BAYESIAN INVERSE LEARNING OF MILKY WAY MODEL PARAMETERS USING MATRIX-VARIATE GAUSSIAN PROCESS-BASED METHOD

BY DALIA CHAKRABARTY[*,§], MUNMUN BISWAS[†,¶], SOURABH BHATTACHARYA[‡,¶],

University of Warwick[§], Indian Statistical Institute, Kolkata[¶]

The modelling of the Milky Way galaxy is an integral step in the study of galactic dynamics; this owes not only to our natural inquisitiveness about our residence in the Universe, but also because knowledge of model parameters that define the Milky Way directly influences our understanding of the evolution of our galaxy. Since the nature of the phase space in the neighbourhood of the Sun is affected by distinct Milky Way features, measurements of phase space coordinates of individual stars that live in the neighbourhood of the Sun, will bear the influence of such features. Then, inversion of such measurements can help us learn the parameters that describe such Milky Way features.

In the past this has been attempted via the "calibration method", a data and computational cost intensive method that is limited by such logistical shortcomings. We argue in this paper that the calibration method is unsatisfactory because of reasons both methodological and computational. Here we develop a Bayesian inverse problem approach, where we model the available stellar velocity information matrix as an unknown function of the Milky Way model parameters, where this function is inverted using Bayesian techniques to predict the model parameters. This unknown function turns out to be matrix-variate, which we model as a matrix-variate Gaussian Process. We develop a general method to perform inverse, nonparametric learning, using matrix-variate Gaussian Processes. For the inference we use the recently advanced Transformation-based Markov Chain Monte Carlo (TMCMC).

Application of our method to observed stellar velocity data results in estimates that are consistent with those in astrophysical literature. That we could obtain these results using far smaller data sets compared to those required for the calibration approach, is encouraging in terms of projected applications to external galaxies.

1. Introduction. Curiosity about the nature of the phase space that we earthlings live in, is only natural. On astronomical length scales, this translates to the pursuit of the topology of phase space in the Solar neighbourhood. By phase space of the stars in the Milky Way, is implied $\mathcal{W}$, the space of the spatial $\mathbf{X}$ and velocity $\mathbf{P}$ vectors of all the stars in the Galaxy, where the spatial location of the i-th star is $\mathbf{X}_i \in \mathbb{R}^3$, $\mathbf{X}_i = (X_1^{(i)}, X_2^{(i)}, X_3^{(i)})^T$, and its velocity vector $\mathbf{P}_i \in \mathbb{R}^3$, $\mathbf{P}_i =$

[*]Research fellow at Department of Statistics, University of Warwick; supported by a Warwick Centre for Analytical Sciences Research Fellowship

[†]Munmun Biswas is a PhD student in Statistics and Mathematics Unit, Indian Statistical Institute

[‡]Sourabh Bhattacharya is an Assistant Professor in Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute

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

$(P_1^{(i)}, P_2^{(i)}, P_3^{(i)})^T$. The phase space of the Milky Way in the vicinity of the Sun, has been studied and astrophysical modelling indicates that in this local patch, the phase space structure is sculpted by different dynamical features of the Milky Way (Antoja et al., 2009; Chakrabarty, 2007; Minchev et al., 2009), such as a central, rotating, stellar bar in the Galaxy as well as the spiral pattern of the Galaxy. Thus, the phase portrait of the solar neighbourhood is affected by the parameters that define these Galactic features, though this parameter space might need to be expanded to include other parameters that govern features that have been hitherto ignored from such modelling. Relevant parameters include the the orientation of us, i.e. the Sun, with respect to the major axis of the bar, the distance from us. i.e. the Sun, to the centre of the Galaxy, structural parameters of the bar and the spiral pattern, as well as parameters that describe the dynamics of these features, such as the frequencies of rotation of the spiral pattern Ω_s and that of the bar Ω_b, etc. Given that these features sculpt the local patch of our Galaxy, if phase space coordinates of a sample of stars that live in the local patch are made available, the data will bear information about the handiwork of these features.

It is to be noted that 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 spatial vector of the i-th star $(X_1^{(i)}, X_2^{(i)})^T$, and the velocity vector $(P_1^{(i)}, P_2^{(i)})^T$. Given the disky spatial geometry, we find it sometimes useful to use the polar, rather than the Cartesian coordinate system to scan the Galaxy. Then the radial coordinate of a particle is R and azimuthal coordinate is Φ, such that $X_1 = R\cos\Phi$, $X_2 = R\sin\Phi$, $R \geq 0$, $\Phi \in [0, 2\pi]$. The radial and transverse components of the velocity vector are then U and V.

Now, phase space data of a sample of stars in the vicinity of the Sun, are available (Francis and Anderson, 2009; Fux, 2001). Given this observed phase space data, we propose a nonparametric Bayesian methodology to help learn the Galactic parameters $\mathbf{S} \in \mathbb{R}^d$, that affect the structure of the phase space around the Sun, in our work. Essentially, we are attempting an inverse problem, namely that of learning $\mathbf{S}$ from the sampled discrete phase space stellar data, $\{\mathbf{w}_i\}_{i=1}^j$, where a total of j stars have been observed. We realise that the phase space data is drawn from the probability density function of the galactic phase space. However, the data-driven nature of our method allows for the learning of $\mathbf{S}$ without requiring modelling of this phase space density. In particular, here we demonstrate our methodology on inversely learning the Milky Way model parameters that are the coordinates of the spatial location of the Sun in the Milky Way disk, with respect to the centre of the Galaxy. Thus, for the application dealt with herein, $\mathbf{S}$ is 2-dimensional. The sought model parameters are then the radial $R_\odot$ and azimuthal $\Phi_\odot$ coordinates of the Sun with respect to the centre of the Galaxy; $R_\odot > 0$, $\Phi_\odot \in [0, 2\pi]$.

While the Milky Way model parameters are sought with respect to the centre of the Galaxy, the measurements that are available, have been taken by us, i.e. with respect to the Sun. In other words, the measured stellar velocities $(U^{(i)}, V^{(i)})^T$, $i = 1, \ldots, j$, are with respect to the solar velocity and the measured spatial locations of the stars are with respect to the location of the Sun. Furthermore, the

spatial locations of the sampled stars are within a small distance ϵ from the Sun (Fux, 2001). We interpret these data to have been sampled from a closed 2-ball with radius ϵ centred at the Sun, living in the space of the Milky Way disk. Also, the azimuthal distribution of the sampled stars, inside this ball, is approximately uniform. Thus, the summary of the distribution of the measured spatial locations of the sampled stars coincides with the origin of a heliocentric spatial coordinate system and these measured spatial locations cannot constrain the sought galactocentric location of the Sun.

So the measurements (from the Sun) of spatial locations of the sampled stars do not provide any constraints on the unknown $(R_\odot, \Phi_\odot)$, though the measured heliocentric stellar velocities are indeed affected by the choice of $(R_\odot, \Phi_\odot)$ - the motion of a given star, if tracked from different observer locations, would register differently. In principle, beyond just the galactocentric solar location, if there are Milky Way parameters that physically affect the structure of the phase space, velocity information drawn from that phase space will bear the signature of such Milky Way parameters. Thus, the available velocity information $\mathbf{V}$ is considered to bear the signature of Milky Way parameters $\mathbf{S}$. This is equivalent to saying that $\mathbf{V} = \boldsymbol{\xi}(\mathbf{S})$, where $\boldsymbol{\xi}(\cdot)$ is unknown and matrix-variate function, given that the velocity information $\mathbf{V}$ is a matrix, with j rows and k columns. We will inversely learn $\mathbf{S}$ within a Bayesian nonparametric framework. The measured data used in our work is $\mathcal{D}_{test} := \{(U_i, V_i)^T\}_{i=1}^{j}$. In this work, we supplement information about the unknown Milky Way parameters via a training data set, $\mathcal{D}_s$, that is generated from dynamical simulations of astronomical models of our Galaxy. The nature of generation of the training data is described below.

The inverse method that we advance below is a generic methodology that can be applied to estimate the d-dimensional vector $\mathbf{S}$, given $\mathcal{D}_{test}$ and $\mathcal{D}_s$, where d is bounded from below by the number of parameters in the astrophysical models that can be called upon to bear influence on the phase space structure of the Galaxy in the solar neighbourhood. As we have said above, we demonstrate an application of our developed inverse learning technique to the learning of the solar position; the reason for restricting our application to the case of d=2 is the existence of stellar phase space data generated in dynamical simulations of astrophysical models of the Milky Way which involve scanning over chosen guesses for $R_\odot$ and $\Phi_\odot$. If simulated data distinguished by choices of other Milky Way model parameters become available, then the implementation of these data as training data will be possible. Then the learning of Milky way model parameters in addition to $R_\odot$ and $\Phi_\odot$ will be possible. The methodology that we advance is indeed a generic one, with computational costs being the only concern in extending to cases of $d > 2$; that extending to a higher dimensional $\mathbf{S}$ only linearly scales computational costs, is discussed later in Section 7.

The training data $\mathcal{D}_s^{(q)}$ constitutes of a sequence of n simulated stellar velocity data matrices, generated using the q-th astrophysical model of the Milky Way $q = 1, \ldots, 4$. The i-th of these simulated velocity data matrices is characterised by the i-th model parameter vector value of $\mathbf{s}^{(i)}$ which in our implementation, reduces to the i-th choice of the solar spatial location, $i = 1, \ldots, n$. In fact, for each i, j simulated

stellar velocity vectors is generated. The simulations of the astrophysical models discussed above are from Chakrabarty (2007). Each training data set generated for any of the 4 astrophysical models, is analysed one at a time; consequently, from now, we will drop any explicit reference to the chosen astrophysical model in the notation of the training data. Details of the dynamical simulations are given in the supplementary section **S-1**.

Of course, it is of aesthetic interest to learn our location inside our galaxy, but the relevance of this exercise goes much beyond. Any modelling of the Milky Way invokes the solar location - in that sense, the solar location is of vital importance in the study of the Milky Way. Importantly, the $\Phi = 0$ axis is chosen in our models to coincide with the long axis of the central stellar bar that exists in the Galaxy. Thus, learning $\Phi_\odot$ would allow for an understanding of the separation of the Sun-Galactic centre line from the bar; this in turn will allow for improved understanding of the effect of the bar on the dynamics of the solar neighbourhood (Englmaier and Gerhard, 1999; Fux, 2001; Minchev et al., 2010). At the same time, learning $R_\odot$ can crucially constrain the rotational frequency of the bar in the Milky Way, which affects the understanding of some crucial physical processes in the Milky Way. The azimuthal location of the Sun - given that this is with respect to the Φ=0 axis that is chosen to coincide with the long axis of the bar - is interpretable as the azimuthal separation of the Sun from the bar.

The rest of the paper is structured as follows. In Section 2 we discuss some salient differences between our methodology and other approaches found in the literature. The motivation for a Gaussian Process-based inverse learning, is clarified in Section 3 while in Section 4 we present the details of this new inverse method. Details of the inference strategies used in our work are given in Section 5 and Section 6 contains results obtained from using real data. The paper is rounded up with a discussions section.

2. Comparison of the Bayesian approach with other available approaches. In this paper we view velocity as a (unknown) function of the Milky Way model parameters, and model this unknown function as a Gaussian Process, subsequently inverting the function using the Bayesian methodology that we develop here. As we elucidate subsequently, our approach to this inverse problem has some significant advantages over the approach based on density estimation that was adopted by Chakrabarty (2007). Moreover, the inverse method that we advance here can be applied to estimate the d-dimensional vector $\mathbf{S}$, though here we demonstrate an application of our developed technique to the learning of $(R_\odot, \Phi_\odot)^T$, given the test and the training data sets - i.e., we will illustrate the case of $d = 2$.

2.1. *Calibration approach.* Chakrabarty (2007) undertook an approach that is an example of the "calibration method". The objective was to predict the solar location given the available stellar velocity measurements and $\mathcal{D}_s$. The methodology entailed pairwise comparison of the achieved empirical density (kernel density estimate) estimated using the observed discrete velocities of $\mathcal{D}_{test}$ and that using the j simulated velocities in the i-th simulated velocity matrix $\mathbf{V}_i$ corresponding to the

choice $\mathbf{S} = \mathbf{s}_i$, for each $i = 1, \ldots, n$. The distribution of the strength of match between the two densities, in the space of $\mathbf{S}$, can be identified using a non-parametric statistical test of hypothesis that was developed for the purpose - see Chakrabarty (2011). This empirical distribution was used to obtain an interval estimate of the unknown observer location.

Since the best match is sought over a very large collection of training data points, obtained by constructing a grid over the relevant sample space, the method requires computational effort and resources. This shortcoming had compelled Chakrabarty (2007) to resort to an unsatisfactory coarse gridding of the space of $\mathbf{S}$. This problem gets acute enough for the method to be rendered useless when the dimensionality of the vector $\mathbf{S}$ that we hope to learn, is increased. Moreover, the method of quantification of uncertainty of the estimate of the location is also unsatisfactory, dependent crucially on the binning details of the space of $\mathbf{S}$, which in turn is bounded by cost and memory considerations.

2.2. *Classical approach.* In classical statistics, the training data set typically provides information about the unknown parameters, given the postulated parametric/nonparametric model. In other words, the purpose of the training data set is to train the model; statistically this boils down to estimation of the model parameters. Using the trained model, it is then possible to predict unknown features of the test data set. The model-based classical approach has a significant advantage over the calibration method; since the information contained in the training data is supplemented by the assumed model, training data sets with sizes far smaller than those required in the calibration approach suffice in the classical statistical paradigm.

As indicated above, we are interested in the inverse problem of estimating $\mathbf{S}$, given velocity information, where we express the velocity information matrix as a function of $\mathbf{S}$: $\mathbf{V} = \boldsymbol{\xi}(\mathbf{S})$. It could potentially be modelled non-parametrically using perhaps splines or wavelets. Our interest lies in the prediction of $\mathbf{S} = \boldsymbol{\xi}^{-1}(\mathbf{V})$, given the measured and training data. For this inversion it is first required to estimate $\boldsymbol{\xi}(\cdot)$ from the data using well-known nonparametric techniques, and then use numerical methods to achieve the inverse solution. The complexity of both these exercises - particularly the latter - increases as the dimensionality of the function space increases, unless simplifying assumptions are made.

The classical approach has been criticised on the ground that it ignores parameter uncertainty, in that the approach provides only a point estimate of the set of model parameters, and based on that single point estimate, attempt is made to predict the unknown observations. As a result, the variances of the predicted parameters, even if available by any means, are likely to be deceptively small. Furthermore, modelling high-dimensional functions using splines/wavelets require restrictive assumptions and in particular, may fail to adequately take into account the correlation structure between the component functions.

2.3. *Bayesian approach.* One Bayesian equivalent of this approach could involve the modelling of the unknown underlying function $\boldsymbol{\xi}(\cdot)$ with splines or wavelets, and then computing the posterior distribution of $\mathbf{S}$, conditional on the training and

the test data sets, after marginalising over the other model parameters. This computation, as well as the marginalisation, are in principle straightforward to implement using well-known Markov chain Monte Carlo (MCMC) methods. Thus, the Bayesian approach avoids the difficulties of nonparametric model parameter estimation and subsequent inversion, both of which turn out to be deterrent to the classical approach.

Additionally, unlike the classical paradigm the Bayesian approach does not ignore parameter uncertainty while predicting the unknown location. Furthermore, quantification of uncertainty in the unknown parameters in the form of prior distribution, in addition to model specification, enhances the information contained in the training data. Thus, in principle, a training data set of even smaller size than that required in the classical approach, will be adequate in the Bayesian paradigm.

However, as already discussed in Section 2.2, modelling high-dimensional functions using splines/wavelets is somewhat restrictive in that this approach fails to provide an appropriate correlation structure for the component-wise functions. Such an approach is also not of direct relevance, given that here, we aim to predict the model parameter vector $\mathbf{S}$ when training and observed data are available and do not have an explicit interest in the learning of $\boldsymbol{\xi}(\mathbf{s})$ or its inverse. In this paper, we circumvent this problem by introducing a matrix-variate Gaussian Process approach to modelling the high-dimensional functions; in this advanced paradigm, the unknown velocity function $\boldsymbol{\xi}(\cdot)$ is treated as high-dimensional matrix, the elements of which are unknown functions. The model and the methodology that we propose are flexible, robust, and amenable to high-dimensional situations. For the inference, we adopt the transformation-based MCMC (TMCMC) methodology, recently introduced by Dutta and Bhattacharya (2011).

3. Motivation for inverse learning using matrix-variate Gaussian Process. In the Bayesian approach that we adopt here, we learn the unknown velocity function $\boldsymbol{\xi}(\mathbf{s})$ by imposing a Gaussian Process ($\mathcal{GP}$) prior $\eta(\mathbf{s})$ on it. The application that we focus upon in this work demands new methodology from two points of view. Firstly, standard implementation of $\mathcal{GP}$ is typically on data that is a vector. However, in our application, the variable that is measured, $\mathbf{P}$, is a vector which renders the velocity information comprising multiple measured velocity vectors, a matrix. This in turn renders the unknown function $\boldsymbol{\xi}(\cdot)$ matrix-variate, thus requiring us to develop the matrix-variate Gaussian Process.

Importantly, we are attempting the inverse problem of predicting the new value of the model parameter given the training and test data, i.e. predict $\mathbf{s}^{(new)}|\mathcal{D}_s, \mathcal{D}_{test}$. In contrast to this, existing methodology that involve implementation of Gaussian Processes are often motivated to solve the problem of predicting the new data $\mathbf{v}^{(new)}$, given the data and the known new model parameter $\mathbf{s}^{(new)}$, i.e. attempt the forward problem. However, in an inverse problem like the one we attempt here, the interest lies in the prediction of $\mathbf{s}^{(new)}$ given $\mathbf{v}^{(new)}$ and the training data set $\mathcal{D}_s$. In this case, the posterior predictive distribution $[\mathbf{s}^{(new)} \mid b, \mathbf{v}^{(new)}, \mathcal{D}_s]$, which is the required solution, does not have an analytic form unlike in the forward case (see supplementary section **S-2**). In the next section we advance a new inverse

methodology that implements matrix-variate $\mathcal{GP}$.

4. The matrix-variate function. In this section we present a method that accommodates our twosome ambition of inverse modelling of data that is in the form of a matrix. We recall $n := N_R \times N_\phi$ and the training data comprises j simulated stellar velocity vectors with k number of components, generated at each of the n values of the Milky Way model parameter $\mathbf{S}$. This training data matrix is referred to as $\mathcal{D}_s$. It will be shown below that in this Bayesian approach, a much smaller j (=50) allows for inference on $\mathbf{s}$ that concurs with the inference obtained for $j \sim$3000 in the aforementioned calibration approach.

We express the velocity information as a $j \times k$-dimensional velocity matrix $\mathbf{V}$. That the model parameters have a bearing on the data, allows us to express as

$$\mathbf{V} = \boldsymbol{\xi}(\boldsymbol{s}), \tag{4.1}$$

where the unknown velocity function $\boldsymbol{\xi}(\cdot)$ is a $j \times k$-dimensional matrix. We define $\boldsymbol{\xi}^{(j\times k)}(\cdot) = (\boldsymbol{\zeta}_1^{(j\times 1)}(\cdot)\vdots\cdots\vdots\boldsymbol{\zeta}_k^{(j\times 1)}(\cdot))$, with a $\mathcal{GP}$ prior is placed on $\boldsymbol{\zeta}_i(\cdot)$, $i = 1,\ldots,k$, such that the prior on $\boldsymbol{\zeta}_i(\cdot)$ is $(\eta_{i1}(\cdot),\cdots,\eta_{ik}(\cdot))^T$, i.e. $\eta_{it}(\cdot)$ is a $\mathcal{GP}$, $t = 1,\ldots,k$, $i = 1,\ldots,j$. In fact, the prior on this $j \times k$-dimensional random matrix $\boldsymbol{\xi}$ is more conveniently represented as a jk-variate Gaussian Process $\boldsymbol{\eta}(\cdot)$, i.e.

$$\boldsymbol{\eta}(\cdot) \sim \mathcal{N}_{jk}(\boldsymbol{\mu}(\cdot), a(\cdot,\cdot)\boldsymbol{\Omega}), \tag{4.2}$$

where the mean function $\boldsymbol{\mu}(\cdot)$ is given by

$$\mu(\cdot) = \boldsymbol{B}^T\boldsymbol{h}(\cdot), \tag{4.3}$$

with $\boldsymbol{B} = (\boldsymbol{\beta}_{11},\ldots,\boldsymbol{\beta}_{j1},\ldots,\boldsymbol{\beta}_{1k},\ldots,\boldsymbol{\beta}_{jk})$, and for $p = 1,\ldots,j$, $q = 1,\ldots,k$, $\boldsymbol{\beta}_{pq}$ is an m-dimensional column vector. Hence, $\boldsymbol{B}$ is a matrix with m rows and jk columns; $\boldsymbol{h}^{(m\times 1)}(\cdot) = (h_1(\cdot),\ldots,h_m(\cdot))^T$, where, for $i = 1,\ldots,m$, $h_i(\cdot)$ could be any function. For our purpose, setting $m = d+1$, we let $h(\boldsymbol{s}) = (1,\boldsymbol{s})^T$ for all $\boldsymbol{s}$. The motivation for this structure of the mean function is discussed in the Supplementary section **S-2**.

For any two model parameter vector values $\boldsymbol{s}_1, \boldsymbol{s}_2$, the covariance function is chosen to be

$$\begin{aligned} \text{cov}(\boldsymbol{\eta}(\boldsymbol{s}_1),\boldsymbol{\eta}(\boldsymbol{s}_2)) &= a(\boldsymbol{s}_1,\boldsymbol{s}_2)\boldsymbol{\Omega} \\ &= \exp\{-(\boldsymbol{s}_1-\boldsymbol{s}_2)^T\boldsymbol{Q}(\boldsymbol{s}_1-\boldsymbol{s}_2)\}\boldsymbol{\Sigma}\otimes\boldsymbol{C}, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}\otimes\boldsymbol{C}$ and $\boldsymbol{Q}$ is a diagonal matrix consisting of d smoothness parameters denoted by $b_1,\ldots,b_d$, $b_i > 0$, $i = 1,\ldots,d$.

Thus, the matrix-variate $\mathcal{GP}$ prior for $\boldsymbol{\xi}(\cdot)$, which, for any given $\boldsymbol{s}$, has a marginal that is matrix normal with mean matrix $\boldsymbol{\mu}^{(j\times k)}(\mathbf{s})$ - the (p,q)-th element (function) of which is $\mu_{pq}(\cdot) = \boldsymbol{\beta}_{pq}^T\boldsymbol{h}(\cdot)$ - and two covariance matrices $\boldsymbol{C}$ and $\boldsymbol{\Sigma}$.

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 $\boldsymbol{\Sigma}$. The matrix-variate Gaussian Process yields the following correlation structures:

(4.5)
$$corr\left(\eta_{i_1,t}(\boldsymbol{s}),\eta_{i_2,t}(\boldsymbol{s})\right)=\frac{\sigma_{i_1,i_2}}{\sqrt{\sigma_{i_1,i_1}\sigma_{i_2,i_2}}} \text{ for all } t,\boldsymbol{s} \text{ and } i_1\neq i_2$$

(4.6)
$$corr\left(\eta_{i,t_1}(\boldsymbol{s}),\eta_{i,t_2}(\boldsymbol{s})\right)=\frac{c_{t_1,t_2}}{\sqrt{c_{t_1,t_1}c_{t_2,t_2}}} \text{ for all } i,\boldsymbol{s} \text{ and } t_1\neq t_2$$

(4.7)
$$corr\left(\eta_{i,t}(\boldsymbol{s}),\eta_{i_1,t_1}(\boldsymbol{s})\right)=\frac{c_{t,t_1}\sigma_{i,i_1}}{\sqrt{c_{t,t}\sigma_{i,i}c_{t_1,t_1}\sigma_{i_1,i_1}}} \text{ for all } \boldsymbol{s} \text{ and } i\neq i_1, t\neq t_1$$

(4.8)
$$corr\left(\eta_{i,t}(\boldsymbol{s}_1),\eta_{i,t}(\boldsymbol{s}_2)\right)=a(\boldsymbol{s}_1,\boldsymbol{s}_2) \text{ for all } i,t \text{ and } \boldsymbol{s}_1\neq\boldsymbol{s}_2$$

4.1. *The training data set and its distribution.* In order to discuss the method, we treat the velocity information simulated for the choice of $\mathbf{S}=\mathbf{s}_i$ as a $j\times k$-dimensional velocity matrix $\mathbf{V}_i$. However, hereafter, we treat this simulated velocity information as a jk-dimensional velocity vector, $\mathbf{v}_i$, $\forall\, i=1,\ldots,n$. Also, we assume the model parameter vector $\mathbf{S}$ that we attempt to learn, (the solar position in our case), to be d-dimensional, so that $\boldsymbol{s}^T=(s_1,s_2,\ldots,s_d)$ is the unknown model parameter in general. We define the training data matrix $\mathcal{D}_s$ corresponding to the design set $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\}$, under the assumption of zero error in the generated velocities, as

$$\mathcal{D}_s^{(n\times jk)}=\begin{pmatrix}\mathbf{v}_1^T\\ \mathbf{v}_2^T\\ \cdots\\ \mathbf{v}_n^T\end{pmatrix}=\begin{pmatrix}\boldsymbol{\eta}^T(\boldsymbol{s}_1)\\ \boldsymbol{\eta}^T(\boldsymbol{s}_2)\\ \cdots\\ \boldsymbol{\eta}^T(\boldsymbol{s}_n)\end{pmatrix}$$

In the same light, the observed velocity sample $\mathcal{D}_{test}$ is represented as a jk-dimensional velocity vector $\mathbf{v}^{(test)}$. There are 4 distinct training data sets, $\mathcal{D}_s^{(q)}$, $q=1,2,3,4$, corresponding to 4 distinct astrophysical models of the Milky Way.

Writing $\boldsymbol{H}_D^{(n\times m)^T}=[\boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_1),\ldots,\boldsymbol{h}^{(m\times 1)}(\boldsymbol{s}_n)]$ and $\boldsymbol{A}_D^{(n\times n)}=[a(\boldsymbol{s}_r,\boldsymbol{s}_\ell)]$, it follows from the Gaussian Process assumption of $\boldsymbol{\eta}(\cdot)$ that $\mathcal{D}_s$ is matrix normal with mean matrix $\boldsymbol{H}_D\boldsymbol{B}$, left covariance matrix $\boldsymbol{A}_D$ and right covariance matrix $\boldsymbol{\Omega}$.

(4.9)
$$[\mathcal{D}_s\mid\boldsymbol{B},\boldsymbol{\Sigma},\boldsymbol{C},\boldsymbol{Q}]\sim\mathcal{MN}_{n,jk}(\boldsymbol{H}_D\boldsymbol{B},\boldsymbol{A}_D,\boldsymbol{\Omega})$$

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

(4.10)
$$[\mathcal{D}_s\mid\boldsymbol{B},\boldsymbol{\Sigma},\boldsymbol{C},\boldsymbol{Q}]=\frac{1}{(2\pi)^{\frac{njk}{2}}|\boldsymbol{A}_D|^{\frac{jk}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Omega}^{-1}(\mathcal{D}_s-\boldsymbol{H}_D\boldsymbol{B})^T\boldsymbol{A}_D^{-1}(\mathcal{D}_s-\boldsymbol{H}_D\boldsymbol{B})\right]\right\}$$

Then, the r-th row of $[\mathcal{D}_s|\mathbf{B},\mathbf{\Sigma},\mathbf{C},\mathbf{Q}]$ is multivariate normal with mean corresponding row of the mean matrix $\boldsymbol{H}_D\boldsymbol{B}$ and with covariance matrix $\mathbf{\Omega}$. Rows r and ℓ of $[\mathcal{D}_s|\mathbf{B},\mathbf{\Sigma},\mathbf{C},\mathbf{Q}]$ has covariance matrix $a(\boldsymbol{s}_r,\boldsymbol{s}_\ell)\mathbf{\Omega}$. 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 $\mathbf{\Omega}$. The covariance between columns r and ℓ is given by the matrix $\omega_{r,\ell}\boldsymbol{A}_D$.

As we expressed above, the aim of this work is to find the posterior predictive distribution of a new value $\mathbf{s}^{(test)}$ of the model parameter $\mathbf{S}$, given the observed velocity matrix $\mathcal{D}_{test}$ and training data. In fact, in our work, we allow for the learning of the d smoothing parameters (the diagonal elements of $\mathbf{Q}$), as well as the covariance matrix $\mathbf{\Sigma}$, along with $\mathbf{s}^{(test)}$. In fact, in this sense our work stands in contrast to previous work on Gaussian Process-based learning of unknown functions which assume $\boldsymbol{Q}$ to be fixed and typically estimated by maximum likelihood methods; we adopt a fully Bayesian approach by allowing $\boldsymbol{Q}$ to be a random quantity, the posterior of which may be learned if desired. Thus, the posterior predictive distribution $[\mathbf{s}^{(test)},\mathbf{\Sigma},\mathbf{Q}|\mathbf{v}^{(test)},\mathcal{D}_s]$ is sought, by marginalising the posterior $[\mathbf{S},\mathbf{\Sigma},\mathbf{Q},\mathbf{B},\mathbf{C}|\mathbf{v}^{(test)},\mathcal{D}_s]$ over the posteriors of $\mathbf{B}$ and $\mathbf{C}$. In order to achieve this marginalisation, we first learn the posteriors of $\mathbf{B}$ and $\mathbf{C}$.

4.2. *Posteriors of $\boldsymbol{B}$ and $\boldsymbol{C}$ given training data.* We impose a simple non-informative prior $\pi(\boldsymbol{B},\mathbf{\Sigma},\boldsymbol{C},\mathbf{Q}) \propto |\ \mathbf{\Sigma}\ |^{-(k+1)/2}|\ \boldsymbol{C}\ |^{-(j+1)/2}$, to write

$$[\boldsymbol{B}\mid \mathbf{\Sigma},\boldsymbol{C},\boldsymbol{Q},\mathcal{D}_s] \sim \mathcal{MN}_{m,jk}(\hat{\boldsymbol{B}}_{GLS},(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1},\mathbf{\Omega}), \tag{4.11}$$

where, $\hat{\boldsymbol{B}}_{GLS} := (\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1}(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\mathcal{D}_s)$; note that this can be interpreted as the generalised least square (GLS) estimate.

We define $\boldsymbol{M}_D := \boldsymbol{A}_D^{-1}-\boldsymbol{A}_D^{-1}\boldsymbol{H}_D(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1}\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}$ and let $\mathcal{D}_s^T\boldsymbol{M}_D\mathcal{D}_s = [\boldsymbol{M}_{it}^*; i,t=1,\ldots,k]$, where $\boldsymbol{M}_{it}^*$ is a matrix with j rows and j columns. Given $\mathbf{\Sigma}$, let $(n-m)k\hat{\boldsymbol{C}}_{GLS} = \sum_{i=1}^k\sum_{t=1}^k \psi_{it}^{-1}\boldsymbol{M}_{ti}^*$, where $m = d+1$ and ψ_{it}^{-1} is the (i,t)-th element of $\mathbf{\Sigma}^{-1}$. Then it can be shown that the posterior distribution $[\boldsymbol{C}|\mathbf{\Sigma},\boldsymbol{Q},\mathcal{D}_s]$ is inverse Wishart with parameters $(n-m)k\hat{\boldsymbol{C}}_{GLS}$ and $(n-m)k$ (degrees of freedom).

Since prediction of $\boldsymbol{s}^{(test)}$, given the test and the training data sets, is of interest, this constitutes an inverse problem. The Bayesian solutions of this inverse problem are encapsulated in the posterior distribution $[\boldsymbol{s}^{(test)}\mid \mathbf{v}^{(test)},\mathcal{D}_s]$.

To construct an expression for the posterior distribution of $\boldsymbol{s}^{(test)}$, we construct an augmented data set $\mathcal{D}_{aug}^T = ((\mathbf{v}^{(test)})^T\vdots\mathcal{D}_s^T)$. Then, assuming the aforementioned non-informative prior for $(\boldsymbol{B},\mathbf{\Sigma},\boldsymbol{C},\boldsymbol{Q})$, we observe that

$$
\begin{aligned}
&[\boldsymbol{s}^{(test)}, \boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C} \mid \mathbf{v}^{(test)}, \mathcal{D}_s] \propto [\mathcal{D}_{aug} \mid \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}][\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}] \\
&\propto \; |\boldsymbol{A}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{\Sigma}|^{-\frac{j(n+1)+k+1}{2}} |\boldsymbol{C}|^{-\frac{k(n+1)+j+1}{2}} \\
&\times \; \exp\left[-\frac{1}{2} tr\left\{\boldsymbol{\Omega}^{-1}(\mathcal{D}_{aug} - \boldsymbol{H}_{\mathcal{D}_{aug}}\boldsymbol{B})^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}(\mathcal{D}_{aug} - \boldsymbol{H}_{\mathcal{D}_{aug}}\boldsymbol{B})\right\}\right] \\
&= \; |\boldsymbol{A}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{\Sigma}|^{-\frac{j(n+1)+k+1}{2}} |\boldsymbol{C}|^{-\frac{j(n+1)+k+1}{2}} \\
&\times \; \exp\left[-\frac{1}{2} tr\left\{\boldsymbol{\Omega}^{-1}\mathcal{D}_{aug}^T \boldsymbol{M}_{aug}\mathcal{D}_{aug}\right\}\right] \\
&\times \; \exp\left[-\frac{1}{2} tr\left\{\boldsymbol{\Omega}^{-1}\left\{\boldsymbol{B} - (\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}})^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\mathcal{D}_{aug}\right\}^T\right.\right.\right. \\
&\qquad \left.\left.\left. \boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}\left\{\boldsymbol{B} - (\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}^T)^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\mathcal{D}_{aug}\right\}\right\}\right] \qquad (4.12)
\end{aligned}
$$

where we have recalled Equation 4.10.

In all the quantities above, the suffix aug denotes operation with respect/corresponding to the augmented matrix $\mathcal{D}_{aug}$. Marginalising $[\boldsymbol{s}^{(test)}, \boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C} \mid \mathbf{v}^{(test)}, \mathcal{D}_s]$ over $\boldsymbol{B}$ and $\boldsymbol{C}$ yields the posterior $[\boldsymbol{s}^{(test)}, \boldsymbol{Q} \mid \mathbf{v}^{(test)}, \boldsymbol{\Sigma}, \mathcal{D}_s]$, the form of which is obtained as follows:

$$
\begin{aligned}
&[\boldsymbol{s}^{(test)}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathbf{v}^{(test)}, \mathcal{D}_s] \\
&\qquad = \; \int\int [\boldsymbol{s}^{(test)}, \boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \mid \mathbf{v}^{(test)}, \mathcal{D}_s] d\boldsymbol{B} d\boldsymbol{C} \\
&\qquad \propto \; |\boldsymbol{A}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{\Sigma}|^{-\frac{j(n+1-m)+k+1}{2}} \\
&\qquad \times \; \int |\boldsymbol{C}|^{-\frac{k(n+1-m)+j+1}{2}} \exp\left[-\frac{1}{2} tr\left\{\boldsymbol{\Omega}^{-1}\mathcal{D}_{aug}^T \boldsymbol{M}_{aug}\mathcal{D}_{aug}\right\}\right] d\boldsymbol{C} \qquad (4.13) \\
&\qquad = \; |\boldsymbol{A}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}} \\
&\qquad \times \; |\boldsymbol{\Sigma}|^{-\frac{j(n+1-m)+k+1}{2}} |(n+1-m)k\hat{\boldsymbol{C}}_{GLS,aug}|^{-\frac{(n+1-m)k}{2}} \qquad (4.14)
\end{aligned}
$$

Equation (4.13) follows from the definition of matrix normal distribution, noting that

$$
\begin{aligned}
&\int \exp\left[-\frac{1}{2} tr\left\{\boldsymbol{\Omega}^{-1}\left\{\boldsymbol{B} - (\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}})^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\mathcal{D}_{aug}\right\}^T\right.\right.\right. \\
&\left.\left.\left. \boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}\left\{\boldsymbol{B} - (\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}})^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\mathrm{D}_{aug}\right\}\right\}\right] d\boldsymbol{B} \\
&\propto \; |\boldsymbol{\Omega}|^{\frac{m}{2}} |\boldsymbol{H}_{\mathcal{D}_{aug}}^T \boldsymbol{A}_{\mathcal{D}_{aug}}^{-1}\boldsymbol{H}_{\mathcal{D}_{aug}}|^{-\frac{jk}{2}}
\end{aligned}
$$

Equation (4.14) follows from the definition of inverse Wishart distribution, noting that

$$
\begin{aligned}
&\int |\boldsymbol{C}|^{-\frac{(n+1-m)k+j+1}{2}} \exp\left[-\frac{1}{2} tr\left\{\boldsymbol{\Omega}^{-1}\mathcal{D}_{aug}^T \boldsymbol{M}_{aug}\mathcal{D}_{aug}\right\}\right] d\boldsymbol{C} \\
&\propto \; |(n+1-m)k\hat{\boldsymbol{C}}_{GLS,aug}|^{-\frac{(n+1-m)k}{2}}
\end{aligned}
$$

We thus obtain, the joint posterior of $(\boldsymbol{s}^{(test)}, \boldsymbol{Q}, \boldsymbol{\Sigma})$ as given by (4.14), up to a normalising constant. We sample from this posterior using MCMC techniques - in fact, we implement the transformation-based MCMC method recently by Dutta and Bhattacharya (2011).

In our application, the errors in the measurement of the stellar velocities 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 stellar velocity 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.

5. Posterior inference using TMCMC.

5.1. *Overview of TMCMC.* 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 (2011) proposed a novel methodology that works by constructing proposals that are deterministic bijective transformations of some arbitrary random vector drawn from a chosen distribution. The random vector can be chosen to be of dimensionality between 1 and the dimensionality of the parameters under the target posterior. Thus, high-dimensional parameter spaces can be explored by constructing bijective deterministic transformations of a low-dimensional random vector. Interestingly, whatever the transformation, the acceptance ratio in TMCMC does not depend upon the distribution of the chosen random vector.

Indeed, Dutta and Bhattacharya (2011) show that for additive transformations, the TMCMC-based acceptance rate decreases at a slower rate compared to random walk Metropolis algorithms. Furthermore, TMCMC includes the hybrid Monte Carlo (HMC) algorithm as a special case and in one-dimensional situations, 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 (2011) show that these transformations require far less number of "move types" compared to non-additive transformations and do not require calculation of the Jacobian, thus ensuring computational superiority over other transformations.

5.2. *TMCMC based on additive transformations.* Our implementation of additive TMCMC is as follows:

(i) Initialise the unknown quantities by fixing arbitrarily initial values $\left(\boldsymbol{s}^{(test,0)}, \boldsymbol{Q}^{(0)}, \boldsymbol{\Sigma}^{(0)}\right)$. In our case, $\boldsymbol{s}^{(test,0)} = (s_1^{(test,0)}, \ldots, s_d^{(test,0)})$, $\boldsymbol{Q}^{(0)}$ is characterised by the

initial values of the d smoothness parameters, which we denote by $(b_1^{(0)}, \ldots, b_d^{(0)})$ and $\boldsymbol{\Sigma}^{(0)}$ denotes the initial choice of the $k \times k$ matrix $\boldsymbol{\Sigma}$.

(ii) Assume that at iteration t, the state of the unknown parameters is $(\boldsymbol{s}^{(test,t)}, \boldsymbol{Q}^{(t)}, \boldsymbol{\Sigma}^{(t)})$.

(iii) (a) 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. Update the d components of location $\boldsymbol{s}^{(test)}$ by setting, with probabilities π_j and $(1-\pi_j)$, $s_j^{(test,t+1)} = s^{(test,t)} + c_j \epsilon$ (forward transformation) and $\mathbf{s}_j^{(test,t+1)} = \mathbf{s}_j^{(test,t)} - c_j \epsilon$ (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 $j_1 \in \mathcal{S}$, $\mathbf{s}_{j_1}^{(test,t)}$ gets the positive transformation, while for $j_2 \in \mathcal{S}^c$, $\mathbf{s}_{j_2}^{(test,t)}$ gets the backward transformation. Here $\mathcal{S} \cup \mathcal{S}^c = \{1, \ldots, d\}$.

(b) Accept the new proposal $\boldsymbol{s}^{(test,t+1)}$ with acceptance probability

$$\alpha_s = \min \left\{ 1, \frac{\prod_{j_1 \in \mathcal{S}} (1 - \pi_{j_1}) \prod_{j_2 \in \mathcal{S}^c} \pi_{j_2}}{\prod_{j_1 \in \mathcal{S}} \pi_{j_1} \prod_{j_2 \in \mathcal{S}^c} (1 - \pi_{j_2})} \times r_s \right\} \tag{5.1}$$

where r_s denotes the ratio of $|\boldsymbol{A}_{D_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{H}'_{D_{aug}} \boldsymbol{A}^{-1}_{D_{aug}} \boldsymbol{H}_{D_{aug}}|^{-\frac{jk}{2}} \times |\boldsymbol{\Sigma}|^{-\frac{k(n+1-m)+k+1}{2}} |(n+1-m) k \hat{\boldsymbol{C}}_{GLS,aug}|^{-\frac{(n+1-m)k}{2}}$, evaluated at the new value ($\boldsymbol{s}^{(test,t+1)}$) and the current value ($\boldsymbol{s}^{(test,t)}$) of $\boldsymbol{s}^{(test)}$ respectively, with $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$ remaining fixed at the current values $\boldsymbol{b}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$, respectively. In our applications, we shall choose $\pi_j = 1/2 \; \forall j$, which simplifies the acceptance ratio by eliminating the first factor altogether.

(c) If the new proposal $\boldsymbol{s}^{(test,t+1)}$ is not accepted, we set $\boldsymbol{s}^{(test,t+1)} = \boldsymbol{s}^{(test,t)}$.

(iv) Using the same TMCMC strategy, update $\boldsymbol{b}$. Note that the probability of acceptance is zero if some components of the proposed value of $\boldsymbol{b}$ is negative.

(v) Update $\boldsymbol{\Sigma}$ by first decomposing it into $\boldsymbol{C}\boldsymbol{C}'$, where $\boldsymbol{C}$ is the appropriate lower-triangular matrix. Then update all the non-zero elements of $\boldsymbol{C}$ in a single block using the same TMCMC strategy. The form of the acceptance ratio remains same. Again, reject the moves that yield negative diagonal elements.

(vi) Cycle over steps (ii)–(v) for $K_1 + K_2$ iterations, assuming that the Markov chain converges after K_1 iterations. Store the realisations $\left\{ \left(\boldsymbol{s}^{(test,t)}, \boldsymbol{Q}^{(t)}, \boldsymbol{\Sigma}^{(t)} \right) : t = K_1 + 1, \ldots, K_2 \right\}$ as samples obtained (approximately) from $[\boldsymbol{s}^{(test)}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_{aug}]$. In particular, the realisations $\{\boldsymbol{s}^{(test,t)} : t = K_1 + 1, \ldots, K_2\}$ are samples (approximately) from the marginal distribution $[\boldsymbol{s}^{(test)} \mid \boldsymbol{V}^{(test)}, \mathcal{D}_{aug}]$.

In the supplementary section **S-4**, we present a method that can be used to validate our models, using a leave-one-out cross-validation technique implemented within the proposed Importance Resampling MCMC framework.

6. Analysis of real data. In our case, $\mathbf{S}$ is a 2-dimensional vector, i.e. d=2. Also, $\mathbf{P}$ is a two-dimensional vector, i.e. k=2. Lastly, we consider 50 stellar vectors at any $\mathbf{s}$, i.e. j=50.

For each of the 4 astrophysical models of the Milky Way, we construct a training dataset that comprises 216 independent values of the model parameter vector (solar position vector). For each $\mathbf{s}$, 50 2-component stellar velocity vectors are randomly drawn from the velocity data that are generated at that $\mathbf{s}$ from dynamical simulations performed using the given astrophysical model of the Milky Way. In other words, for each $\mathbf{s}$, a 50×2-dimensional velocity matrix is set up. The test dataset contains a single observed 50×2-dimensional velocity matrix (that is, we considered 50 stars, each with an observed two-component velocity vector), corresponding to an unknown location that we need to predict, using our matrix-variate Gaussian Process based method.

It will be seen in Section 6.2 that, in spite of our relatively smaller data sets, the summary of the posterior distributions of the unknown solar position concur with existing knowledge in the astrophysics literature. In this connection it is important to discuss some issues of data-sufficiency arising in our Gaussian Process-based Bayesian approach. Note that since the predicted function must pass through all the points of the training data set, $\mathbf{\Sigma}$ must be close to the null matrix *a posteriori* (the diagonal elements must be close to zero *a posteriori* signifying that the individual functions are almost deterministic, and the covariances between any two such almost deterministic functions must be close to zero), if the input (here solar position vector) design and the number of design points are adequately chosen, hereby indicating the importance of the design and the number of design points for inference about $\mathbf{\Sigma}$. However, increasing the number of stars does not contribute to the design and need not appreciably enhance inference about $\mathbf{\Sigma}$.

Also, since the smoothness parameters determine the smoothness of the function with respect to the sample path, design points being the discretised version, information about the smoothness parameters is clearly more enhanced by increasing the number of design points rather than the number of stars. These arguments show that it makes more sense to increase the number of design points than the number of stars.

Understanding these data-sufficiency issues turns out to be very rewarding from the computational point of view – increasing the number of stars implies increasing the dimension of the matrix $\boldsymbol{C}$, which would have made our computation prohibitively slow, but keeping the dimension of $\boldsymbol{C}$ at some reasonable level, and increasing the number of data points is computationally much less expensive.

In our case, expected results regarding the unknown location are obtained by our approach with 50×2 velocity functions, which are of reasonable dimensionality, and 216 well-chosen design points. Our results, which are astrophysically interpretable, demonstrate that these choices are adequate.

6.1. *TMCMC details.* The general working of the TMCMC algorithm provided in Section 5.2 is modified in our application, to ensure that all unknowns are updated at once, using additive transformations of a one-dimensional random

TABLE 1
Summary of the posterior distributions of the unknown location $(R_\odot, \Phi_\odot)$ for the 4 models.

Model	R		
	Mode	95% HPD	50% HPD
$bar6$	2.20	$[2.04, 2.30]$	$[2.16, 2.24]$
$sp3bar3$	1.73	$[1.70, 2.26] \cup [2.27, 2.28]$	$[1.71, 1.79] \cup [1.96, 1.97] \cup [1.99, 2.05] \cup [2.10, 2.21]$
$sp3bar3\_18$	1.76	$[1.70, 2.29]$	$[1.72, 1.86] \cup [1.98, 2.09]$
$sp3bar3\_25$	1.95	$[1.70, 2.15]$	$[1.86, 1.98]$

variable. This, along with adequate choice of the scale parameters, ensure fast computations and reasonable convergence.

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 1,00,000, starting with arbitrary initial values and guesses of the scale parameters; 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 following proposal mechanism: at each iteration we generated $\epsilon \sim \mathcal{N}(0,1)I_{\{\epsilon>0\}}$. Then, with equal probabilities, for $i = 1, 2$, we set $s_i^{(test,t)} = s_i^{(test,t-1)} \pm \epsilon$ and $b_i^{(t)} = b_i^{(t-1)} \pm 0.007\epsilon$. For updating $\boldsymbol{\Sigma}$, we decompose $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}'$ as before, but here we set the scale factor to be 0.07 for all the non-zero elements, using the additive transformation with equal move-type probabilities.

Once the proposal mechanism and the initial values are thus decided, we then discarded the next 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.

6.2. *Results from real data.* For prediction of the unknown location $\boldsymbol{s}$ we use the prior information available in the astrophysical literature (Chakrabarty, 2007) for the corresponding polar co-ordinates $S_1 \in [1.7, 2.3]$ and $S_2 \in [0, 90]$ (in degrees). It is worth noting that the locations associated with the training data sets satisfy these restrictions. Accordingly, in our TMCMC algorithm we reject those moves that fail to satisfy the above constraints.

The marginal posterior densities of $(R_\odot, \Phi_\odot)$ corresponding to the 4 astrophysical models of the Milky Way, are shown in Figures 1, 2, 3 and 4. Tables 1 and 2 present the posterior mode, the 95% and the 50% highest posterior density (HPD) credible region of R and θ respectively, associated with the four models. The HPDs are computed using the methodology discussed in Carlin and Louis (1996). Disjoint HPD regions, particularly in the case of 50% HPD level, characterise the highly multi-modal posterior distributions of the unknown location.

Summaries of the posteriors (mean, variance and 95% credible interval) of the smoothness parameters b_1, b_2 and $\boldsymbol{\Sigma}$ are presented in Tables 3, 4, 5 and 6. Notable in all these tables are the small posterior variances of the quantities in question;

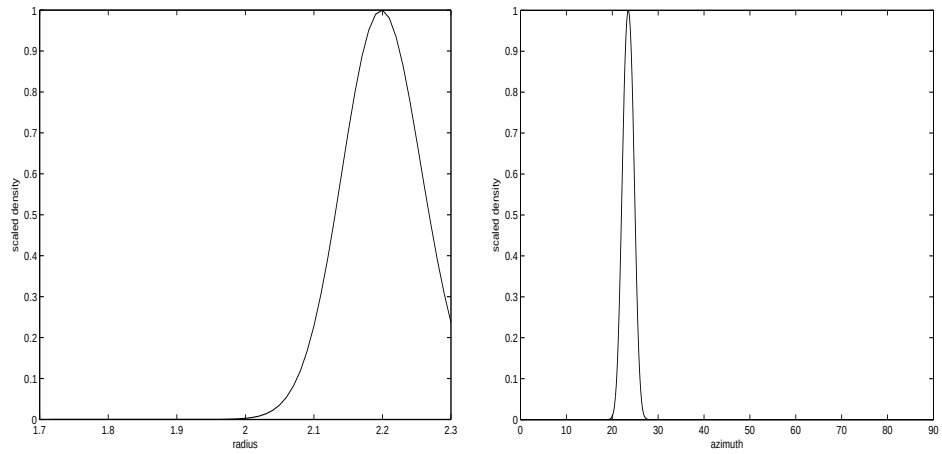

FIG 1. *Posteriors of R and θ for the model $bar6$.*

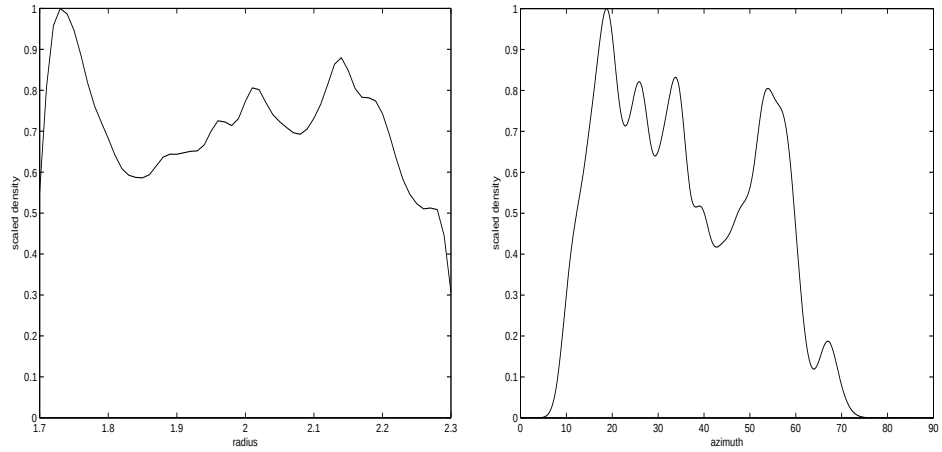

FIG 2. *Posteriors of R and θ for the model $sp3bar3$.*

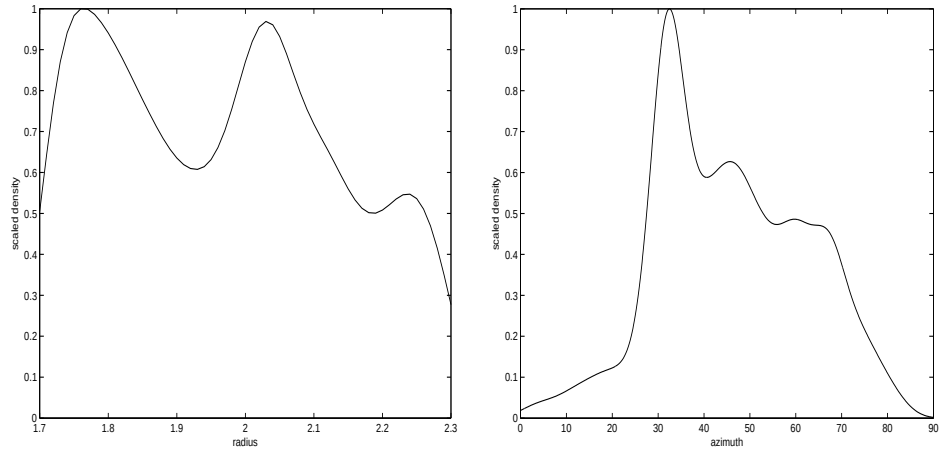

FIG 3. *Posteriors of R and θ for the model $sp3bar3\_18$.*

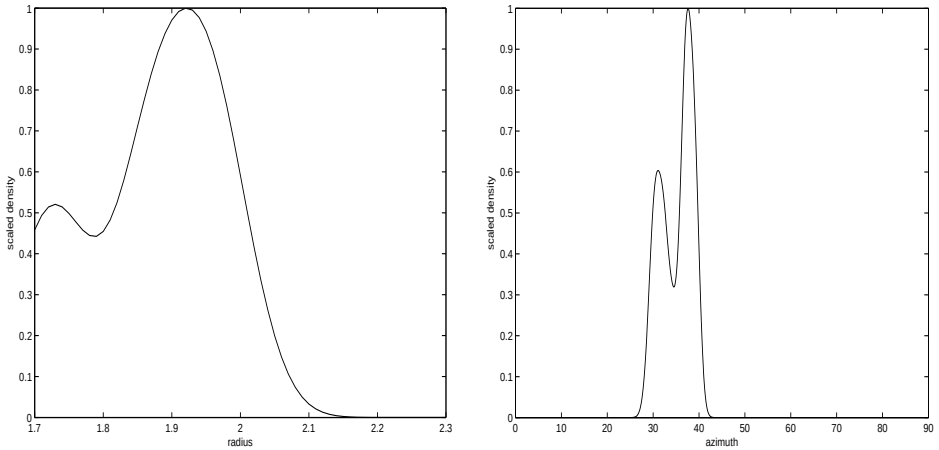

FIG 4. *Posteriors of R and θ for the model $sp3bar3\_25$.*

TABLE 2
Summary of the posterior distributions of the azimuthal component $\Phi_\odot$, of the unknown observer location vector $(R_\odot \;\; \Phi_\odot)^T$ for the 4 models.

Model	θ		
	Mode	95% HPD	50% HPD
$bar6$	23.50	$[21.20, 25.80]$	$[22.60, 24.30]$
$sp3bar3$	18.8	$[9.6, 61.5]$	$[15.10, 22.50] \cup [23.20, 27.80] \cup [31.30, 35.50] \cup [52.00, 57.80]$
$sp3bar3\_18$	32.5	$[17.60, 79.90]$	$[27.9, 49.9]$
$sp3bar3\_25$	37.6	$[28.80, 40.40]$	$[30.70, 31.50] \cup [36.00, 39.60]$

TABLE 3
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
$bar6$	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 4
Summary of the posterior distribution of the first diagonal element of $\mathbf{\Sigma}$ for the 4 models.

Model	σ_{11}		
	Mean	Var	95% CI
$bar6$	1.80×10^{-4}	1.21×10^{-8}	$[5.40 \times 10^{-5}, 4.0 \times 10^{-4}]$
$sp3bar3$	5.53×10^{-3}	3.15×10^{-6}	$[3.66 \times 10^{-3}, 1.03 \times 10^{-2}]$
$sp3bar3\_18$	2.37×10^{-2}	1.49×10^{-3}	$[1.45 \times 10^{-3}, 1.68 \times 10^{-1}]$
$sp3bar3\_25$	3.00×10^{-4}	1.50×10^{-8}	$[1.21 \times 10^{-4}, 5.69 \times 10^{-4}]$

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 matrix-variate $\mathcal{GP}$-based Bayesian approach.

The posterior means (the small variances imply that mean, median, mode are very much the same value) show that the velocity function corresponding to model $sp3bar3\_18$ is the smoothest, followed, in order, by $sp3bar3$, $sp3bar3\_18$, and $bar6$. In each case b_2 has a higher posterior mean than b_1, the former being almost equal to one.

As discussed in the beginning of Section 6, thanks 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 and the number of design points are adequate. Quite encouragingly, tables 4, 5 and 6 show 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.

6.3. *Comparison with results in astrophysical literature.* The posterior summaries that we have presented in the last section, for each of the 4 astrophysical models of the Milky Way, compare favourably with results obtained by Chakrabarty

TABLE 5
Summary of the posterior distribution of the second diagonal element of $\mathbf{\Sigma}$ for the 4 models.

Model	σ_{22}		
	Mean	Var	95% CI
$bar6$	2.14×10^{-4}	1.67×10^{-8}	$[6.20 \times 10^{-5}, 4.76 \times 10^{-4}]$
$sp3bar3$	9.87×10^{-3}	1.01×10^{-5}	$[6.53 \times 10^{-3}, 1.83 \times 10^{-2}]$
$sp3bar3\_18$	2.13×10^{-2}	1.20×10^{-3}	$[1.29 \times 10^{-3}, 1.50 \times 10^{-1}]$
$sp3bar3\_25$	2.75×10^{-4}	1.25×10^{-8}	$[1.13 \times 10^{-4}, 5.21 \times 10^{-4}]$

TABLE 6
Summary of the posterior distribution of the off-diagonal element of $\mathbf{\Sigma}$ for the 4 models.

Model	σ_{12}		
	Mean	Var	95% CI
$bar6$	3.86×10^{-6}	1.20×10^{-11}	$[0, 1.30 \times 10^{-5}]$
$sp3bar3$	7.98×10^{-5}	6.89×10^{-9}	$[-6.40 \times 10^{-5}, 2.68 \times 10^{-4}]$
$sp3bar3\_18$	2.86×10^{-4}	4.18×10^{-7}	$[-1.19 \times 10^{-4}, 2.16 \times 10^{-3}]$
$sp3bar3\_25$	5.08×10^{-6}	1.65×10^{-11}	$[-1.00 \times 10^{-6}, 1.50 \times 10^{-5}]$

(2007), using the calibration method. However, the all important difference in our implementation is the vastly smaller data set that we needed to invoke, in order to achieve the learning of the two-dimensional vector $\mathbf{S}$ - in fact while in the calibration approach, 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 $R_\odot \in [1.9375, 2.21]$ while $\Phi_\odot \in [0°, 30°]$, where these ranges correspond to the presented uncertainties on the estimates. The values of the components of $\mathbf{S}$, learnt in our work, overlap well with these results. As mentioned in Section 1, learning $R_\odot$ and $\Phi_\odot$ have crucial astronomical significance for the azimuthal separation of the Sun from the long axis of the bar and for the rotational frequency of the bar in the Milky Way. That the models $sp3bar3\_18$, $sp3bar3$ and $sp3bar3\_25$ are distinguished by distinct values of the ratios of the rotational frequencies of the spiral pattern to the bar in the Galaxy, the derived estimate for the bar frequency in turn suggests a possible values of the frequency of the Milky Way spiral.

However, the concurrence of our results with the results reported in astrophysical literature 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 $\mathbf{S}$ that we advance. Chakrabarty and Sideris (2008) suggest minimum chaos in the $bar6$ 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 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 $\mathbf{S}$) that are multi-modal, though less so, than that for the model $sp3bar3$.

The equivalent in the work by Chakrabarty (2007), of our multi-modal posterior distributions (of $R_\odot$ and $\Phi_\odot$) is the distribution of the p-value of the chosen test

statistic that was used by Chakrabarty (2007) to test for the null that the simulated velocity data matrix $\mathbf{V}_t$ is drawn from the pdf of the velocity space that is estimated from the observed data, for $t = 1, \ldots, n$. When the p-value distribution was sought in the relevant space of $\mathbf{S}$, the distribution was indeed found to be most multi-modal for the $sp3bar3$ case while the modes were found to be less scattered than for the other three models. In line with this result, in our work we notice that the model agrees best with the observed and simulated data, when the simulated data are generated at scattered values of the observer location vector $\mathbf{S}$. The exact physical reason for the correlation between chaos in the models and the multi-modality of the posterior distribution is suggested in Chakrabarty (2007) and here we focus upon the concurrence of our results and what is already published in the astrophysics domain.

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.

The implication of our demonstrated success with a smaller data set is profound. Firstly, our method is advanced as a template for the analysis of the stellar phase space 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. At the same time, the planned flight of the GAIA - a mission of the European Space Agency - later in 2012, is set to provide large sets of phase space data all over the Milky Way. Our method, in conjunction with astronomical models, can allow for fast learning of local and global model parameters of the Galaxy. In fact, the greater strength of our method lies in its proposed applicability to galaxies other than our own, for which, only small data sets ($\lesssim 100$) can be expected. The learning of global model parameters of such systems is possible with a method such as ours, once data generated from simulations of such systems are available.

7. Discussions. We have presented a new method to allow for inverse Bayesian learning of model parameter vectors, given matrix-valued data, using the matrix-variate Gaussian Process. The biggest advantage of our method is that it can work even when data sufficiency might be a concern under other methods. The learning is rendered possible, owing to the matrix-variate Gaussian Process that we introduce here and demonstrate to be successful towards such inverse learning.

It is to be noted that by avoiding a direct inference on the inverse of the velocity function, given the inferential approach that we adopt, we have managed to avoid problems that may arise out of lack of consistency between the learnt values of $\mathbf{S}$, as we scan across the measured velocity sample - in fact, in such a case, the uncertainty in our learnt value of $\mathbf{S}$ would potentially increase with the size

of the velocity data sample. Our inferred uncertainty on the learnt parameters, is on the other hand, a signature of the dispersion in the chain upon convergence. Had we included velocity measurement errors in our analysis, this too would have contributed towards the inferred credible region. However, in this application, the errors in the velocity measurements were low and therefore neglected, though we suggest (in Section 4.2), how measurement errors could be incorporated.

Following on from what we said in the introduction, the Milky Way model parameter vector $\mathbf{S}$, that we seek to learn, can be high-dimensional. In the introductory section, we suggested that based on astrophysical considerations, the dimensionality d of $\mathbf{S}$ could be high - at least 8. As long as adequate training data exists - velocity data generated from dynamical simulations of the Milky Way that use distinct choices of each of the d components of $\mathbf{S}$ - the learning of $\mathbf{S}$ is possible. A strength of the used method is that increasing d even 4-folds would add not increase the run time drastically, given the velocity information. In fact, it is the $\mathbf{A}_{\mathcal{D}_{aug}}$ above that will be affected by the increase in the value of d; the computational complexity involved in the computation of this matrix scales linearly with d. Thus, a 4-fold increase in d is still very much within tractable time frames. This is a crucial advantage of the method and allows for expansion of its application to several other contexts.

In contrast to the situation with increasing the dimensionality of the unknown model parameter, increasing the dimensionality of the measurable will 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 measurable variable $\mathbf{P}$. This is because of the dimensionality of the aforementioned $\mathbf{\Sigma}$ matrix is $k \times k$, and the inverse of this enters the computation of the posterior. Thus, for example, increasing the measurable from a 2-dimensional to 4-dimensional vector 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 the choice of fixed values of the same.

Inverse learning similar to our application, is essentially of wide applicability - for example, one might need to learn the required dosage of a drug in order to achieve a chosen efficacy, using available data from a trial (the training data). Alternatively, the price of a commodity might need to be fixed, given a parametrisation of the current demand for the item, guided by training data comprising such a demand parameter and the price, at time points in the past. In another astrophysical context, Chakrabarty and Jackson (2009) learnt the distribution of the density of the total (i.e. luminous+dark) gravitational matter $\rho : \mathbb{R}^3 \longrightarrow \mathbb{R}$, $\rho(x, y, z) \geq 0$, in distant galaxies belonging to an ubiquitous astronomical class of real galactic systems which is parametrised by the shape parameter $n \in \mathbb{R}$ and a length scale parameter $X_e \in \mathbb{R}$, $n, X_e \geq 0$. The measurements include the observed galactic

image and central velocity dispersion measurement in the galaxy, while training data comprising synthetic image data and central velocity dispersion, generated from simulations of such systems carried out at chosen values of n, X_e and other relevant model variables. In such applications, as the measurables are rendered vectors and the data a matrix, inverse learning is possible using the method that we present here.

As more Galactic simulations spanning a greater range of model parameters become available, the learning of greater details of the features of the Milky Way will become possible. That our method allows for such learning even for under-abundant data, 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 velocity data sets that are then available, compared to the situation in our Galaxy.

Acknowledgments. We thank Prof. Ayanendranath Basu, Prof. Smarajit Bose, Mr. Pronoy K. Mondal, and Mr. Sayan Patra for helpful discussions. That the matrix $\hat{C}_{GLS}$ is positive definite has been proved formally by Mr. Pronoy Mondal. DC acknowledges the support of a Warwick Centre for Analytical Sciences Fellowship.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian Inverse Learning of Milky Way Parameters Using Matrix-Variate Gaussian Process-based Method"
(http://lib.stat.cmu.edu/aoas/???/???). Section S-1 discusses the details of the dynamical simulations that lead to the training data set used in our supervised learning of the Milky Way model parameters. In Section S-2, some background on Gaussian Processes in presented. Section S-3 contains details of the forward problem as well as a proof that the matrices $(n-m)k\hat{C}_{GLS}$ and $(n-m)kC^{*}_{GLS}$ are positive definite. In Section S-4, a detailed leave-out cross-validation exercise with simulation experiments is discussed.

References.

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.

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

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. (2011). Galactic Phase Spaces. In *Proceedings of the 7th Workshop on Data Analysis in Astronomy–Science: IMAGe IN AcTION Ettore Majorana Foundation and Centre for Scientific Culture, Erice, April 15-22, 2011*.

CHAKRABARTY, D. and JACKSON, B. (2009). Total Mass Distributions of Sersic Galaxies from Photometry & Central Velocity Dispersion. *Astronomy & Astrophysics* **498** 615.

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

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

DUTTA, S. and BHATTACHARYA, S. (2011). Markov Chain Monte Carlo Based on Deterministic Transformations. Submitted, available at http://arxiv.org/abs/1106.5850.

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.

FRANCIS, C. and ANDERSON, E. (2009). Calculation of the local standard of rest from 20 574 local stars in the New Hipparcos Reduction with known radial velocities. *New Astronomy* **14** 615-629.

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

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.

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.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY CV4 7AL, U.K.
d.chakrabarty@warwick.ac.uk

INDIAN STATISTICAL INSTITUTE
203, B. T. ROAD,
KOLKATA 700108
INDIA
munmun.biswas08@gmail.com

BAYESIAN & INTERDISCIPLINARY RESEARCH INSTITUTE
INDIAN STATISTICAL INSTITUTE
203, B. T. ROAD,
KOLKATA 700108
INDIA
sourabh@isical.ac.in

SUPPLEMENT TO "BAYESIAN INVERSE LEARNING OF MILKY WAY MODEL PARAMETERS USING MATRIX-VARIATE GAUSSIAN PROCESS-BASED METHOD"

BY DALIA CHAKRABARTY[*,§], MUNMUN BISWAS[†,¶], SOURABH BHATTACHRYA BHATTACHARYA[‡,¶],

University of Warwick[§], Indian Statistical Institute, Kolkata[¶]

Throughout, we refer to our main manuscript Chakrabarty, Biswas and Bhattacharya (2012) as CBB.

S-1. Details of dynamical simulations of astrophysical models. In Chakrabarty (2007), the modelling involves the following. A sample of phase space coordinates $\{\mathbf{w_0}\}$, is drawn from a chosen (to mimic real disc galaxies') phase space density $g(\mathbf{w})$ at $t = 0$, and is evolved in a (chosen) parametric Galactic gravitational potential $\Psi : \mathbb{R}^3 \times \mathcal{T} \longrightarrow \mathbb{R}$, where the time variable $T \in \mathcal{T} \subset \mathbb{R}_{\geq 0}$. $\Psi(\mathbf{X}, t)$ is chosen to emulate a realistic background Galactic potential $\Psi_0(\mathbf{x}) \in \mathbb{R}$, perturbed by the chosen parametric forms (rigidly rotating, quadrupolar) of the gravitational potential of the bar ($\Psi_b(\mathbf{x}) \in \mathbb{R}$) and the (logarithmic spiral) gravitational potential of the spiral pattern ($\Psi_s(\mathbf{x}) \in \mathbb{R}$) in the Milky Way. The perturbation strengths of these features ($\varepsilon_b : \mathcal{T} \longrightarrow \mathbb{R}$ and $\varepsilon_s : \mathcal{T} \longrightarrow \mathbb{R}$) are slowly increased, to saturation values over time $T_s \in \mathbb{R}$, $T_s > 0$ which is chosen such that $T_s \ll T_{sim}$, where $T_{sim} \in \mathbb{R}$ is the total simulation time. Thus, in the q^{th} astrophysical model, $\Psi^{(q)}(\mathbf{x}, t) = \Psi_0^{(q)}(\mathbf{x}) + \varepsilon_b(t)\Psi_b^{(q)}(\mathbf{x}, t) + \varepsilon_s(t)\Psi_s^{(q)}(\mathbf{x}, t)$, for each $q = 1, \ldots, 4$. At $T = T_{sim}$, orbits are sampled in phase and recorded stroboscopically in the rotating frame of the bar at times when $(\Omega_b - \Omega_s)t$=0.

In order to sort the orbits by the value of the solar location in a galactocentric coordinate frame, chosen ranges of $R \in [1.7, 2.3]$ in simulation units and $\Phi \in [0, 90]$ in degrees, are used to define a regular, 2-D polar grid with $N_R \times N_\phi$ cells of radial and azimuthal widths of δ_R and δ_ϕ, $\delta_R,\ \delta_\phi \in \mathbb{R}$. The i^{th} cell in this grid represents the i^{th} value of the model parameter vector, $\mathbf{S}^{(i)}$, with $n = N_R \times N_\phi$, $i = 1, \ldots, N_R \times N_\phi$. In order to find stars that end up in the simulations with spatial coordinates within the interval around $\mathbf{s}^{(i)}$ that defines the i^{th}-cell, we identify stars with radius $\in [s_1^{(i)}, s_1^{(i)} + \delta_1)$ and azimuth $\in [s_2^{(i)}, s_2^{(i)} + \delta_2)$; these stars comprise the simulated velocity information $\mathbf{V}_i$ in the i^{th}-cell. This is performed for each i, $i = 1, \ldots, n$. The same 2-D polar grid is used for each of the Milky Way astrophysical models.

[*]Research fellow at Department of Statistics, University of Warwick; supported by a Warwick Centre for Analytical Sciences Research Fellowship.

[†]Munmun Biswas is a PhD student in Statistics and Mathematics Unit, Indian Statistical Institute.

[‡]Sourabh Bhattacharya is an Assistant Professor in Bayesian and Interdisciplinary Research Unit, Indian Statistical Institute.

S-2. Some background material. In our work we impose a Gaussian Process prior $\eta(\mathbf{s})$ on the unknown function $f(\mathbf{s})$. A Gaussian Process is an example of a stochastic process, and has been discussed and applied within statistical literature for over four decades now, (Abrahamsen, 1997; Adler, 1981; MacKay, 1998; Mehr and McFadden, 1965; Rasmussen and Williams, 2006; Santner, Williams and Notz, 2003; Shepp, 1971).

The mean function $\mu(\mathbf{s})$ and covariance function of a $\mathcal{GP}$ are $\text{cov}(\mathbf{s}, \mathbf{s}^{'})$, and are for the practitioner to choose. Often, the mean function is set to zero for simplicity's sake - after all, setting the mean to zero does not limit the mean function of the posterior process to zero. As Rasmussen and Williams (2006) clarifies, while it might be naively appealing to choose a deterministic mean function, the specification of such a fixed mean function might be difficult. Instead, it is easier to define the mean function in terms of a specific (small) set of basis functions, the coefficients of which are treated as process parameters that are then learnt. This construct is due to Blight and Ott (1975) who performed a $\mathcal{GP}$-based polynomial regression analysis. In our work, model flexibility is improved upon over that achieved with fixed mean in that we choose to work with a mean function of the form $\boldsymbol{h}(\mathbf{s})^T\boldsymbol{\beta}$ for the case of the vector-valued $\eta(\mathbf{s})$, where, for any $\mathbf{s}$, we limit the set of basis functions to the linear function, so that $\boldsymbol{h}(\mathbf{s}) = (1, \mathbf{s})^T$, and the coefficients are then treated as the hyperparameter $\boldsymbol{\beta} \in \mathbb{R}^2$.

The covariance between the functions generated by a $\mathcal{GP}$ is $\text{cov}[f(\mathbf{s}, \mathbf{s}^{'})] = k(\mathbf{s}, \mathbf{s}^{'})$. As per the popular square-exponential choice (Rasmussen and Williams, 2006; Scholkopf and Smola, 2002) for the covariance function, $k(\mathbf{s}, \mathbf{s}^{'})$ is chosen to be $\sigma^2 a(\mathbf{s}, \mathbf{s}^{'})$ where $\sigma^2 > 0$ is the maximal process variance and the correlation function $a : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}$ is defined as $a(\mathbf{s}, \mathbf{s}^{'}) = \exp\{-b(\mathbf{s} - \mathbf{s}^{'})^2\}$, where $b > 0$. As we can see from the form of the correlation function, b is the square inverse length scale of this correlation function, implying that the strength of influence of $f(\mathbf{s})$ at point $\mathbf{s}^{'}$ is determined by how the distance between $\mathbf{s}$ and $\mathbf{s}^{'}$ compares to $1/\sqrt{b}$. In other words, b is the smoothness parameter that characterises this example univariate $\mathcal{GP}$. The square-exponential covariance function corresponds to a Bayesian linear regression model with infinite basis functions and is infinitely differentiable.

As an illustration, the case of vector-valued $\eta(\cdot)$ is considered. In the forward problem, the solution involves the computation of the posterior predictive distribution $[\eta(\mathbf{s}^{(new)})|b, \mathcal{D}_s]$[1]. To achieve this, we define the $n \times n$-dimensional correlation matrix $\mathbf{A}_D$ given the training data $\mathcal{D}_s$, which we view - in the absence of measurement errors - as the n-dimensional vector $(\eta(\mathbf{s}^{(1)}, \ldots, \eta(\mathbf{s}^{(n)}))^T$ corresponding to n-values of $\mathbf{S}$ at which measurements are available: $\{\mathbf{s}^{(i)}\}_{i=1}^n$. Along with this correlation matrix, we define vectors σ_D and $\boldsymbol{H}_D$, as follows.

$$
\begin{aligned}
(\mathbf{A}_D^{(n\times n)})_{ij} &:= a(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}) \\
\sigma_D(\cdot) &:= (a(\cdot, \mathbf{s}^{(1)}), \ldots, a(\cdot, \mathbf{s}^{(n)}))^T \\
\boldsymbol{H}_D^T &:= (\boldsymbol{h}(\mathbf{s}^{(1)}), \ldots, \boldsymbol{h}(\mathbf{s}^{(n)})).
\end{aligned} \tag{1}
$$

Now, $(\eta(\mathbf{s}^{(new)}), \mathcal{D}_s)$ is a realisation from a $\mathcal{GP}$. This implies that it is jointly

multivariate normal:

(2)
$$\begin{pmatrix} \eta(\mathbf{s}^{(new)}), \\ \mathcal{D}_s \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{h}(\mathbf{s}^{(new)})^T\boldsymbol{\beta} \\ \boldsymbol{H}_D\boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \sigma_D(\mathbf{s}^{(new)})^T \\ \sigma_D(\mathbf{s}^{(new)}) & \boldsymbol{A}_D \end{bmatrix} \right).$$

Then matrix algebra implies that

(3)
$$\begin{aligned} [\eta(\mathbf{s}^{(new)}) \mid b, \mathcal{D}_s] \quad \sim \quad & \mathcal{N}(\boldsymbol{h}(\mathbf{s}^{(new)})^T\boldsymbol{\beta} + (\mathcal{D}_s - \boldsymbol{H}_D\boldsymbol{\beta})^T\boldsymbol{A}_D^{-1}\sigma_D(\mathbf{s}^{(new)}) \ , \\ & \sigma^2[1 - \sigma_D(\mathbf{s}^{(new)})^T\boldsymbol{A}_D^{-1}\sigma_D(\mathbf{s}^{(new)})]) \end{aligned}$$

This posterior can be then marginalised with respect to posteriors of $\boldsymbol{\beta}$ and σ^2 to arrive at a Student's t distribution. The smoothness parameter b can be estimated by maximum likelihood methods, i.e. the posterior predictive distribution is actually derived conditional on the maximum likelihood estimate of b.

S-3. The forward problem - posterior predictive distribution of output given the data. Letting $\boldsymbol{\sigma}_D(\cdot) = [a(\cdot, \boldsymbol{s}_1), \ldots, a(\cdot, \boldsymbol{s}_n)]^T$ it follows that $[\boldsymbol{\eta}(\cdot) \mid \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s]$ is a jk-variate normal distribution with mean function

(4)
$$\boldsymbol{\mu}_1(\cdot) = \boldsymbol{B}^T\boldsymbol{h}(\cdot) + (\mathcal{D}_s - \boldsymbol{H}_D\boldsymbol{B})^T\boldsymbol{A}_D^{-1}\boldsymbol{s}_D(\cdot)$$

and the covariance function, for any $\boldsymbol{s}_1, \boldsymbol{s}_2$, is given by $a_1(\boldsymbol{s}_1, \boldsymbol{s}_2)\boldsymbol{\Omega}$, where

(5)
$$a_1(\boldsymbol{s}_1, \boldsymbol{s}_2) = a(\boldsymbol{s}_1, \boldsymbol{s}_2) - \boldsymbol{\sigma}_D(\boldsymbol{s}_1)^T\boldsymbol{A}_D^{-1}\boldsymbol{\sigma}_D(\boldsymbol{s}_2)$$

As discussed in Section 5.2 of CBB, the prior $\pi(\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}) \propto |\ \boldsymbol{\Sigma}\ |^{-(k+1)/2}|\ \boldsymbol{C}\ |^{-(j+1)/2}$, implies $[\boldsymbol{B} \mid \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s] \sim \mathcal{MN}_{m,jk}(\hat{\boldsymbol{B}}_{GLS}, (\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1}, \boldsymbol{\Omega})$, where we had defined $\hat{\boldsymbol{B}}_{GLS} := (\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1}(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\mathcal{D}_s)$. In the same section, it had also been noted that $[\boldsymbol{C}|\boldsymbol{\Sigma}, \boldsymbol{Q}, \mathcal{D}_s]$ is inverse Wishart with parameters $(n-m)k\hat{\boldsymbol{C}}_{GLS}$ and $(n-m)k$ (degrees of freedom).

Marginalising the conditional $[\boldsymbol{\eta}(\cdot) \mid \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s]$ with respect to the posteriors of $\boldsymbol{B}$ and $\boldsymbol{C}$, it can be shown that $[\boldsymbol{\eta}(\cdot) \mid \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s]$ is a jk-variate normal distribution given by

(6)
$$[\boldsymbol{\eta}(\cdot) \mid \boldsymbol{\Sigma}, \boldsymbol{C}, \boldsymbol{Q}, \mathcal{D}_s] \sim \mathcal{N}_{jk}(\boldsymbol{\mu}_2(\cdot), a_2(\cdot, \cdot)\boldsymbol{\Omega})$$

where

(7)
$$\begin{aligned} \boldsymbol{\mu}_2(\cdot) &= \hat{\boldsymbol{B}}_{GLS}^T\boldsymbol{h}(\cdot) + (\mathcal{D}_s - \boldsymbol{H}_D\hat{\boldsymbol{B}}_{GLS})^T\boldsymbol{A}_D^{-1}\boldsymbol{\sigma}_D(\cdot) \\ a_2(\boldsymbol{s}_1, \boldsymbol{s}_2) &= a_1(\boldsymbol{s}_1, \boldsymbol{s}_2) + [\boldsymbol{h}(\boldsymbol{s}_1) - \boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{s}_D(\boldsymbol{s}_1)]^T(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1} \end{aligned}$$
(8)
$$[\boldsymbol{h}(\boldsymbol{s}_2) - \boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{s}_D(\boldsymbol{s}_2)]$$

[1]This posterior can be easily arrived at, using known ideas in matrix algebra that tell us the conditional probability of the vector $\mathbf{P}$ given the vector $\mathbf{Q}$, where $\mathbf{P}$ and $\mathbf{Q}$ are consecutive sub-parts of the vector $\mathbf{R} = (\mathbf{P} \vdots \mathbf{Q})^T$. If $\mathbf{R} \sim \mathcal{N}(\mathbf{m_0}, \mathbf{A}_D^{'})$, then $\mathbf{Q}|\mathbf{P}$ can be expressed in terms of the matrix of regression coefficients and the Schur complement (Boyd and Vandenberghe, 2004), of one of the subcomponent covariance matrices that $\mathbf{A}_D^{'}$ can be split into, when written as a 2×2 matrix.

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 $\boldsymbol{M} = \boldsymbol{A}_D^{-1} - \boldsymbol{A}_D^{-1}\boldsymbol{H}_D(\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}\boldsymbol{H}_D)^{-1}\boldsymbol{H}_D^T\boldsymbol{A}_D^{-1}$).

Also, let $\mathcal{D}_s^T\boldsymbol{M}\mathcal{D}_s = [\boldsymbol{M}_{it}^*; i,t = 1,\ldots,k]$, where $\boldsymbol{M}_{it}^*$ is a matrix with j rows and j columns. Given $\boldsymbol{\Sigma}$, let $(n-m)k\hat{\boldsymbol{C}}_{GLS} = \sum_{i=1}^k\sum_{t=1}^k \psi_{it}^{-1}\boldsymbol{M}_{ti}^*$, where ψ_{it}^{-1} stands for the (i,t)-th element of $\boldsymbol{\Sigma}^{-1}$. Then it can be shown that the posterior distribution of $\boldsymbol{C}$, given $\boldsymbol{\Sigma}$, $\boldsymbol{Q}$ and $\mathcal{D}_s$ is inverse Wishart with parameters $(n-m)k\hat{\boldsymbol{C}}_{GLS}$ and $(n-m)k$ (degrees of freedom). Integrating (6) with respect to the posterior of $\boldsymbol{C}$, and writing $(\boldsymbol{\eta}-\boldsymbol{\mu}_2)(\boldsymbol{\eta}-\boldsymbol{\mu}_2)^T = [\mathbf{N}_{it}^*; i,t = 1,\ldots,k]$, we obtain the following:

$$[\boldsymbol{\eta}(\cdot) \mid \boldsymbol{Q}, \mathcal{D}_s] \propto |(n-m)k\hat{C}_{GLS}^*|^{-\frac{(n+1-m)k}{2}}, \tag{9}$$

where $(n-m)kC_{GLS}^* = \sum_{i=1}^k\sum_{t=1}^k \psi_{it}^{-1}\left(\boldsymbol{M}_{ti}^* + \frac{\mathbf{N}_{ti}^*}{a_2(\cdot,\cdot)}\right)$.

S-3.1. Now marginalisation of the posterior $[\mathbf{s}^{(test)}, \boldsymbol{\Sigma}, \mathbf{Q}, \mathbf{B}, \mathbf{C}|\mathbf{v}^{(test)}, \mathcal{D}_s]$ are sensitive to the posterior $[\mathbf{C}|\boldsymbol{\Sigma}, \mathbf{Q}, \mathcal{D}_s]$. Thus, we need to ensure that $(n-m)k\hat{\boldsymbol{C}}_{GLS}$ is positive definite in order for the posterior of $\mathbf{C}$ to be well-defined. We now present the proof that $(n-m)k\hat{\boldsymbol{C}}_{GLS}$ is positive definite.

PROOF. Since $\mathcal{D}_s^T\boldsymbol{M}\mathcal{D}_s = [\boldsymbol{M}_{it}^*; i,t = 1,\ldots,k]$ is positive definite, $\boldsymbol{B}^T\boldsymbol{M}\mathcal{D}_s = \boldsymbol{P}\boldsymbol{P}'$, where $\boldsymbol{P}$ is a lower triangular matrix with strictly positive diagonal entries. Writing $\boldsymbol{P} = (\boldsymbol{P}_1^T\vdots\boldsymbol{P}_2^T\vdots\cdots\boldsymbol{P}_k^T)^T$, where each $\boldsymbol{P}_t$ is of order $j\times j$, it then follows that $\boldsymbol{M}_{it}^* = \boldsymbol{P}_i'\boldsymbol{P}_t$. Hence, for any non-null j-component vector $\boldsymbol{s}$,

$$\sum_{i=1}^k\sum_{t=1}^k \sigma_{it}^{-1}\boldsymbol{s}^T\boldsymbol{M}_{ti}^*\boldsymbol{s} = \sum_{i=1}^k\sum_{t=1}^k \sigma_{it}^{-1}(\boldsymbol{P}_i\boldsymbol{s})^T(\boldsymbol{P}_t\boldsymbol{s}). \tag{10}$$

Here σ_{it} is the it-th element of the matrix $\boldsymbol{\sigma}_D(\cdot) := [a(\cdot,\boldsymbol{s}_1),\ldots,a(\cdot,\boldsymbol{s}_n)]^T$ The right hand side of the last equation is $\boldsymbol{y}^T(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_j)\boldsymbol{y}$, where $\boldsymbol{y} := (\boldsymbol{y}_1^T,\ldots,\boldsymbol{y}_k^T)^T$, with $\boldsymbol{y}_t = \boldsymbol{P}_t\boldsymbol{s}$, and $\boldsymbol{I}_j$ is the j-th order identity matrix. Since Kronecker product of positive definite matrices yields positive definite matrices, it follows that the last expression is positive, implying that $(n-m)k\hat{C}_{GLS}$ is positive definite. □

S-4. Importance Re-sampling MCMC (IRMCMC) for Cross-validation in Inverse Problems. We assess the validity of our model and methodology using leave-one-out cross-validation. In other words, we successively leave out data point i and the corresponding $j\times k$-dimensional velocity matrix $\mathcal{V}^{(i)}$ from the training data set, and using the remaining training data set, along with the test data, predict the i-th model parameter $\boldsymbol{S}_i$, $i = 1,\ldots,n$. Here $\mathcal{V}^{(i)}$ is the i-th row of the training data $\mathcal{D}_s^{(n\times jk)}$. Clearly, in our case, this requires a TMCMC run for each data point, implying that n many TMCMC runs in all are required. This is computationally burdensome, in spite of the attractive features of TMCMC. Bhattacharya and Haslett (2007) show 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) suggests a fast methodology for cross-validation exercise 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.

Before embarking on a discussion of the procedure for model validation, we define the data set formed after deleting $\mathcal{V}^{(i)}$ from $\mathcal{D}_s$ as $\mathcal{D}_s^{(-i)}$, $i = 1, \ldots, n$.

1. Choose an initial i^* where $\pi(\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i^*)}, \mathcal{D}_{test})$ 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:

$$d(i) = \sum_{t=1}^{n} \left\{ \sum_{u=1}^{d} \frac{(\boldsymbol{S}_t^{(u)} - \boldsymbol{S}_i^{(u)})^2}{\nu_{S_u}^2} + \sum_{\ell=1}^{j} \sum_{u=1}^{k} \frac{(\mathcal{V}_{\ell,u}^{(t)} - \mathcal{V}_{\ell,u}^{(i)})^2}{\nu_{V_u}^2} \right\}, \tag{11}$$

   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 $\boldsymbol{V}$, respectively.
2. From this density, sample, using TMCMC, as described in Section 6.2 of CBB, $(\boldsymbol{S}^{(\ell)}, \boldsymbol{Q}^{(\ell)}, \boldsymbol{\Sigma}^{(\ell)}); \ell = 1, \ldots, N$ for large N.
3. For $i \in \{1, \ldots, i^* - 1, i^* + 1, \ldots, n\}$ do,
   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

$$w_{i^*,i}(\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma}) = \frac{L(\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i)}, \mathcal{D}_{test})}{L(\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i^*)}, \mathcal{D}_{test})}, \tag{12}$$

      where $L(\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i)}, \mathcal{D}_{test}) = |\boldsymbol{A}_{D_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{H}'_{D_{aug}} \boldsymbol{A}_{D_{aug}}^{-1} \boldsymbol{H}_{D_{aug}}|^{-\frac{jk}{2}} \times |\boldsymbol{\Sigma}|^{-\frac{j(n-m)+k+1}{2}} |(n-m)k\hat{\boldsymbol{C}}_{GLS,aug}|^{-\frac{(n-m)k}{2}}$, which is the posterior $[\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i)}, \mathcal{D}_{test}]$ without the normalising constant.
   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 J_2 times from $[\boldsymbol{S} \mid \boldsymbol{Q}, \boldsymbol{\Sigma}, \mathcal{D}_s^{(-i)}, \mathcal{D}_{test}]$ using TMCMC, where

$$\begin{aligned} [\boldsymbol{S} \mid \boldsymbol{Q}, \boldsymbol{\Sigma}, \mathcal{D}_s^{(-i^*)}, \mathcal{D}_{test}] &\propto [\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{D}_s^{(-i^*)}, \mathcal{D}_{test}] \\ &\propto |\boldsymbol{A}_{D_{aug}}|^{-\frac{jk}{2}} |\boldsymbol{H}'_{D_{aug}} \boldsymbol{A}_{D_{aug}}^{-1} \boldsymbol{H}_{D_{aug}}|^{-\frac{jk}{2}} \\ &\times |\boldsymbol{\Sigma}|^{-\frac{j(n-m)+k+1}{2}} |(n-m)k\hat{\boldsymbol{C}}_{GLS,aug}|^{-\frac{(n-m)k}{2}}. \end{aligned}$$

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

S-4.1. *Simulation study.* We contrive a situation where there are $j = 3$ stars, each having $k = 2$ velocity components, and the true velocity function $\mathcal{V} = \xi(\boldsymbol{s})$, with the model parameter $\boldsymbol{S}$ being of dimension $d = 2$, $\boldsymbol{s} = (s^{(1)}, s^{(2)})$ (the two coordinates of the solar position), is of the following form:

$$\xi_{11}(\boldsymbol{s}) = \alpha s^{(1)} + \beta \frac{s^{(2)}}{1 + (s^{(1)})^2} + \gamma \cos(1.2 s^{(2)}) \tag{13}$$

$$\xi_{12}(\boldsymbol{s}) = \alpha s^{(1)} + \beta \frac{s^{(2)}}{1 + (s^{(1)})^2} \tag{14}$$

$$\xi_{21}(\boldsymbol{s}) = \alpha + \beta s^{(1)} \tag{15}$$

$$\xi_{22}(\boldsymbol{s}) = \gamma \cos(1.2 s^{(2)}) \tag{16}$$

$$\xi_{31}(\boldsymbol{s}) = \alpha s^{(1)} + \gamma \cos(1.2 s^{(2)}) \tag{17}$$

$$\xi_{32}(\boldsymbol{s}) = \gamma \cos(s^{(2)} + \sin(s^{(2)})) \tag{18}$$

Most of the above forms of the element-wise velocity functions 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. (2011).

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 element-wise functional forms (13)—(18). Here we set $\alpha = 0.05$, $\beta = 0.1$ and $\gamma = 0.2$. We thus obtained 100 data points $(\boldsymbol{s}_i, \mathcal{V}_i); i = 1, \ldots, 100$.

To demonstrate that our matrix-variate Gaussian process model fits this data very well, we implemented IRMCMC-based cross-validation detailed in Section S-4. We left 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 Section S-4 we obtained $i^* = 43$; hence, we used $[\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{\Sigma} \mid \mathcal{V}_{43}, \mathcal{D}_{test}]$ as the importance sampling density.

S-4.2. *Details of IRMCMC implementation to simulated data.* We implemented TMCMC following the details provided in Section 6.2 of CBB in order to simulate from the posterior corresponding to $i^* = 43$. Specifically, for updating $\boldsymbol{S}$, we chose the scale factors (c_1, c_2) to be $(0.1, 50)$, where $\epsilon \sim N(0, 1) I_{\{\epsilon > 0\}}$. For the smoothness parameters we simulated ϵ from a zero mean normal distribution with variance 0.005, restricted to $\mathbb{R}_+$, and selected $(0.1, 1)$ as the scale factors. For updating $\boldsymbol{\Sigma}$, we simulated from the positively restricted zero mean normal distribution with variance 0.005 and set the scale factor to be 1 for the non-zero elements of the lower triangular matrix $\boldsymbol{C}$ associated with the Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}'$. These choices are made after assessing TMCMC convergence in several pilot runs.

We discarded the first 100,000 TMCMC runs corresponding to i^* as burn-in and stored the next 100,000 runs for IR purpose. Informal convergence diagnostics indicated reasonable convergence; see, for example, the trace plots of s_1 and

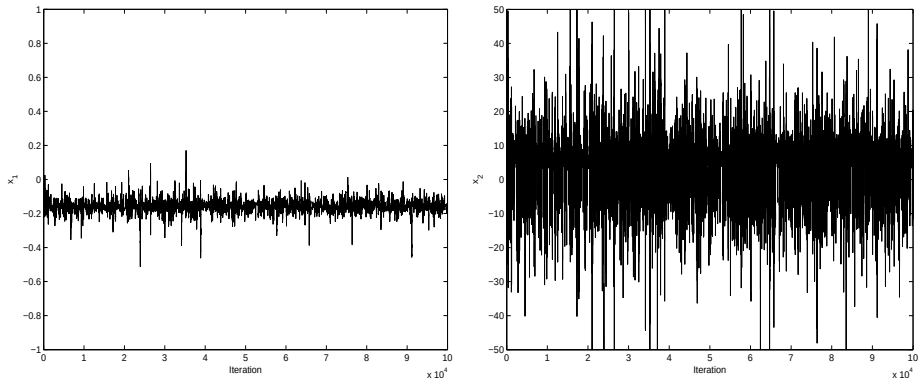

FIG S-1. **Simulation study:** *Trace plots of s_1 and s_2 corresponding to $i^* = 43$.*

s_2 in Figure S-1. 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; corresponding to the first IR-realisation, we used a burn-in of 100,000 iterations of $\boldsymbol{s}$, starting from an initial value generated uniformly from $[-1, 1] \times [-1, 1]$. Thereafter, for the remaining 99 IR-realisations, we used the last realisation of $\boldsymbol{s}$ corresponding to the previous IR-realisation of $(\boldsymbol{Q}, \boldsymbol{\Sigma})$ as the initial value for the first realisation of $\boldsymbol{s}$, without discarding any iteration as burn-in. That this is a valid and efficient strategy, has been established by Bhattacharya and Haslett (2007). Thus, we obtain 100,000 IRMCMC realisations corresponding to each location omitted from the data site.

This entire exercise took around 49 hours on a laptop—posterior simulation corresponding to i^* took around one hour, while the remaining exercise, beginning from computation of the importance weights till IRMCMC for each of the 100 data points took around 48 more hours. 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 drastically reducing the computation time.

S-4.3. *Results of cross-validation on simulated data.* In all 100% 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 S-2 and S-3. the results indicate a good fit to the data - that this is the case despite the true matrix of velocity functions consisting of highly non-linear functions as elements, and treated as unknown, is encouraging as far as the application of our matrix-variate Gaussian process-based approach to the observed stellar velocity data is concerned.

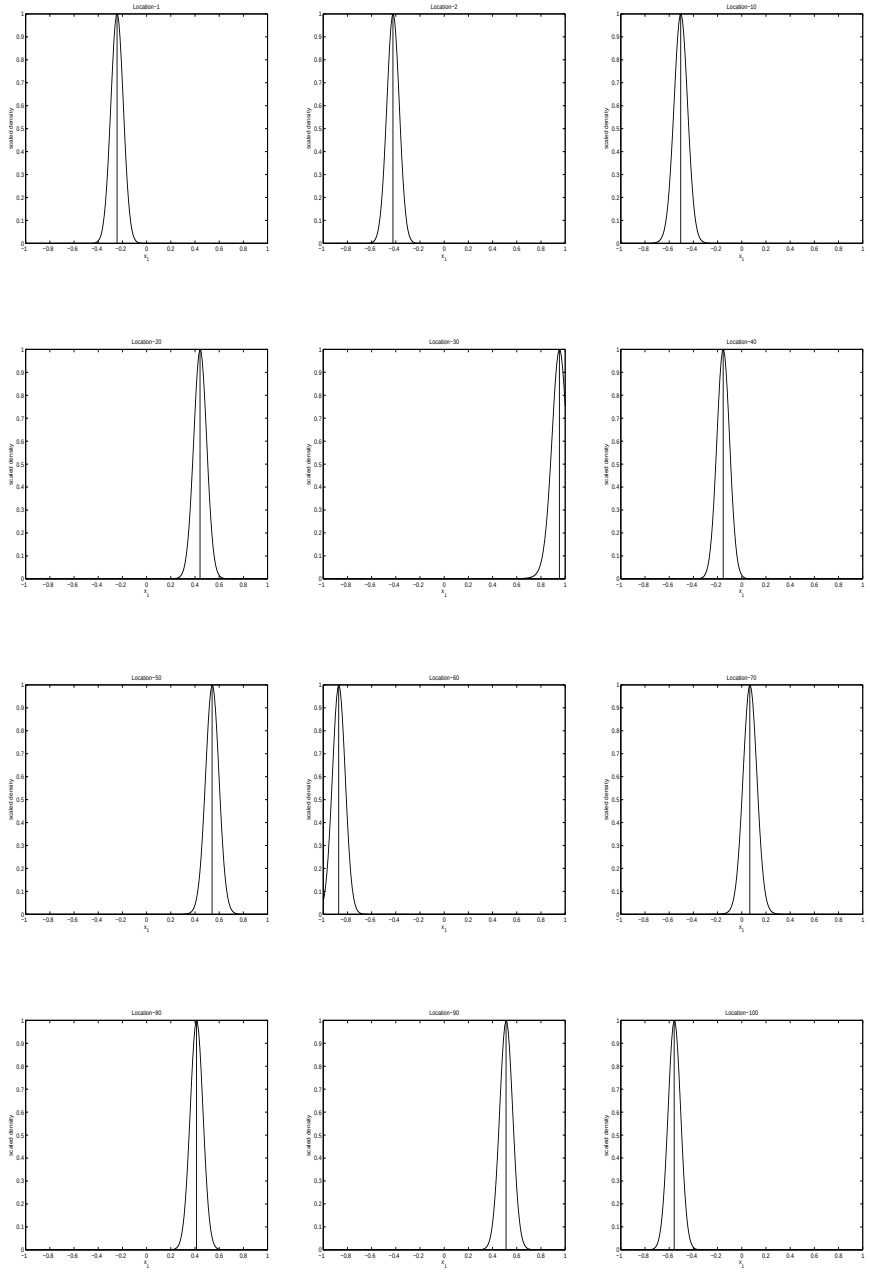

FIG S-2. **Simulation study:** *Leave-one-out cross-validation posteriors of s_1; the vertical line indicates the true value.*

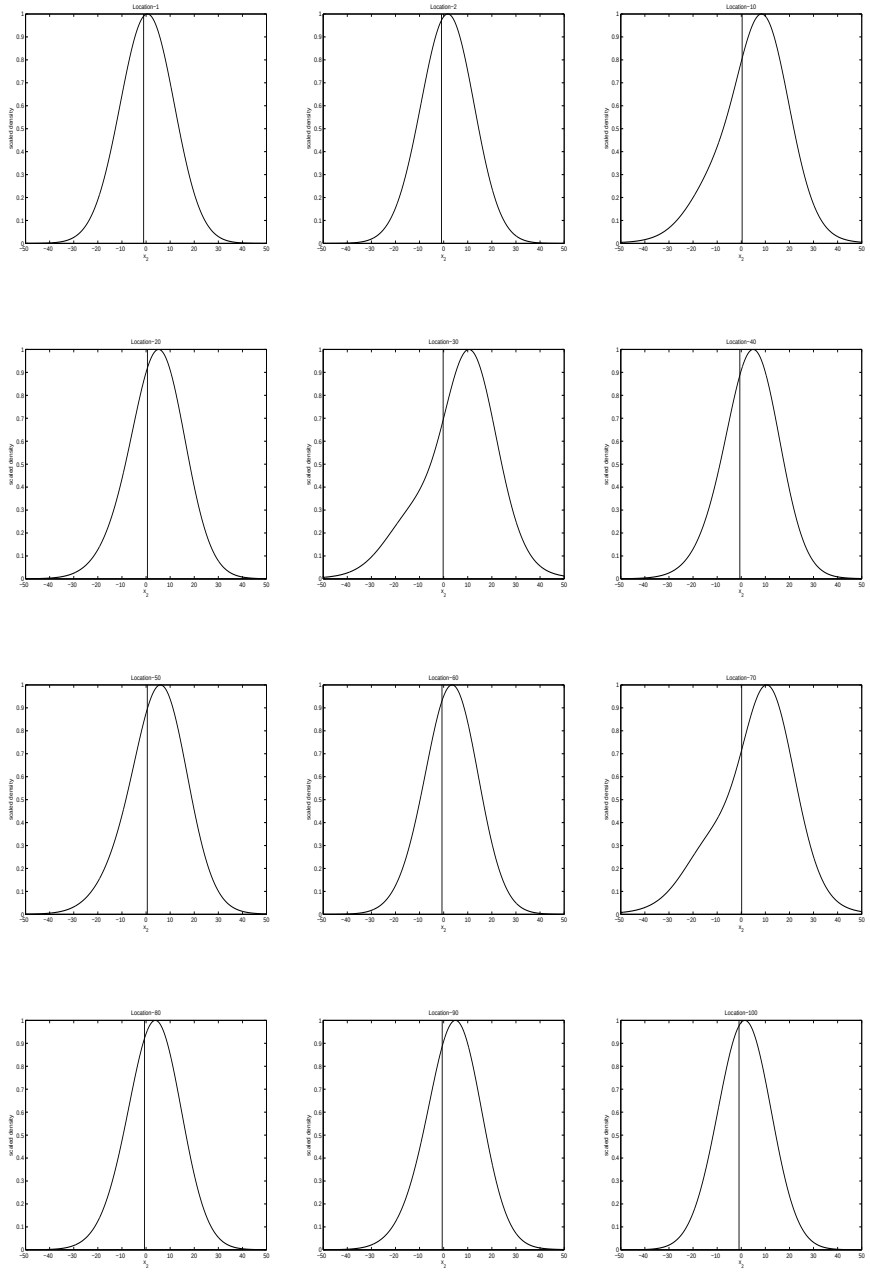

FIG S-3. **Simulation study:** *Leave-one-out cross-validation posteriors of s_2; the vertical line indicates the true value.*

References.

ABRAHAMSEN, P. (1997). A Review of Gaussian Random Fields and Correlation Functions Technical Report. Technical Report 917, Norwegian Computing Center, Oslo, Norway, http://publications.nr.no/917 Rapport.pdf.

ADLER, R. J. (1981). *The Geometry of Random Fields*. Wiley, Chichester.

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.

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

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

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., BISWAS, M. and BHATTACHARYA, S. (2012). Bayesian Learning of Milky Way Parameters Using New Matrix-Variate Gaussian Process-based Method Technical Report. Submitted.

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. (2011). Bayesian Inference in Nonparametric Dynamic State-Space Models. Submitted, available at http://arxiv.org/abs/1108.3262.

MACKAY, D. J. C. (1998). Introduction to Gaussian Processes. In *Neural Networks and Machine Learning* (C. M. BISHOP, ed.). Springer-Verlag.

MEHR, C. B. and MCFADDEN, J. A. (1965). Certain properties of Gaussian processes and their first passage times. *J. Roy. Statist. Soc. Ser. B* **27** 505–522.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, MIT.

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The design and analysis of computer experiments*. *Springer Series in Statistics*. Springer-Verlag, New York, Inc.

SCHOLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*. MIT Press, MIT.

SHEPP, L. A. (1971). First Passage Time for a Particular Gaussian Process. *The Annals of Mathematical Statistics* **42** 946–951.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY CV4 7AL, U.K.
E-MAIL: d.chakrabarty@warwick.ac.uk

INDIAN STATISTICAL INSTITUTE
203, B. T. ROAD,
KOLKATA 700108
INDIA
E-MAIL: munmun.biswas08@gmail.com

BAYESIAN & INTERDISCIPLINARY RESEARCH INSTITUTE
INDIAN STATISTICAL INSTITUTE
203, B. T. ROAD,
KOLKATA 700108
INDIA
E-MAIL: sourabh@isical.ac.in