Way disc, this observer will register its velocity to be entirely radial, with zero angular component of the velocity vector. Here “radial” component of the velocity vector is the component along the line-of-sight joining the observer to the star and the component orthogonal to the line-of-sight is referred to as the “angular” component. On the contrary, had the observer been at a different point B, the velocity vector of this star would have registered to have had a radial as well as an angular component, in general. Thus, the observed stellar velocities will bear information about the location of the observer, i.e. the Sun. Then the available velocity data $\mathbf{V}$ can be considered to bear the signature of the unknown $\boldsymbol{S}$. In principle, beyond just the galactocentric solar location, if there are Milky Way feature parameters that physically affect stellar motions, observed velocity data will

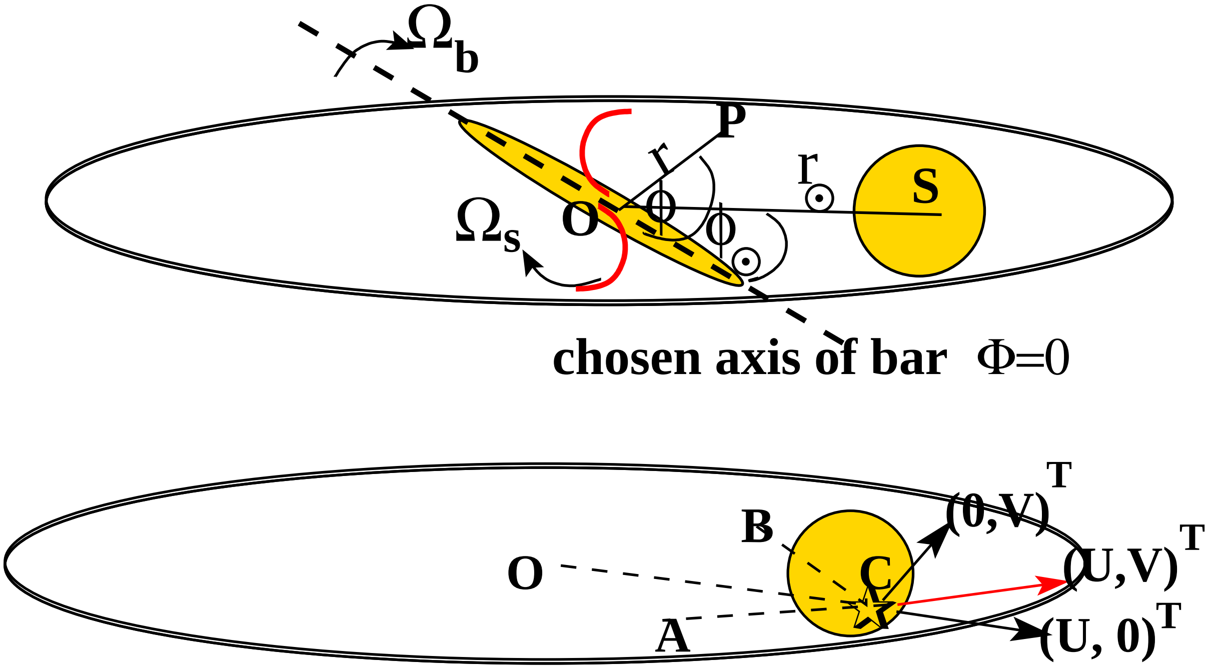


FIG 1. *Top: schematic diagram of the Milky Way disk centred at point marked O with the Sun at the point marked S. The long axis of the central rotating stellar bar (marked in broken lines) is chosen to define $\Phi = 0$, where Φ is the variable that measures the angular separation of a point on the disk from this axis. The variable R measures the distance of a point on the disk from the Galactic centre O. Thus, at a general point P, $R = r$ and $\Phi = \phi$. At the Sun at S, $R = r_\odot$, $\Phi = \phi_\odot$. The 2-dimensional velocity vectors of a sample of stars lying within a circle centred at S, are observed by us from Earth, i.e. from the Sun. The Galactic spiral arms are schematically shown in red. The bar and spiral rotate with angular speeds of Ω_b and Ω_s respectively. Bottom: one of the sampled stars (marked by the $\star$ symbol within the circle centred at the Sun) at point C, would appear to have velocity $(u, v)^T$ with respect to the Galactic centre at O. This means that an observer at O would register a "radial" component $(u, 0)^T$ along her line-of-sight to the star, i.e. along the line OC, and a "angular" component $(0, v)^T$ orthogonal to the line-of-sight. However, if another observer at the point A viewed this star - where the line segment AC lies along vector $(u, v)^T$ - she will regard the projection of the $(u, v)^T$ vector onto a line orthogonal to the line segment AC, to be zero, i.e. will view the velocity of the star to be entirely along her line of sight. The radial component (component along line of sight) of this star's velocity according to her would be $\sqrt{u^2 + v^2}$ while the angular component (orthogonal to line of sight) is 0, so that she will observe this star to move with velocity $(\sqrt{u^2 + v^2}, 0)^T$. Another viewer at point B will however infer different values of the radial and transverse components of the velocity of this star, given the orientation of the location B with respect to the vector $(u, v)^T$.*

bear the signature of such Milky Way parameters.

2. Details of dynamical simulations of astrophysical models

In Chakrabarty (2007), the simulations involve the following. A sample of stellar 2-dimensional location and 2-dimensional velocity coordinates $\{r^{(a)}, \phi^{(a)}, u^{(a)}, v^{(a)}\}_{a=1}^{j}$, is drawn from a chosen (to mimic real disc galaxies') density function $g(R, \Phi, U, V)$ at $T = 0$, and is evolved in a (chosen) parametric Galactic gravitational potential $\Psi(R, \Phi, T)$ where we recall that the strength and shape of the gravitational influence of the system is given by the gravitational potential. Here T is time variable. In fact, the gravitational influence of the system is modelled as mostly due to the Milky Way disk, but perturbed by the gravitational potential of the central bar in the Galaxy (see Fig 1 above) as well as that of the Galaxy's spiral arms. The potential of the disk is assumed to be stationary and chosen to emulate a realistic, time-independent background Galactic potential $\Psi_0(R,\Phi)$. The contribution of the potential of the rotating (and therefore time-dependent) Galactic bar is $\varepsilon_b(T)\Psi_b(R,\Phi,T)$ where the scalar $\varepsilon_b(T)$ represents the strength of the disturbance that the bar imposes on the disk's gravitational influence at time T. Again, $\varepsilon_b(T)$ is chosen to emulate the growth of the bar inside the Galaxy. We recall that the bar is chosen to rotate about the centre of the disk at a rate of Ω_b. Similarly, the potential of the rotating spiral pattern is $\varepsilon_s(T)\Psi_s(R,\Phi,T)$. Again, Ω_s defines the rotation rate of the spiral pattern. Thus at any time and at any location on the disk, the net gravitational potential in the q-th base astronomical model is $\Psi_0^{(q)}(R,\Phi)+\varepsilon_b(T)\Psi_b^{(q)}(R,\Phi,T)+\varepsilon_s(T)\Psi_s^{(q)}(R,\Phi,T)$, for each $q=1,\ldots,4$. The sampled stellar location and velocity coordinates are made to evolve using Newtonian equations of motion under the influence of the aforementioned net gravitational potential. At the end of a chosen period of time, when $T=t_{sim}$, evolved orbits are sampled and recorded in the rotating frame of the bar at times when $(\Omega_b-\Omega_s)t$=0.

The relevant subset of the space of the design vectors (i.e. chosen solar location vector) is discretised and the recorded orbits are sorted by their final locations into the discretised bins; thus stars with final locations in the neighbourhood of the centroid of the i-th discretised bin were slotted into the i-th bin. The interpretation of this is that stars in the i-th bin share a similar galactocentric location $\boldsymbol{s}_i^\star$, and their j number of k-dimensional velocity vectors comprise the i-th synthetic velocity data set, (which we treat as the jk-dimensional vector). Here $i=1,2,\ldots,n$ and Chakrabarty (2007) used n=216. The n synthetic velocity vectors, each thus generated at the n grid points, form the training data set $\mathcal{D}_s$. The grid that the design vectors are grid points of is defined by the ranges of $r\in[1.7,2.3]$ in model units and $\phi\in[0,90]$ in degrees. The same 2-D grid is used for each of the base astrophysical models.

In fact, in any base astrophysical model of the Milky Way, all distances are in units of the "co-rotation radius" r_{CR} of the central bar in the Milky Way disc. This is the radius at which the rotational rate Ω_b of the rotating bar, equals the radius-dependent rotational rate $\Omega(R)$ of the stars at distance R from the centre of the Galaxy, which in turn is determined by the choice of inputs in the base astrophysical models. In all 4 of the base astrophysical models, this stellar rotational rate $\Omega(R)$ is defined as $\dfrac{v_0}{R}$, where v_0 is a constant that is set to unity in the base astrophysical models by Chakrabarty (2007); Ω_b is also set to unity. Then co-rotation occurs when $\Omega(R) = \Omega_b$, i.e. $\dfrac{v_0}{r_{CR}} = \Omega_b$ which implies that $r_{CR} = 1$ in each of the 4 base astrophysical models.

To connect any distance in these base models to a physically realised distance measured in units of kilo parsec–or kpc–one needs to

- scale the constant v_0 to its real astronomical value of 220 kms^{-1} (Binney and Merrifield, 1998) and
- scale the bar rotational rate Ω_b to its real astronomical value so that $r_{CR} = \dfrac{v_0}{\Omega_b}$ is computed in real physical units. However it is the value of Ω_b in astronomical units that remains elusive and is sought. So, we
- scale our estimate of the solar radial location to the Galactic centre, $\hat{r_\odot}$, to the distance $\mathcal{R}$ measured in kpc, as cited in astronomical literature (obtained using ancillary information in independent astronomical modelling). This gives the scaling between model units and astronomical units (kpc) so that a distance of 1 in model units then follows as $\dfrac{\mathcal{R}}{\hat{r}_\odot}$kpc, i.e. r_{CR} in real units is $\dfrac{\mathcal{R}}{\hat{r}_\odot}$kpc. Independent astronomical studies have suggested $\mathcal{R}$=8kpc (Binney and Merrifield, 1998). We then use this real value of r_{CR} in $\dfrac{v_0}{r_{CR}} = \Omega_b$ to get an estimate of the sought Ω_b. Thus, $\Omega_b = \dfrac{v_0}{\frac{\mathcal{R}}{\hat{r}_\odot}}$. Using v_0=220kms^{-1} and $\mathcal{R}$=8kpc, we get $\Omega_b = \dfrac{220}{\frac{8}{\hat{r}_\odot}}$kms^{-1}/kpc. Learning the rotational rate Ω_b of the bar is the ulterior benefit of learning the solar radial location as in our approach.

3. TMCMC algorithm

Motivated by the fact that the performance of traditional MCMC methods - including the Metropolis-Hastings algorithm - can be less than satisfactory in high dimensions, both in terms of convergence and computational time, Dutta and Bhattacharya (2013) proposed the Transformation based MCMC or TMCMC. Dutta and Bhattacharya (2013) show that for additive transformations, the TMCMC-based acceptance rate decreases at a slower rate compared to block random walk Metropolis algorithms. Furthermore, TMCMC includes the hybrid Monte Carlo (HMC) algorithm as a special case and in one-dimensional situations while it boils down to the Metropolis-Hastings algorithm with a specialised proposal mechanism.

For our purpose, we shall consider TMCMC based on additive transformations, since Dutta and Bhattacharya (2013) show that these transformations require far less number of "move types" compared to non-additive transformations.

TMCMC allows updating the entire block $(\boldsymbol{s}^{(new)}, \boldsymbol{Q}, \boldsymbol{\Sigma})$ at the same time. The algorithm is as follows.

(i) Initialise the unknown quantities by fixing arbitrarily initial values $\left(\boldsymbol{s}^{(new,0)}, \boldsymbol{Q}^{(0)}, \boldsymbol{\Sigma}^{(0)}\right)$. In our case, $\boldsymbol{s}^{(new,0)} = (s_1^{(new,0)}, \ldots, s_d^{(new,0)})$, $\boldsymbol{Q}^{(0)}$ is characterised by the initial values of the d smoothness parameters, which we denote by $\boldsymbol{b} := (b_1^{(0)}, \ldots, b_d^{(0)})^T$ and $\boldsymbol{\Sigma}^{(0)}$ denotes the initial choice of the $k \times k$ matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is decomposed into $\boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is the appropriate lower-triangular matrix

(ii) Let $\boldsymbol{\varphi} = ((\boldsymbol{s}^{(new)})^T, \boldsymbol{b}^T, (\boldsymbol{L}^*)^T)^T$, where $\boldsymbol{L}^*$ denotes the column vector consisting of the non-zero elements of $\boldsymbol{L}$.

(iii) Next we propose $\epsilon \sim g(\cdot)I_{\{\epsilon>0\}}$, where $g(\cdot)$ is some arbitrary distribution, and I denotes the indicator function. In our applications, we shall choose $g(\cdot) = N(0,1)$, so that, $\epsilon > 0$ is drawn from a truncated normal distribution.

(iv) Assume that at iteration t, the state of the unknown parameters is $(\boldsymbol{s}^{(new,t)}, \boldsymbol{Q}^{(t)}, \boldsymbol{\Sigma}^{(t)}) := \boldsymbol{\varphi}^{(t)}$. Update $\boldsymbol{\varphi}^{(t)}$ by setting, with probabilities π_j and $(1-\pi_j)$, $\varphi_j^{(t+1)} = \varphi_j^{(t)} \pm c_j\epsilon$ (forward transformation) and $\boldsymbol{\varphi}_j^{(t+1)} = \boldsymbol{\varphi}_j^{(t)} - c_j\epsilon$ (backward transformation), respectively, where, for $j = 1, \ldots, d$, π_j are appropriately chosen probabilities and c_j are appropriately chosen scaling factors. Assume that for $j_1 \in \mathcal{U}$, $\boldsymbol{\varphi}_{j_1}^{(t)}$ gets the positive transformation, while for $j_2 \in \mathcal{U}^c$, $\boldsymbol{\varphi}_{j_2}^{(t)}$ gets the backward transformation. Here $\mathcal{U} \cup \mathcal{U}^c = \{1, \ldots, d^*\}$, where $d^* = 2d + \frac{k(k+1)}{2}$.

(v) We accept the new proposal $\boldsymbol{\varphi}^{(t+1)}$ with acceptance probability

$$\alpha_{\boldsymbol{\varphi}} = \min\left\{1, \frac{\prod_{j_1 \in \mathcal{U}}(1-\pi_{j_1})\prod_{j_2 \in \mathcal{U}^c}\pi_{j_2}}{\prod_{j_1 \in \mathcal{U}}\pi_{j_1}\prod_{j_2 \in \mathcal{U}^c}(1-\pi_{j_2})} \times r_{\boldsymbol{\varphi}}\right\} \tag{3.1}$$

where $r_{\boldsymbol{\varphi}}$ denotes the ratio of $\left[|\boldsymbol{A}_{D_{aug}}(\boldsymbol{s}_{aug})|^{-\frac{jk}{2}}|\boldsymbol{H}_{D_{aug}}^T(\boldsymbol{s}_{aug})\boldsymbol{A}_{D_{aug}}^{-1}(\boldsymbol{s}_{aug})\boldsymbol{H}_{D_{aug}}(\boldsymbol{s}_{aug})|^{-\frac{jk}{2}}\right] \times \left[|\boldsymbol{\Sigma}|^{-\frac{k(n+1-m)+k+1}{2}}|(n+1-m)k\hat{\boldsymbol{C}}_{GLS,aug}|^{-\frac{(n+1-m)k}{2}}\right]$, evaluated at the new value $(\boldsymbol{\varphi}^{(t+1)})$ and the current value $(\boldsymbol{\varphi}^{(t)})$ of $\boldsymbol{\varphi}$ respectively. We only need to bear in mind that the acceptance probability is zero if $b_j \leq 0$ for any j or if any diagonal element of $\boldsymbol{L}$ is negative.

For proper choices of the scale parameters of the additive transformation and the initial values of the parameters we conducted several initial "pilot" TMCMC runs of length around 100,000, starting with arbitrary initial values and guesses of the scale parameters such that all the runs converged to the same distribution as indicated by informal diagnostics such as trace plots. For the final TMCMC run, we chose those scale parameters that yielded the best convergence (with respect to empirical diagnostics such as trace plots) among the pilot runs, and selected the final values of the parameters obtained in this best pilot run as the initial values for the final run of TMCMC. The pilot runs yielded the proposal mechanism that we worked with.

4. Cross-validation

We employ leave-one-out cross-validation to assess the validity of our model and methodology. We leave out the i-th value $\boldsymbol{s}_i$ of the model parameter vector $\boldsymbol{S}$ and predict this $\boldsymbol{s}_i$ using the data that comprises motion vector $\mathbf{v}_i$, along with the remaining training data set from which $\mathbf{v}_i$ is omitted. Then for this prediction, the test data is really $\mathbf{v}_i$; to emphasise this form of the test data in notation similar to what we have used above, we denote the test data as $\mathbf{v}^{(test,i)}$ where $\mathbf{v}^{(test,i)} := \mathbf{v}_i$. The training data set relevant to this exercise is obtained by omitting the i-th row from $\mathcal{D}_s$, i.e. the training data $\mathcal{D}_s^{(-i)}$ is constructed as $\mathcal{D}_s$ bereft of the jk-dimensional motion vector $\mathbf{v}_i$. The aim is to compute the posterior probability density of $\boldsymbol{s}_i$, given the relevant test and training data sets. We perform such leave-one-out cross-validation for each $i = 1, 2, \ldots, n$. To perform inference with $\mathbf{v}_i$ omitted, a TMCMC run is required, implying that the full cross-validation would then demand n many TMCMC runs. Such is however computationally burdensome. Bhattacharya and Haslett (2007) have shown that the usual importance sampling/resampling methods suggested by Gelfand, Dey and Chang (1992) and Gelfand (1996), which may be effective in the case of forward problems, are not appropriate for inverse problems because of the technical intricacies of the latter. Bhattacharya and Haslett (2007) suggested a fast methodology for implementing cross-validation in inverse problems, by combining importance resampling (IR) and low-dimensional MCMC runs in an effective manner. We adopt this methodology, which the above authors termed IRMCMC, but replace the MCMC part with the more effective TMCMC methodology.

In the following we discuss the procedure for model validation.

1. Choose an initial i^* where $\pi(\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i^*)}, \mathbf{v}^{(test,i^*)})$ as the importance sampling density. Bhattacharya and Haslett (2007) demonstrate that an appropriate i^* may be obtained by minimising the following distance function with respect to i:
$$d(i) = \sum_{t=1}^{n} \left\{ \sum_{u=1}^{d} \frac{(\boldsymbol{s}_t^{(u)} - \boldsymbol{s}_i^{(u)})^2}{\nu_{s_u}^2} + \sum_{\ell=1}^{jk} \frac{(\mathbf{v}_t^{(\ell)} - \mathbf{v}_i^{\ell})^2}{\nu_{v_\ell}^2} \right\}, \tag{4.1}$$
where $\nu_{s_u}^2$ and $\nu_{v_u}^2$ are the data-based standard deviations corresponding to the u-th coordinate of $\boldsymbol{s}$ and $\mathbf{v}$, respectively.
2. From this importance sampling density, following Section 3, use TMCMC to sample $(\boldsymbol{s}^{(\ell)}, \boldsymbol{Q}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)})$, $\ell = 1, \ldots, N$ for large N.
3. For $i \in \{1, \ldots, i^* - 1, i^* + 1, \ldots, n\}$,
   a. for each sample value $(\boldsymbol{s}^{(\ell)}, \boldsymbol{Q}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)})$, compute importance weights $w_{i^*,i}^{(\ell)}$=$w_{i^*,i}(\boldsymbol{s}^{(\ell)}, \boldsymbol{Q}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)})$, where the importance weight function is given by
$$w_{i^*,i}(\boldsymbol{s}, \boldsymbol{Q}, \boldsymbol{\Sigma}) = \frac{L(\boldsymbol{s}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i)}, \mathbf{v}^{(test,i)})}{L(\boldsymbol{s}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i^*)}, \mathbf{v}^{(test,i^*)})}, \tag{4.2}$$
   where $L(\boldsymbol{s}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i)}, \mathbf{v}^{(test,i)})$ is proportional to the posterior $[\boldsymbol{s}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i)}, \mathbf{v}^{(test,i)}]$ which is given in Equation 2.13 of CBB.

   b. For $j \in \{1, \ldots, J_1\}$
      (i) Sample $(\tilde{\boldsymbol{Q}}^{(j)}, \tilde{\boldsymbol{\Sigma}}^{(j)})$ from $\{(\boldsymbol{Q}^{(1)}, \boldsymbol{\Sigma}^{(1)}), \ldots, (\boldsymbol{Q}^{(N)}, \boldsymbol{\Sigma}^{(N)})\}$ *without replacement*, where the probability of sampling $(\boldsymbol{Q}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)})$ is proportional to $w_{i^*,i}^{(\ell)}$.
      (ii) For fixed $(\boldsymbol{Q}, \boldsymbol{\Sigma}) = (\tilde{\boldsymbol{Q}}^{(j)}, \tilde{\boldsymbol{\Sigma}}^{(j)})$, draw $\boldsymbol{s}$ J_2 times from posterior density $[\boldsymbol{s} \mid \boldsymbol{Q}, \boldsymbol{\Sigma}, \mathcal{D}_s^{(-i)}, \mathbf{v}^{(test,i)}]$ using TMCMC, where for this choice of $\boldsymbol{Q}$ and $\boldsymbol{\Sigma}$,
$$[\boldsymbol{S} \mid \boldsymbol{Q}, \boldsymbol{\Sigma}, \mathcal{D}_s^{(-i)}, \mathbf{v}^{(test,i)}] \propto [\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i)}, \mathbf{v}^{(test,i)}]$$
      In this way, J_2 samples of $\boldsymbol{s}$ are obtained at each J_1.

   c. Store the $J_1 \times J_2$ draws of $\boldsymbol{s}$ as the $J_1 J_2$ number of posterior samples for $\boldsymbol{s}_i$ as $\hat{\boldsymbol{s}}_i^{(1)}, \ldots, \hat{\boldsymbol{s}}_i^{(J_1 J_2)}$.

4.1. *Simulation study*

In order to perform the cross-validation discussed above, on simulated data, we contrive a situation where there are $j = 3$ stars, each having $k = 2$ velocity components where velocity $\mathbf{v} = \boldsymbol{\xi}(\boldsymbol{s}) =$

$(\xi_1(\boldsymbol{s}), \ldots, \xi_6(\boldsymbol{s}))^T$, with the model parameter $\boldsymbol{S}$ being of dimension $d = 2$, $\boldsymbol{s} = (s^{(1)}, s^{(2)})$ (the two coordinates of the solar position). We assign the following forms to the component functions of the 6-dimensional vector-valued function $\boldsymbol{\xi}(\cdot)$.

$$\xi_1(\boldsymbol{s}) = \alpha s^{(1)} + \beta \frac{s^{(2)}}{1 + (s^{(1)})^2} + \gamma \cos(1.2 s^{(2)}) \tag{4.3}$$

$$\xi_2(\boldsymbol{s}) = \alpha s^{(1)} + \beta \frac{s^{(2)}}{1 + (s^{(1)})^2} \tag{4.4}$$

$$\xi_3(\boldsymbol{s}) = \alpha + \beta s^{(1)} \tag{4.5}$$

$$\xi_4(\boldsymbol{s}) = \gamma \cos(1.2 s^{(2)}) \tag{4.6}$$

$$\xi_5(\boldsymbol{s}) = \alpha s^{(1)} + \gamma \cos(1.2 s^{(2)}) \tag{4.7}$$

$$\xi_6(\boldsymbol{s}) = \gamma \cos(s^{(2)} + \sin(s^{(2)})), \tag{4.8}$$

where α, β and γ are chosen constants. Most of these above forms are modified versions of the functional forms used in Carlin, Polson and Stoffer (1992) and Bhattacharya (2007) in connection with dynamic models; see also Ghosh et al. (2013).

We generated 100 data points by first simulating $\boldsymbol{s}_i = (s_i^{(1)}, s_i^{(2)}); i = 1, \ldots, 100$ independently from $Uniform(-1, 1) \times Uniform(-1, 1)$, and then evaluating $\xi(\cdot)$ at each $\boldsymbol{s}_i$, using the component-wise functional forms given above (4.3)—(4.8). Here we set $\alpha = 0.05$, $\beta = 0.1$ and $\gamma = 0.2$. We thus obtained 100 data points $(\boldsymbol{s}_i, \mathcal{V}_i); i = 1, \ldots, 100$.

We leave out each data point in turn, predicting the corresponding location using the remaining data points and the corresponding velocity matrix. Using the distance minimisation method discussed in the last sub-section we obtain $i^* = 43$; hence the importance sampling density is $[\boldsymbol{s}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathbf{v}_{43}, \sqsubseteq^{(test,i)}]$.

4.2. *Details of IRMCMC implementation to simulated data*

We implemented TMCMC following the details provided in Section 3.1 of CBB in order to simulate from the importance sampling density at $i^* = 43$. Specifically, for updating $\boldsymbol{s}$ using TMCMC, the parameter ϵ–a scaled value of which the proposed state is an additive transformation–is chosen to be $\epsilon \sim N(0, 1) I_{\{\epsilon > 0\}}$ while the scale factors c_1 and c_2 are chosen to be 0.1 and 50. For the smoothness parameters, i.e. the elements of the diagonal matrix $\boldsymbol{Q}$, we simulated ϵ from a zero mean normal distribution with variance 0.005, restricted to $\mathbb{R}_+$, and selected 0.1 and 1 as the scale factors. For

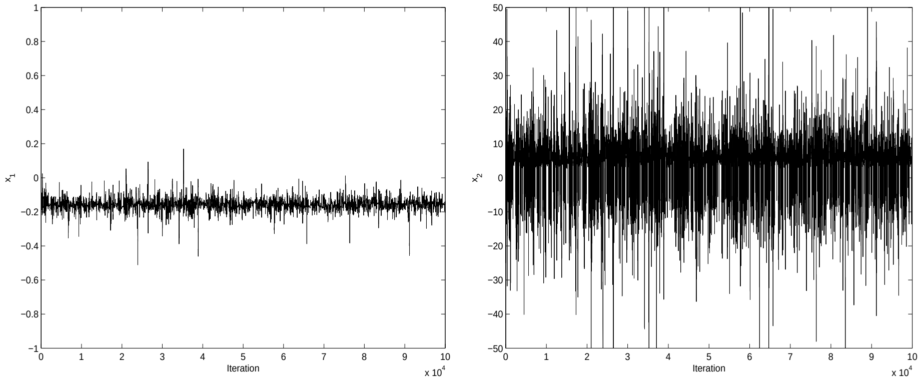


FIG 2. **Simulation study:** *Trace plots of s_1 and s_2 corresponding to $i^* = 43$.*

updating $\boldsymbol{\Sigma}$, we simulated the transformation parameter ϵ from the positively restricted zero mean normal distribution with variance 0.005 and set the scale factors to be 0.07 for the non-zero elements. These choices are arrived at after assessing TMCMC convergence in several pilot runs.

We discarded the first 100,000 TMCMC runs corresponding to the obtained i^* as burn-in and stored the next 100,000 runs for IR purposes. Informal convergence diagnostics indicated reasonable convergence; see, for example, the trace plots of s_1 and s_2 in Figure 2. From these 100,000 samples we simulated 100 realisations of $(\boldsymbol{Q}, \boldsymbol{\Sigma})$ using IR without replacement. For each IR-realised $(\boldsymbol{Q}, \boldsymbol{\Sigma})$ we simulated 1000 realisations of $\boldsymbol{s}$ using TMCMC; In this implementation of TMCMC we used a burn-in of 100,000 iterations of $\boldsymbol{s}$, starting from an initial value generated uniformly over $[-1, 1] \times [-1, 1]$. Thereafter, for the remaining 99 IR-realisations, we used the last realisation of $\boldsymbol{s}$ as the initial value for the first realisation of $\boldsymbol{s}$, without discarding any iteration as burn-in. This was done at the previous IR-realisation of $(\boldsymbol{Q}, \boldsymbol{\Sigma})$. That this is a valid and efficient strategy, has been established by Bhattacharya and Haslett (2007). Thus, we obtain 100×1000=100,000 IRMCMC realisations of $\boldsymbol{s}$. Each such set of 100,000 realisations was generated for each omitted data point.

This entire exercise took around 49 hours on a laptop; posterior simulation corresponding to i^* took around one hour, while the remaining exercise took a further period of 48 hours approximately. It is to be noted that brute-force cross-validation in this example would have taken 100 hours. Hence IRMCMC lives up to the expectation of reducing the computation time.

4.3. Results of cross-validation on simulated data

In all of the 100% simulated cases, the true locations fell within the 95% highest posterior density credible intervals of the corresponding leave-one-out IRMCMC-based posteriors. Some of the leave-one-out cross-validation posteriors, along with the corresponding true values are shown in Figures 3 and 4. The results indicate a good fit to the data and are therefore encouraging as far as the application of our high-dimensional Gaussian process-based method is concerned.

4.4. IRMCMC-based cross-validation using data generated from base astrophysical models

For each of the four base astrophysical models (bar_6, $sp3bar3$, $sp3bar3\_18$ and $sp3bar3\_25$), we have a training data set consisting of 216 observations on 100-dimensional motion vectors, each generated at a distinct value of the design vector. In order to validate our Gaussian process based methodology we perform leave-one-out cross-validation for each of the four training data sets using IRMCMC in conjunction with TMCMC.

4.5. Prior on location in the context of cross-validation

For the cross-validation purpose, we assume a somewhat expanded parameter space for the locations: $(R_\odot, \Phi_\odot) \in (1,3) \times [0,\pi)$, instead of the parameter space $[1.7, 2.3] \times [0, \pi/2]$, which was assumed for actually predicting the unknown location associated with the real, test data set. The reason for expanding the parameter space is that the training data sets consist of many observations that lie almost on the boundary of $[1.7, 2.3] \times [0, \pi/2]$ and our initial cross-validation showed that many boundary values were excluded from the 95% credible regions of their respective cross-validation posteriors. Indeed, for both classical and Bayesian asymptotics the important regularity condition that is typically assumed is that the true value of the parameter lies within the interior of the parameter space (page 436 of Schervish (1995)).

Note that the aforementioned expansion of the parameter space of $(R_\odot, \Phi_\odot)$ for cross-validation purpose is not in conflict with the uniform prior on $[1.7, 2.3] \times [0, \pi/2]$, which we assumed for predicting the unknown location corresponding to the real, training data set. Indeed, guided by the astrophysics literature we believe *apriori* that the true location lies in the interior of $[1.7, 2.3] \times [0, \pi/2]$, not on the boundary.

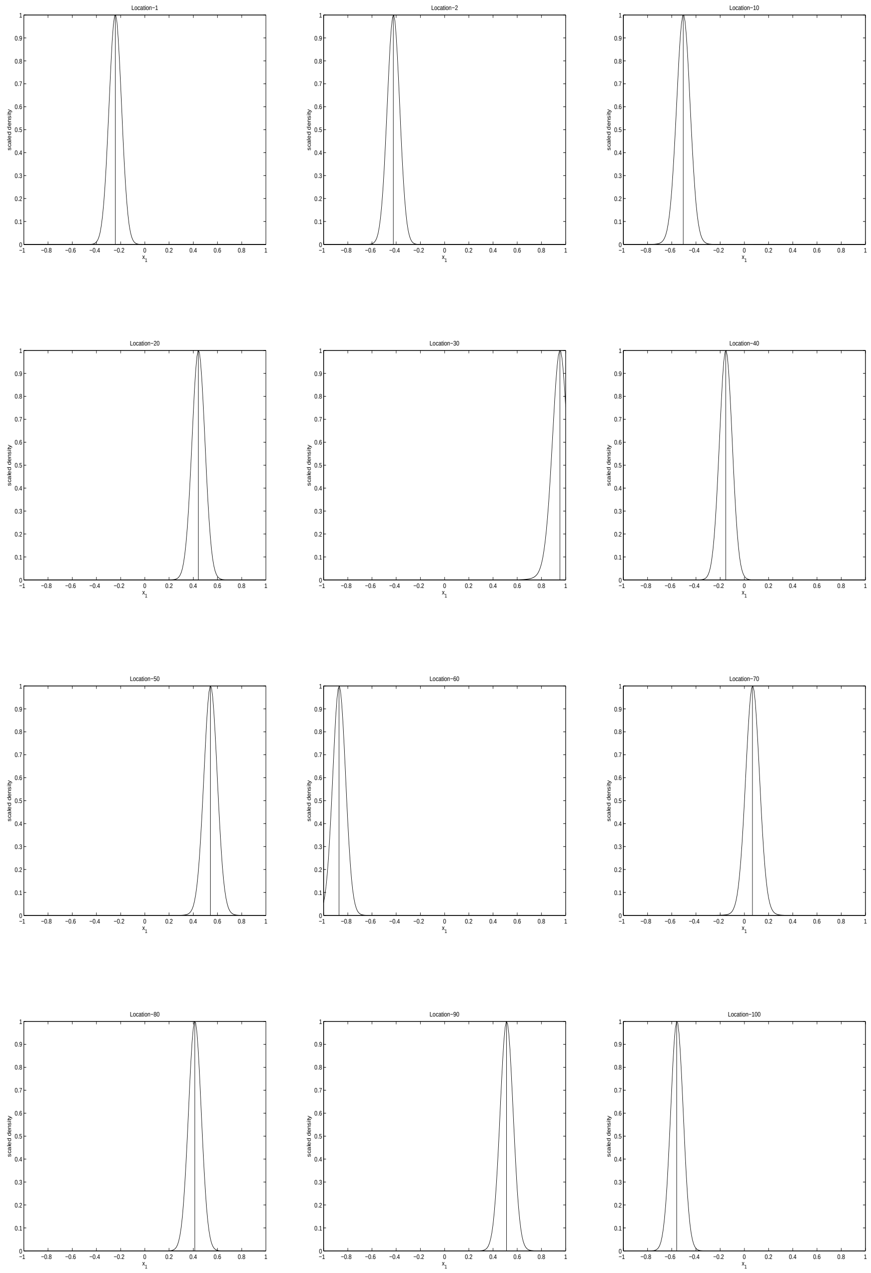


FIG 3. **Simulation study:** *Leave-one-out cross-validation posteriors of s_1; the vertical line indicates the true value.*

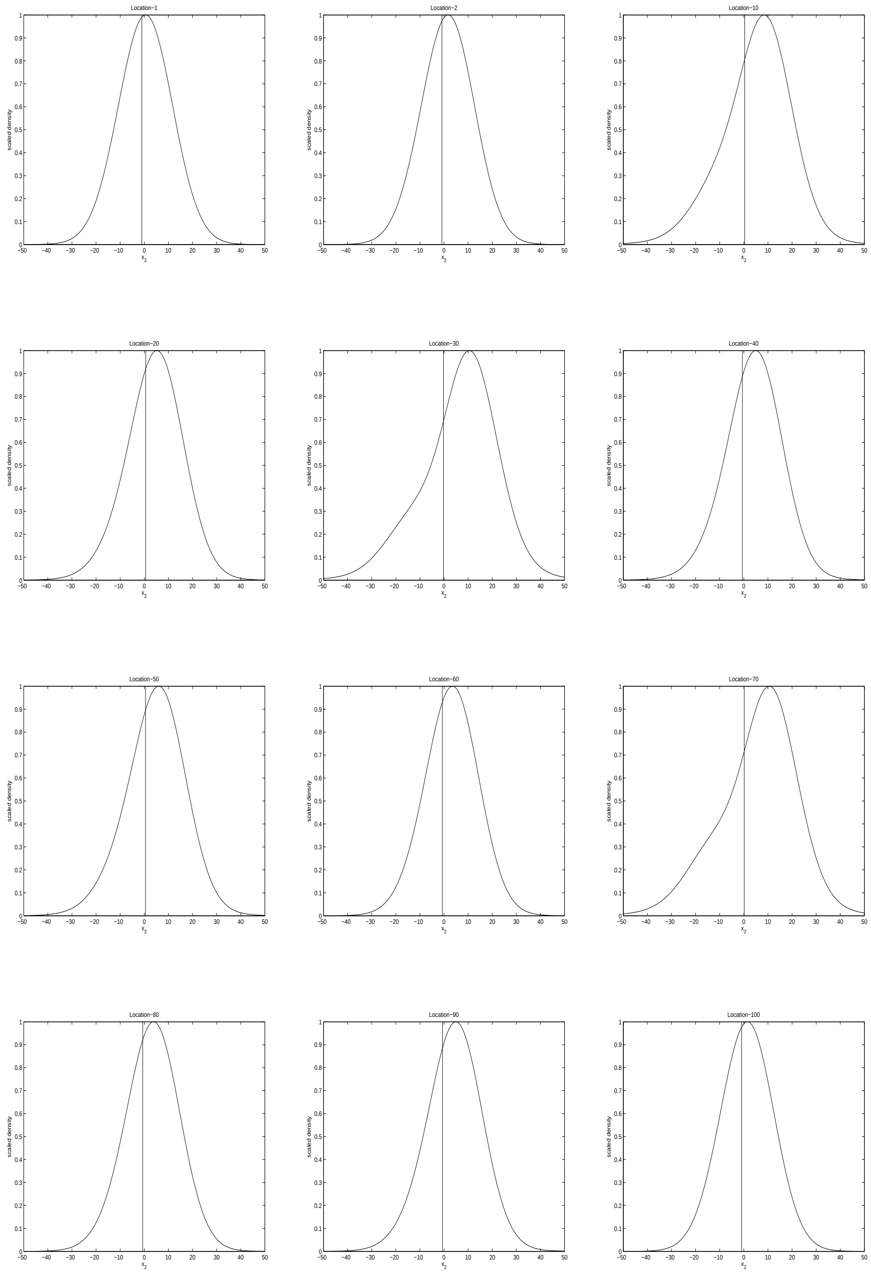


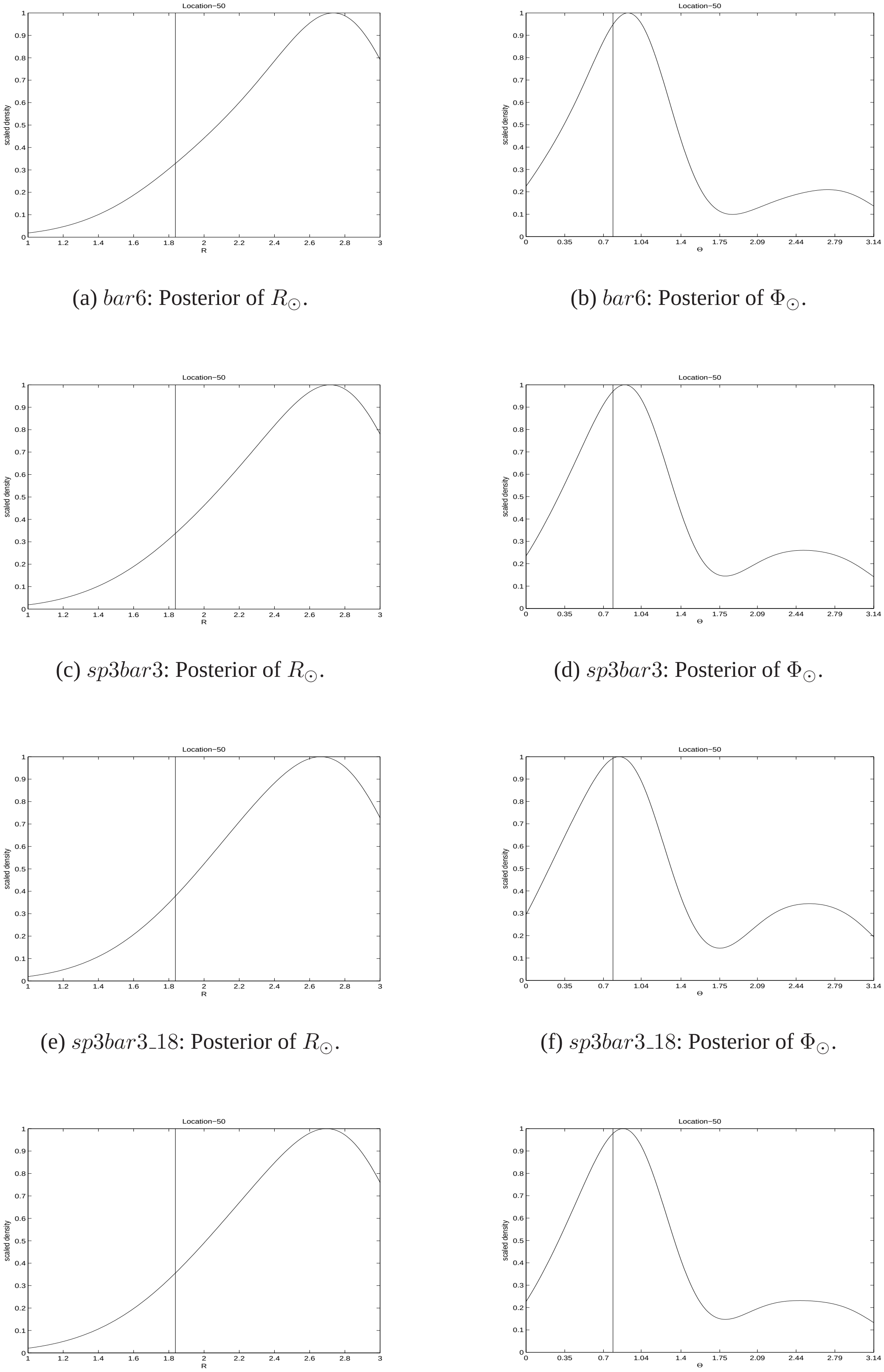
FIG 4. **Simulation study:** *Leave-one-out cross-validation posteriors of s_2; the vertical line indicates the true value.*

IRMCMC has an inbuilt strategy of handling multimodality by incorporating re-starts in Step b (ii) of the IRMCMC algorithm provided in Section 4.1. To clarify, given a $(\tilde{\boldsymbol{Q}}^{(j)}, \tilde{\boldsymbol{\Sigma}}^{(j)})$, we can use *independent starting points of* $\boldsymbol{s}$ *and a subsequent burn-in for every* $j \in \{1, \ldots, J_1\}$, while drawing from the posterior $[\boldsymbol{s} \mid \tilde{\boldsymbol{Q}}^{(j)}, \tilde{\boldsymbol{\Sigma}}^{(j)}, \mathcal{D}_s^{(-i)}, \mathbf{v}^{(test,i)}]$. The independent initialising values can be drawn uniformly from the parameter space of $\boldsymbol{s}$. This multiple re-start strategy ensures that for adequately large J_1, all the modes of the multimodal posterior are explored by IRMCMC; that is, the IRMCMC sample $\{\hat{\boldsymbol{s}}_1^{(1)}, \ldots, \hat{\boldsymbol{s}}^{(J_1 J_2)}\}$ will then adequately represent the multimodal i-th cross-validation posterior (Bhattacharya and Haslett, 2007).

4.6. Results of cross-validation on real stellar velocity data

For our implementation of the above-discussed re-start based IRMCMC, we choose $N = 20,000$, $J_1 = 50$, a burn-in of size $1,000$ for every re-start, and $J_2 = 4,000$. Thus, IRMCMC for every cross-validation posterior yields $J_1 J_2 = 50 \times 4,000 = 20,000$ samples. For each of the four models, 100% observed $r_\odot$ fell within the 95% credible regions of their respective cross-validation posteriors. In the case of $\phi_\odot$, all but the minimum observed value of $\phi_\odot$ in the training data sets, which is about 0.08, are captured by the respective 95% credible regions.

Figure 5 displays the cross-validation posteriors corresponding to the 50-th data point of each of the four training data sets. The true or held out values of $R_\odot$ and $\Phi_\odot$, as inferred using our Bayesian method and TMCMC, are denoted by vertical lines in the panels of Figure 5. The cross-validation posteriors corresponding to the different models are very similar, even though the posteriors of the unknown location associated with the real test data set are quite different (recall Figures 1, 2, 3, 4 of CBB. However, there is no conflict between these two issues. Figures 1, 2, 3, 4 of CBB correspond to different training data sets, but all these have a common test data set. On the other hand, in the cross-validation scenario, while predicting a particular observed location, the held out test data sets are also different for the four different cross-validation studies. Thus for example, the test data employed in predicting the i-th held out data point is $\mathbf{v}^{(test,i)} := \mathbf{v}_i$. The cross-validation results suggest that the four different model-specific test data sets used to predict a location common to all the four models (that is, the original four training data sets) provide similar information regarding the held out location, in conjunction with the remaining model-specific training data set.



(a) $bar6$: Posterior of $R_\odot$.

(b) $bar6$: Posterior of $\Phi_\odot$.

(c) $sp3bar3$: Posterior of $R_\odot$.

(d) $sp3bar3$: Posterior of $\Phi_\odot$.

(e) $sp3bar3\_18$: Posterior of $R_\odot$.

(f) $sp3bar3\_18$: Posterior of $\Phi_\odot$.

(g) $sp3bar3\_25$: Posterior of $R_\odot$.

(h) $sp3bar3\_25$: Posterior of $\Phi_\odot$.

FIG 5. **Real data:** *Leave-one-out cross-validation posteriors of the model parameter (50-th data point of the training data sets left out); the vertical line indicates the true (held out) value of $\boldsymbol{S}$.*

5. Effects of chaos

The concurrence of our results with the results reported in astrophysical literature (see Setion 4 of CBB) goes beyond just the summaries of the posteriors of the solar position vector; remarkable correlation can be noticed between the measure of chaos in these 4 astrophysical models - as estimated by (Chakrabarty and Sideris, 2008) - and the multi-modality of the posterior distribution of $\boldsymbol{S}$ that we advance. Chakrabarty and Sideris (2008) report minimum chaos in the bar_6 model compared to the other three, while we notice the posteriors of both $r_\odot$ and $\phi_\odot$ in this model to be the unimodal. In fact the posteriors of $r_\odot$ and $\phi_\odot$ are unimodal only for this model, out of the 4 astrophysical models that we use to illustrate the efficacy of our method. Perhaps more importantly, the $sp3bar3$ model is noticed to manifest maximum (even global) chaoticity, on theoretical grounds by Chakrabarty (2007), backed by the chaos quantification at higher energies (Chakrabarty and Sideris, 2008). Likewise, in our work, the posterior distributions for $r_\odot$ and $\phi_\odot$ are most multi-modal in this model, compared to the other three. The models $sp3bar3\_18$ and $sp3bar3\_25$ are considered to be of intermediate chaoticity and we find these to correspond to posterior distributions (of $\boldsymbol{s}^{(new)}$) that are multi-modal, though less so, than that for the model $sp3bar3$.

The exact physical reason for the correlation between chaos in the base astrophysical model of the Milky Way and the multi-modality of the posterior distribution of $\boldsymbol{S}$ is understood if we begin with the premise that increased chaos is responsible for increased scatter in the distribution of the stellar velocity vector values that are generated at a chosen design vector. While for zero chaos, a distinct set of data vectors is generated at a given set of experimental conditions (a value of $\boldsymbol{S}$), increased scatter implies that the same data set can result from multiple experimental conditions (multiple values of $\boldsymbol{S}$). In fact, a necessary condition for chaos to occur is the increasing non-injectivity of $\boldsymbol{\xi}(\cdot)$ (Sengupta, 2003) where data vector $\mathbf{v} = \boldsymbol{\xi}(\boldsymbol{s})$. Thus in a base model that has zero chaoticity–eg. the bar_6 model which Chakrabarty and Sideris (2008) found to have near zero chaos–the velocity vectors generated at different values of $\boldsymbol{S}$ are distinct in general. However in the other 3 base models that were reported to bear a very high fraction of chaotic orbits, similar velocity vectors can be generated at different values of $\boldsymbol{S}$.

In summary, the function $\boldsymbol{\xi}(\cdot)$ that is learnt from the training data will be rendered increasingly more non-injective with increasing chaoticity in the base model from which the training data is generated. Thus, with increased chaoticity, $\boldsymbol{\xi}^{-1}(\cdot)$ becomes multivalued, i.e. the same observed velocity is predicted to be realised at multiple values of $\boldsymbol{S}$. The increase in the non-uniqueness of our achieved solution is thus physically motivated by the different amounts of chaos in the base astrophysical models. While this non-uniqueness can only be relieved by invoking further information–if and when such become available–our inference allows for the identification of all $\boldsymbol{s}^{(new)}$ that are consistent with the data.

References

BHATTACHARYA, S. (2007). A Simulation Approach to Bayesian Emulation of Complex Dynamic Computer Models. *Bayesian Analysis* **2** 783–816.

BHATTACHARYA, S. and HASLETT, J. (2007). Importance Resampling MCMC for Cross-Validation in Inverse Problems. *Bayesian Analysis* **2** 385–408.

BINNEY, J. and MERRIFIELD, M. (1998). *Galactic Astronomy*. Princeton University Press, Princeton.

CARLIN, B. P., POLSON, N. G. and STOFFER, D. S. (1992). A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association* **87** 493–500.

CHAKRABARTY, D. (2007). Phase Space around the Solar Neighbourhood. *Astronomy & Astrophysics* **467** 145.

CHAKRABARTY, D. and SIDERIS, I. (2008). Chaos in Models of the Solar Neighbourhood. *Astronomy & Astrophysics* **488** 161.

DUTTA, S. and BHATTACHARYA, S. (2013). Markov Chain Monte Carlo Based on Deterministic Transformations. *Statistical Methodology*. To appear. *Available at arxiv:1106.5850v3 with supplementary section in arxiv.org/pdf/1306.6684*.

FUX, R. (2001). Order and chaos in the local disc stellar kinematics induced by the Galactic bar. *Astronomy & Astrophysics* **373** 511-535.

GELFAND, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (W. GILKS, S. RICHARDSON and D. SPIEGELHALTER, eds.). *Interdisciplinary Statistics* 145–162. Chapman and Hall, London.

GELFAND, A. E., DEY, D. K. and CHANG, H. (1992). Model determination using predictive distributions with implementation via sampling methods(with discussion). In *Bayesian Statistics 4* (J. M. BERNARDO, J. O. BERGER, A. P. DAWID and A. F. M. SMITH, eds.) 147–167. Oxford University Press.

GHOSH, A., MUKHOPADHYAY, S., ROY, S. and BHATTACHARYA, S. (2013). Bayesian Inference in Nonparametric Dynamic State-Space Models. Submitted, available at http://arxiv.org/abs/1108.3262.

SCHERVISH, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.

SENGUPTA, A. (2003). Toward a Theory of Chaos. *International Journal of Bifurcation and Chaos* **13** 3147-3233.