The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2672–2692
https://doi.org/10.1214/22-AOAS1608

MAPPING INTERSTELLAR DUST WITH GAUSSIAN PROCESSES

BY ANDREW C. MILLER[1,a], LAUREN ANDERSON[4,e], BORIS LEISTEDT[4,f], JOHN P. CUNNINGHAM[2,c], DAVID W. HOGG[3,d] AND DAVID M. BLEI[1,b]

[1]*Data Science Institute, Columbia University,* [a]*am5171@columbia.edu,* [b]*david.blei@columbia.edu*
[2]*Department of Statistics, Columbia University,* [c]*jpc2181@columbia.edu*
[3]*Department of Physics, New York University,* [d]*david.hogg@nyu.edu*
[4]*Center for Computational Astrophysics, Flatiron Institute,* [e]*anders.astro@gmail.com,* [f]*boris.leistedt@gmail.com*

Interstellar dust corrupts nearly every stellar observation and accounting for it is crucial to measuring physical properties of stars. We model the dust distribution as a spatially varying latent field with a Gaussian process (GP) and develop a likelihood model and inference method that scales to millions of astronomical observations. Modeling interstellar dust is complicated by two factors. The first is *integrated observations*. The data come from a vantage point on Earth, and each observation is an integral of the unobserved function along our line of sight, resulting in a complex likelihood and a more difficult inference problem than in classical GP inference. The second complication is *scale*; stellar catalogs have millions of observations. To address these challenges, we develop ZIGGY, a scalable approach to GP inference with integrated observations based on stochastic variational inference. We study ZIGGY on synthetic data and the Ananke dataset, a high-fidelity mechanistic model of the Milky Way with millions of stars. ZIGGY reliably infers the spatial dust map with well-calibrated posterior uncertainties.

1. Introduction. The Milky Way galaxy is primarily comprised of dark matter, stars, and gas. Within the gas, in its densest and coldest regions, dust particles form. The stars of the Milky Way are embedded in this field of dust.

Back on Earth, astronomers try to map the stars, measuring the location, apparent brightness, and color of each. But the dust between Earth and a star obscures the light, corrupting the astronomers' observations. Relative to the star's true brightness and color, the dust *dims* the brightness and *reddens* the color. This corruption is called *extinction*, and it complicates inferences about a star's true distance and other true properties. Stellar extinction has hindered many studies of stars in the Milky Way disk, which is where most dust lies, and, therefore, the largest extinctions (Mathis (1990)).

To cope with these corruptions, astronomers need a map of interstellar dust, an estimate of the density of dust at each location in the galaxy. An accurate dust map could be used to correct our astronomical measurements and sharpen our knowledge of the galaxy's stars. Moreover, a dust map may be of independent scientific interest—for example, it may reveal macroscopic properties of the shape of the Milky Way, such as spiral arms.[1]

But constructing such a map is difficult. We are embedded in our own dust field, and so we cannot directly observe it. Rather, we can only observe a noisy *integral* of the field along the line of sight between Earth and a star—the starlight extinction. (Here, the line of sight is the straight line through space between the star and the Earth.) Thus, the inference problem

Received April 2021; revised November 2021.
Key words and phrases. Gaussian process, interstellar dust, astrostatistics, stochastic variational inference.

[1]Though hints of spiral structures have been inferred from other observations (Georgelin and Georgelin (1976), Chen et al. (2019)), whether the Milky Way galaxy is a grand design spiral or if the spiral structures are more flocculent remains an open question. An accurate dust map will shed light on the recent dynamical history of the Milky Way.

is to calculate a single, coherent spatial field of dust that can explain the observed extinction of millions of spatially distributed stars. In this paper we use a large data set of astronomical measurements to infer a three-dimensional map of interstellar dust.

The data comes from standard practice in astronomy which is to estimate the extinction of an individual star from its observed color and brightness using a physical model of star formation and evolution.[2] This procedure results in an estimate of the extinction and an approximate variance of the estimate about the true extinction value. However, this estimate is derived from a single star's color and brightness; it ignores information about the spatial structure of the Milky Way and the fact that all stars are observed through the same three-dimensional density of dust.

The dust map is a spatial distribution of dust—an unobserved spatial function—that can help refine these noisy measurements. Each measurement is modeled conditional on the function and the three-dimensional location of the star.[3] Such a model shrinks estimates toward their true values and reduces uncertainty about them. More formally, the unknown dust map is a function $\rho : \mathbb{R}^D \mapsto \mathbb{R}$ from a three-dimensional location in the universe x to the density of dust at that location. Our goal is to estimate this function from data.

We take a Bayesian nonparametric approach. We place a Gaussian process (GP) prior on the dust map and posit a likelihood function for how astronomical observations arise from it. We develop a scalable approximate posterior inference algorithm to estimate the posterior dust map from large-scale astronomical data. In addition to the scale of the data, the main challenge is the likelihood function. Typical spatial analysis involves data that are noisy evaluations of an unobserved function, and the likelihood is simple. Astronomical data, however, comes from a limited vantage point; we only observe the latent dust map as an *integrated process* along a line of site to a star, and this integral is baked into the likelihood of the observation.

More formally, let e_n be the true extinction of starlight for star n at location x_n; let a_n be the noisily measured extinction with uncertainty σ_n. Given the unknown dust map, we model the noisy measurement as a Gaussian whose mean is an integral from Earth to x_n (Rezaei Kh. et al. (2017)). With covariance function $k_{\boldsymbol{\theta}}(\cdot, \cdot)$, the full model is

$$\rho(\cdot) \sim \mathrm{GP}\big(0, k_\theta(\cdot, \cdot)\big), \tag{1}$$

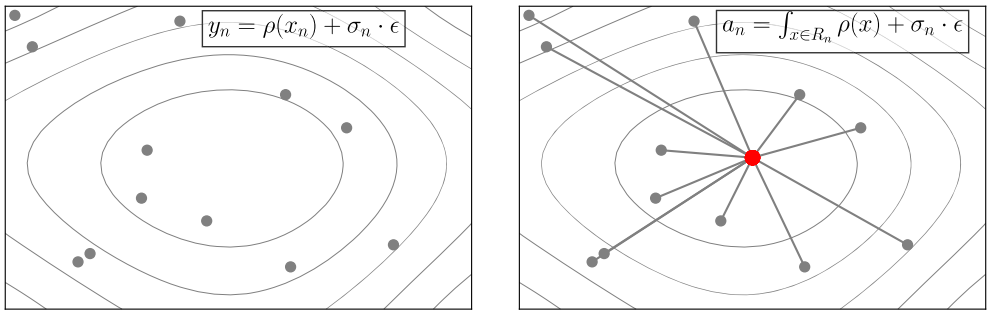
$$a_n \sim \mathcal{N}\big(e_n, \sigma_n^2\big), \quad \text{where } e_n \triangleq \int_{x \in R_n} \rho(x)\, dx. \tag{2}$$

Here, R_n is the set of points (i.e., the ray) from Earth to x_n; the extinction e_n is in an integral of the latent dust map $\rho(\cdot)$, and it is through this integral that the map enters the likelihood. Figure 1 graphically depicts the distinction between pointwise and *integrated* observations.

The data are N locations, extinctions, and measurement errors, denoted $\mathcal{D} = \{a_n, x_n, \sigma_n^2\}_{n=1}^N$. Conditional on the data, we want to calculate the posterior dust map $\pi(\rho \mid \mathcal{D})$. This posterior can estimate different properties of the latent dust, for example, the density of dust $\rho(\cdot)$ at a new location, the integral of dust $\rho(\cdot)$ over new sets, and posterior uncertainty about these values. Such inferences can aid many stellar studies.

[2]The observed color and brightness are directly derived from telescope images (e.g., photometry). Extinction measurements are backed out from a theoretical distribution of dust-free star colors and luminosities, where luminosity is the intrinsic brightness of a star. This distribution can be based on isochrones or empirically derived from a region of the sky known to have no dust (e.g., the Milky Way halo). The extinction corresponds to how far the observed color and brightness are from the set of theoretically plausible colors and luminosities. The extinction uncertainty incorporates both noise in the photometric measurement and prior uncertainty over the range of plausible colors and luminosities.

[3]The spatial locations of stars can be derived from parallax measurements. In this work we consider them fixed but in future work will additionally consider their uncertainty.



(a) Pointwise **(b)** Integrated

FIG. 1. *Reconstructing an unobserved function from pointwise* (*left*) *and integrated* (*right*) *observations. The latent function governs noisy pointwise observations,* y_n*, and integrated observations,* a_n*. The task is to reconstruct the unobserved function* $\rho(x)$*, depicted by the grey contours, everywhere in the domain. Our perspective is limited to the origin* (*large center dot*)*; our observations of this process are integrated along a compact set* (*i.e., a ray*).

But the posterior is difficult to compute, complicated by both the integrated likelihood and the large scale of the data. (Modern catalogs of astronomical data contain observations of millions of stars (Brown et al. (2018), Aguado et al. (2019)).) In theory, the scale is helpful; each star provides information about the unobserved dust map. But scaling a Gaussian process to millions of observations is a significant computational challenge.

We overcome these challenges with a scalable algorithm for Gaussian process inference with integrated observations, which we call ZIGGY.[4] In particular, ZIGGY is a stochastic variational inference algorithm, one that uses stochastic optimization, inducing points, and variational inference to approximate the posterior. It handles general covariance functions and scales to millions of data points. We study ZIGGY on both synthetic data and the Ananke data which comes from a high-fidelity mechanistic model of the Milky Way. We find that ZIGGY accurately reconstructs the dust map, and accuracy continues to improve as the number of observations grows above a million.

In our applied setting the integrated observations introduce technical challenges. ZIGGY builds upon an existing scalable Bayesian inference framework (Hoffman et al. (2013), Hensman, Fusi and Lawrence (2013)) and introduces additional approximations and computational techniques to address the challenges created by integrated observations.

The paper is organized as follows. Section 2 describes related work for estimating interstellar dust. Section 3.1 formally sets up the problem and describes exact inference in GP models with integrated observations. Section 4 develops a stochastic variational inference algorithm that scales to millions of stellar observations. Appendix F in the Supplementary Material (Miller et al. (2022)) studies ZIGGY on a synthetic two-dimensional example, comparing various settings of the algorithm that trade computation for accuracy and flexibility. Section 5 studies ZIGGY with the Ananke data set, showing that it recovers a well-calibrated three-dimensional dust map that accurately predicts extinctions and global dust structure. Section 6 concludes the paper and discusses directions for future research.

2. Related research. This work builds on a foundation of research in both astronomy and statistics.

Estimating the latent dust map. The seminal work of Schlegel, Finkbeiner and Davis (1998) estimates the two-dimensional map of dust across the full sky using dust emission (rather

[4]ZIGGY is named for Ziggy Stardust, David Bowie's alter ego.

than dust absorption, which we use here). Emission is a more direct estimate of the dust, but doesn't allow for direct inferences of 3D structure. Stars within the Milky Way, however, are embedded in the three-dimensional dust field. Estimating per-star extinctions requires characterizing the dust field as a function of distance, motivating the construction of three-dimensional dust maps.

More recent approaches use hierarchical spatial models of noisy integrated observations to reconstruct three-dimensional maps. These approaches, however, rely on *discretizing space* and modeling integrated observations as finite sums. Various discretization strategies have been proposed, each with different computational demands.

One approach models discretized lines of sight to every star (Rezaei Kh. et al. (2017)), but this approach can only accommodate a few thousand stellar observations. Another approach uses a fine mesh of cubic voxels to describe the dust distribution. This introduces a different computational trade-off. Larger voxels introduce unrealistic spatial artifacts (Green et al. (2018, 2019)), while smaller voxels significantly increase the computational burden (Leike and Enßlin (2019)). Moreover, the theoretical length scale of the dust distribution is small; discretizing three-dimensional space with fine enough resolution to capture this small length scale will have to rely on additional approximations to scale to the entire volume of the Milky Way. The approach we develop here avoids discretization, works directly with continuous space, and scales to millions of observations.

Spatial statistics and Gaussian processes. The field of spatial statistics has developed many tools to estimate unobserved functions from noisy measurements (Cressie (1992)). One ubiquitous and fundamental method is Gaussian process regression, or *kriging*, which interpolates or smooths a latent function, given noisy observation at nearby locations (Krige (1951), Rasmussen and Williams (2006) Matheron (1963, 1973), Cressie (1990)). A GP defines a probability distribution over the unobserved function that encodes prior assumptions about some properties—continuity, smoothness, and amplitude—while remaining flexible. GPs admit analytically tractable inference routines in typical settings, making them a useful prior for unobserved functions.

Scaling GPs to massive datasets is a more recent challenge. One approach is to approximate the GP with inducing point methods (Quiñonero-Candela and Rasmussen (2005), Snelson and Ghahramani (2006)), where process values at specific places in the space, the inducing points, are learned from the data, and predictions are then produced using the inducing point process values. Inference using this approximation scales better than exact GP inference. Bolstering their use, recent theoretical work has characterized the error of inducing point GP approximations (Burt, Rasmussen and van der Wilk (2019)). However, inducing point methods still require analyzing the entire dataset simultaneously; this requirement limits their applicability to datasets where the number of observations is in the millions or billions.

Another thread of GP research in spatial statistics embeds them in more complicated models and scales them (Cressie (1992), Banerjee, Carlin and Gelfand (2015)). For example, nearest-neighbor GP's are a scalable approximation for estimating the posterior of a latent spatial field (Datta et al. (2016)). Similarly, the integrated nested Laplace approximations (INLA) framework is an approximate inference methodology that computes posterior marginal uncertainty in latent Gaussian models, including Gaussian process models with computationally tractable precision matrices (Rue, Martino and Chopin (2009)). Several scalable Gaussian process approximations are reviewed in Heaton et al. (2019).

Gaussian processes have also been used within astronomy applications beyond building spatial maps of interstellar dust. Methods to scale one-dimensional GP regression to millions of observations have been developed and deployed with success (Ambikasaran et al. (2014), Foreman-Mackey et al. (2017)).

Scalable Bayesian inference. We build on methods for scaling Bayesian inference to massive datasets. Variational methods (Jordan et al. (1999), Wainwright et al. (2008), Blei, Ku-

cukelbir and McAuliffe (2017)) are a computationally efficient alternative to Monte Carlo methods for posterior approximation. Variational inference treats approximate inference as an optimization problem, fitting a parameterized family to be close to the exact posterior. Stochastic optimization (Robbins and Monro (1951)) scales variational inference to large datasets, iteratively subsampling from the data to produce cheap, noisy gradients of the objective; this strategy is called stochastic variational inference (SVI) (Hoffman et al. (2013)). Building off of SVI, stochastic variational Gaussian processes (SVGP) bring in inducing point methods to scale Gaussian processes to massive datasets (Hensman, Fusi and Lawrence (2013)). We build on the SVGP framework here, adapting it to fit the GP model of astronomical data.

3. A Gaussian process model of starlight extinctions. The data are N stars, each one a tuple of its location x_n, a noisy measurement of the extinction a_n, and the variance of the observation σ_n^2; denote the data set $\mathcal{D} \triangleq \{x_n, a_n, \sigma_n^2\}_{n=1}^N$. Given the latent dust map $\rho(x)$, the likelihood of each observation a_n is defined in equation (2), where the region of integration R_n is the ray that originates at the origin O (i.e., the Earth) and ends at the spatial location of the star, x_n,

$$R_n = \{\alpha \cdot O + (1-\alpha) \cdot x_n : \alpha \in [0,1]\}. \tag{3}$$

To complete the model, equation (1) places a Gaussian process prior on the dust map.

Given the data, the posterior distribution, $p(\rho|\mathcal{D})$, summarizes the evidence about the latent dust map. Via the relationship between $\rho(x_n)$ and e_n, defined in equation (2), we use this posterior over ρ to form estimates of the extinction for observed stars e_n, the extinction at new locations e_*, and the dust map itself at new locations $\rho(x_*)$. Specifically, compute the posterior expectations

$$\hat{\rho}(x_*) \triangleq \mathbb{E}[\rho(x^*)|\mathcal{D}], \quad \hat{e}_n \triangleq \mathbb{E}[e_n|x_n, \mathcal{D}], \quad \hat{e}_* \triangleq \mathbb{E}[e_*|x_*, \mathcal{D}]. \tag{4}$$

The posterior variance, for example, $\mathbb{V}[e_*|x_*, \mathcal{D}]$, describes the uncertainty of the estimates.

The posterior dust map $\rho(\cdot)$ synthesizes information from nearby sources to de-noise or shrink an individual extinction e_n, resulting in more accurate inferences with smaller estimator variance. Moreover, it enables estimating the dust density at a new location x_* or to estimate estimate the extinction along a path to the new point.

In the following sections we discuss how to calculate the posterior and how to approximate it with large data sets of stellar observations. Before that, however, we discuss the choice of a Gaussian process prior.

Why use a Gaussian process? First, Gaussian processes can flexibly model complex unobserved latent functions. The dust map $\rho(\cdot)$ is not easily described by a parametric functional form, but GPs can adapt highly complex, nonlinear functions to describe data. Second, high-level properties of the unobserved dust map can be captured by the choice of the GP's covariance function. In particular, we assume the dust map has specific structure; it is continuous, and two nearby locations are more likely to have similar dust density values than two distant locations. GPs naturally capture such spatial coherence. Third, all stellar observations are derived from the same underlying dust map $\rho(x)$. Two noisy extinction measurements with nearby integration paths must be coherently described by a single estimate of $\rho(\cdot)$. The continuity, smoothness, and spatial coherence across observations give traction in forming an accurate statistical estimate of the unobserved dust map.

3.1. *GP inference with integrated observations.* In a GP, any finite set of function evaluations are multivariate normal distributed with covariance defined by the covariance function, $\mathrm{cov}(\rho(x_i), \rho(x_j)) = k_{\boldsymbol{\theta}}(x_i, x_j)$. When observations are corrupted with Gaussian noise—the

typical setting for GP models—the joint distribution between all observations and latent process quantities remains multivariate normal, and the posterior at a new point $\rho(x^*)$ can be expressed as a few matrix operations (Rasmussen and Williams (2006), Chapter 2). Inference in Gaussian process models with noisy *integrated observations*, defined in equation (2), closely mirrors inference with the more typical, noisy pointwise observations.

Consider the model in equations (1) and (2). We derive the posterior distribution by first characterizing the joint distribution over $\rho(x_*)$ and $\mathbf{a} = (a_1, \ldots, a_N)$ (Rasmussen and Williams (2006)). Observe that linear operations of multivariate normal random variables remain multivariate normal in distribution. As noted in Rasmussen and Williams (2006), Section 9.8, because integration is a linear operation, the value of the integral of ρ over the domain $\mathcal{X}$ will remain Gaussian. This is also true for any collection of definite integrals and the corresponding (Gaussian) noisy observations.

The joint distribution over $\mathbf{a}$ and $\rho(x_*) \triangleq \rho_*$ remains multivariate normal. Given the mean and covariance of this joint distribution, we can compute the posterior distribution over ρ_* (or any set of ρ_* points), given observations $\mathbf{a}$, $\mathbf{x}$, and $\boldsymbol{\sigma}^2$. The joint distribution over $\mathbf{a}$ and ρ_* mimics the standard Gaussian process setup

$$\begin{pmatrix} \mathbf{a} \\ \rho_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K_e} + \boldsymbol{\Sigma_a} & \mathbf{K_{e,*}} \\ \mathbf{K_{*,e}} & \mathbf{K_*} \end{pmatrix}\right), \tag{5}$$

where the covariance matrix blocks are defined

$$(\mathbf{K_e})_{ij} = \mathrm{Cov}(e_i, e_j), \qquad (\mathbf{K_{*,e}})_{*j} = \mathrm{Cov}(\rho_*, e_j), \qquad \mathbf{K_*} = \mathrm{Cov}(\rho_*, \rho_*), \tag{6}$$

and the observation noise matrix $\boldsymbol{\Sigma_a} \equiv \mathtt{diag}(\sigma_1^2, \ldots, \sigma_N^2)$. We note that, because ρ_* is scalar, $\mathbf{K_*}$ is a 1×1 matrix. Computing these covariance entries enables us to populate the joint covariance matrix and use standard multivariate normal conditioning for posterior inference. This conditioning mimics the equations for standard Gaussian process inference,

$$\rho(x_*) \sim \mathcal{N}(\mu_*, \Sigma_*), \tag{7}$$

$$\mu_* = \mathbf{K_{*,e}}(\mathbf{K_e} + \boldsymbol{\Sigma_a})^{-1}\mathbf{a}, \tag{8}$$

$$\Sigma_* = \mathbf{K_*} - \mathbf{K_{*,e}}(\mathbf{K_e} + \boldsymbol{\Sigma_a})^{-1}\mathbf{K_{e,*}}. \tag{9}$$

To compute this posterior, we need to compute the covariances in equation (6).

The chosen covariance function directly specifies $\mathrm{Cov}(\rho_*, \rho_*)$ The other terms, $\mathrm{Cov}(e_i, e_j)$ and $\mathrm{Cov}(\rho_*, e_j)$ are slightly more complex and require integrating the covariance function in one or both of its arguments.

CLAIM 3.1 (Semi-integrated covariance function). *The covariance between a process value $\rho_i \triangleq \rho(x_i)$ and an integrated value $e_j \triangleq \int_{x \in R_j} \rho(x)\, dx$ takes the form*

$$\mathrm{Cov}(\rho_i, e_j) = \int_{x \in R_j} k(x_i, x)\, dx = k^{(\mathrm{semi})}(x_i, x_j). \tag{10}$$

For consistency, we will write the integrated argument second.

The proof is in Appendix B in the Supplementary Material (Miller et al. (2022)).

CLAIM 3.2 (Doubly-integrated covariance function). *The covariance between two integrated process values e_i and e_j takes the form*

$$\mathrm{Cov}(e_i, e_j) = \int_{(x,x') \in R_i \times R_j} k(x, x')\, dx\, dx' = k^{(\mathrm{double})}(x_i, x_j). \tag{11}$$

The proof for the doubly-integrated kernel is similar to the semi-integrated case.

To summarize, if we can compute these two types of covariance values, then we can plug the result into the distribution in equation (5). As for standard GP inference, we can then manipulate the joint distribution using Gaussian conditioning to compute the posterior distribution over the value of the unobserved function $\rho(x_*)$ or functionals over test locations, for example, integrated values e_*. Algorithm 1 in the Supplementary Material (Miller et al. (2022)) summarizes GP inference using integrated measurements.

3.2. *Choice of covariance function.* Assumptions about the dust map $\rho(\cdot)$ can be encoded in the covariance function $k_{\boldsymbol{\theta}}(\cdot, \cdot)$. We focus on kernels that are *stationary* and *isotropic*, which can be written $k_{\boldsymbol{\theta}}(x, y) = \sigma^2 k(\frac{|x-y|}{\ell})$, where $\boldsymbol{\theta} \triangleq (\sigma^2, \ell)$. The parameter ℓ is the *length scale*, and the parameter σ^2 is the process marginal variance (Genton (2002)). The length scale controls how smooth the function is; the marginal variance influences the function's amplitude. We will compare three families of kernel functions: squared exponential, Matérn, and a kernel from Gneiting (Gneiting (2002)). The squared exponential kernel—a commonly used covariance function for Gaussian process regression—admits an analytically tractable form for the semi-integrated kernel but not the doubly-integrated kernel. (See Appendix A in the Supplementary Material (Miller et al. (2022)).) Other kernels, such as the Matérn (Matérn (1960)) or the kernel presented in Gneiting (2002) do not readily admit a semi-integrated form which complicates their use with integrated observations. Similarly, none of the mentioned kernels, including the squared exponential, admit a doubly-integrated form over two line segments. We develop a method to overcome this technical limitation in the following section. Furthermore, we use the analytic tractability of the semi-integrated kernel to validate this method on the squared exponential kernel.

Beyond the choice of covariance function family, the covariance function parameters (e.g,. the length scale and process variance) ought to be tuned. Tuning GP covariance parameters can be complex; the process values $\boldsymbol{\rho}$ can strongly depend on parameters σ^2 and ℓ, making it difficult to explore their joint space. We discuss this phenomenon within the context of our scalable approximate inference algorithm in Section 4.

4. Scaling integrated GP inference. Gaussian process models scale poorly to large datasets (Quiñonero-Candela and Rasmussen (2005), Titsias (2009)). Computing the posterior distribution of the latent function, the posterior predictive distribution of new observations, or the marginal likelihood of observed data (e.g., for model comparison) all require computing the inverse or the determinant of a $N \times N$ matrix — a $O(N^3)$ operation. As N grows larger than a few thousand observations, this computation becomes intractable. To address this bottleneck, a common strategy is to approximate Gaussian process inference using $M \ll N$ *inducing points* or spatial locations in the input space that are used to approximate the full Gaussian process. Inducing point approximations can be interpreted in multiple ways. One interpretation is that the inducing points are used to construct a rank M approximation to the $N \times N$ matrix that can be efficiently inverted (e.g., $O(N \cdot M^3)$) (Quiñonero-Candela and Rasmussen (2005)). Another interpretation (which we adopt below) is that the inducing points are used to define a family of variational distributions for approximating the posterior over ρ, and optimizing the variational objective avoids direct manipulation of the $N \times N$ matrix (Titsias (2009), Hensman, Fusi and Lawrence (2013)).

Integrated observations create an additional computational issue; the semi-integrated or doubly-integrated covariance value between any two values may be unavailable in closed form (except in some special cases) and may require high-precision numerical approximation. With the integrated observations considered here, we have an additional computational issue. Calculating the semi-integrated or doubly-integrated covariance value between any

two values may be unavailable in closed form (except in some special cases) and may require high-precision numerical approximation. The exact GP inference algorithm in Algorithm 1 (in the Supplementary Material) requires computing a $N \times N$ matrix of doubly-integrated covariance values which will be computationally prohibitive when N is a modest size. Furthermore, when tuning covariance function parameters or comparing covariance functions, the $N \times N$ matrix of numerically integrated covariance function values will need to be recomputed many times.

Thus, to use the model to analyze millions of integrated stellar observations, we derive a scalable variational inference algorithm to perform approximate posterior inference. Variational inference approximates a posterior distribution by optimizing a parameterized family—the variational distribution—to be close to the exact posterior (Blei, Kucukelbir and McAuliffe (2017), Jordan et al. (1999), Wainwright et al. (2008)). For the model here, our approach builds on the stochastic variational Gaussian process (SVGP) framework (Hensman, Fusi and Lawrence (2013)). It uses stochastic optimization to approximate the posterior (Hoffman et al. (2013)), operating on small minibatches of observations.

Further, to accommodate the integrated observations, we construct a Monte Carlo estimate of the semi-integrated kernel that generalizes to covariance functions that do not admit a closed form semi-integrated version. Additionally, we use the rotational invariance of stationary and isotropic covariance functions to construct a fast approximation to the doubly-integrated covariance kernel.

This section describes the SVGP framework and our approach to adapt the framework to incorporate integrated observations. Section 4.1 describes relevant details of the SVGP framework, including inducing points, the variational family, and forming efficient mini-batch estimates of the variational objective. Section 4.2 details our approach to do inference with integrated observations for a general class of semi-integrated kernels.

4.1. *Stochastic variational Gaussian processes.* SVGP casts inference in GP models as an optimization problem. Crucially, the SVGP optimization objective is constructed such that it can be written as a sum of N terms, each depending on only one data point. Given this construction, the objective can be optimized using stochastic gradients computed with minibatches of data. This makes inference more computationally and memory efficient.

Here, we discuss the SVGP objective with the standard observations—noisy versions of $\rho(x)$. We then adapt this approximate inference framework to integrated observations in Section 4.2. Our derivation of the variational objective is slightly different from Hensman, Fusi and Lawrence (2013); we first define a structured approximating family and then directly derive the evidence lower bound objective.

For this section's presentation of inducing points, variational inference, and stochastic variational Gaussian processes, consider the typical GP model for N observations with $\rho \sim \mathrm{GP}(0, k_\theta(\cdot,\cdot))$ and standard (i.e., nonintegrated) observations $y_n|\mathbf{x}_n \sim \mathcal{N}(\rho(\mathbf{x}_n), \sigma_n^2)$. We write the observation vector $\mathbf{y} = (y_1, \ldots, y_N)$ and the corresponding latent process vector $\boldsymbol{\rho} = (\rho(\mathbf{x}_1), \ldots, \rho(\mathbf{x}_N))$.

Inducing points. Inducing points are a common tool used to scale Gaussian process inference (Quiñonero-Candela and Rasmussen (2005)). An inducing point is simply a location in the input space, $\bar{x}$, which has a corresponding inducing point value $\rho(\bar{x})$. Inducing points are typically distributed about the input space. In our experiments we place them on a fixed three-dimensional grid throughout space.

We augment the model above with M inducing points and their corresponding values

$$\bar{\mathbf{x}} \triangleq \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_M \quad \text{inducing points,} \tag{12}$$

$$\mathbf{u} \triangleq \rho(\bar{\mathbf{x}}_1), \ldots, \rho(\bar{\mathbf{x}}_M) \quad \text{inducing point values,} \tag{13}$$

where $\mathbf{u}$ is the M-length vector of values of ρ evaluated at each of the M inducing points. Note that the introduction of $\bar{\mathbf{x}}$ and $\mathbf{u}$ has not altered the original model.

The model above specifies two latent variables, $\boldsymbol{\rho}$ and $\mathbf{u}$, whose distributions we wish to infer, given data $\mathbf{y}$. That is, we wish to characterize the posterior distribution $p(\boldsymbol{\rho}, \mathbf{u}|\mathbf{y}, \mathbf{x}, \bar{\mathbf{x}})$ in a computationally efficient way.[5] SVGP uses inducing points and a specific variational approximation to avoid the manipulation of the $N \times N$ observation covariance matrix.

Variational inference. Variational inference (VI) is an optimization-based approach to approximate posterior inference (Jordan et al. (1999), Wainwright et al. (2008), Blei, Kucukelbir and McAuliffe (2017)). VI methods posit a *variational family* of distributions, $q \in Q$, and use optimization techniques to find the optimal approximate distribution from the set Q. Here, each element of Q is a distribution over the latent quantities—$q(\boldsymbol{\rho}, \mathbf{u})$.

The variational inference framework defines an objective to be optimized; the most common VI objective is the *evidence lower bound* (ELBO)

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{u},\boldsymbol{\rho})}\big[\ln p(\boldsymbol{\rho}, \mathbf{u}, \mathbf{y}|\mathbf{x}, \bar{\mathbf{x}}) - \ln q(\boldsymbol{\rho}, \mathbf{u})\big] \leq \ln p(\mathbf{y}|\mathbf{x}, \bar{\mathbf{x}}). \tag{14}$$

Maximizing the ELBO minimizes the KL divergence between $q(\boldsymbol{\rho}, \mathbf{u})$ and the true posterior $p(\boldsymbol{\rho}, \mathbf{u}|\mathbf{y}, \mathbf{x}, \bar{\mathbf{x}})$.

Stochastic variational Gaussian processes. The stochastic variational GP framework (SVGP) (Hensman, Fusi and Lawrence (2013)) uses a particular form for the variational family. Given the set of inducing points $\bar{\mathbf{x}}$, the SVGP variational approximation is

$$q(\boldsymbol{\rho}, \mathbf{u}) = p(\boldsymbol{\rho}|\mathbf{u})q_{\boldsymbol{\lambda}}(\mathbf{u}), \qquad q_{\boldsymbol{\lambda}}(\mathbf{u}) \triangleq \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S}), \quad \text{and} \quad \boldsymbol{\lambda} \triangleq (\mathbf{m}, \mathbf{S}), \tag{15}$$

where $p(\boldsymbol{\rho}|\mathbf{u})$ is specified by the Gaussian process prior and recall that $\mathbf{u}$ and $\boldsymbol{\rho}$ are jointly multivariate normal. The variational parameters $\boldsymbol{\lambda}$ are fit by optimizing the ELBO defined in equation (14).

The SVGP approximation, defined in equation (15), is chosen because it has a very specific property; when plugged into equation (14), the objective decomposes into the sum of N separate terms. This allows us to create unbiased *minibatched* estimators of the objective (and its gradients), enabling efficient inference. We can see through a straightforward algebraic manipulation (we suppress ρ and $\mathbf{u}$'s dependence on $\mathbf{x}$ and $\bar{\mathbf{x}}$ to remove clutter)

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\boldsymbol{\rho},\mathbf{u})}\big[\ln p(\boldsymbol{\rho}, \mathbf{u}, \mathbf{y}) - \ln q_{\boldsymbol{\lambda}}(\rho, \mathbf{u})\big] \tag{16}$$

$$= \mathbb{E}_{q_{\boldsymbol{\lambda}}(\mathbf{u})p(\boldsymbol{\rho}|\mathbf{u})}\big[\ln p(\mathbf{y}|\boldsymbol{\rho}) + \cancel{\ln p(\boldsymbol{\rho}|\mathbf{u})} + \ln p(\mathbf{u}) - \ln q_{\boldsymbol{\lambda}}(\mathbf{u}) - \cancel{\ln p(\boldsymbol{\rho}|\mathbf{u})}\big] \tag{17}$$

$$= \underbrace{\mathbb{E}_{q_{\boldsymbol{\lambda}}(\mathbf{u})p(\boldsymbol{\rho}|\mathbf{u})}\big[\ln p(\mathbf{y}|\boldsymbol{\rho})\big]}_{\text{(i)}} - \mathrm{KL}\big(q_{\boldsymbol{\lambda}}(\mathbf{u})||p(\mathbf{u})\big), \tag{18}$$

and we can write the term (i) as a sum over the N observations

$$\text{(i)} \triangleq \mathbb{E}_{q_{\boldsymbol{\lambda}}(\mathbf{u})}\big[\mathbb{E}_{p(\boldsymbol{\rho}|\mathbf{u})}\big[\ln p(\mathbf{y}|\boldsymbol{\rho})\big]\big] \tag{19}$$

$$= \mathbb{E}_{q_{\boldsymbol{\lambda}}(\mathbf{u})}\left[\mathbb{E}_{p(\boldsymbol{\rho}|\mathbf{u})}\left[\sum_{n=1}^{N} \ln p(y_n|\rho_n)\right]\right] \tag{20}$$

$$= \sum_{n=1}^{N} \underbrace{\mathbb{E}_{q_{\boldsymbol{\lambda}}(\mathbf{u})}\big[\mathbb{E}_{p(\boldsymbol{\rho}|\mathbf{u})}\big[\ln p(y_n|\rho_n)\big]\big]}_{\triangleq \mathcal{L}_n}. \tag{21}$$

[5]Note that we are considering a model where the likelihood and prior are both Gaussian distributions, a conjugate pair, which implies that the posterior distribution will also be Gaussian. Indeed, it will be, but manipulating Gaussian will scale cubically in N. The SVGP approach is constructed to scale linearly in N.

When the likelihood $p(y_n|\rho_n)$ is Gaussian, the expectation in each of the $\mathcal{L}_n$ terms can be computed analytically. Notice the cancellation in equation (17) eliminates the term that involves all N data points and a $N \times N$ matrix inversion, $\ln p(\boldsymbol{\rho}|\mathbf{u})$ (Titsias (2009)).

To complete the derivation, we write a minibatched estimator of the ELBO. Given a set of randomly selected observations B,

$$\hat{\mathcal{L}}(\boldsymbol{\lambda}) = \frac{N}{|B|}\sum_{b\in B}\mathcal{L}_b - \mathrm{KL}\big(q_{\boldsymbol{\lambda}}(\mathbf{u})||p(\mathbf{u})\big), \tag{22}$$

which is an unbiased estimator of the full objective and only touches $|B| \ll N$ observations. We can use gradients of this estimator to optimize the lower bound with respect to variational parameters $\boldsymbol{\lambda} = (\mathbf{m}, \mathbf{S})$. With M inducing points, a minibatched estimator can be computed in $O(|B| \cdot M^3)$ time.

When predicting, note that we only condition on the inducing point values $\mathbf{u}$, relying only on the prior covariance between $\mathbf{u}$ and ρ_*. The structure of this variational approximation forces the inducing point values $\mathbf{u}$ to represent all of the information learned from the data. While this approximation is no longer "nonparametric," in the traditional sense, it has been shown that inducing point methods can provide good approximations to exact GP inference, provided enough inducing points are used (Burt, Rasmussen and van der Wilk (2019)). Empirically, we find that we can adequately tune covariance parameters and predict on held-out data using a modest number of inducing points (on the order of $M = 6000$).

Natural gradients. Given the objective in equation (18), we are now tasked with finding the optimal parameters $\boldsymbol{\lambda}$. We use stochastic optimization with minibatches (Hoffman et al. (2013)) along with natural gradients (Amari (1998)) which are effective for variational inference of Gaussian processes (Hensman, Fusi and Lawrence (2013)).

For exponential family distributions (e.g., the multivariate normal), the natural gradient can be computed by taking the standard gradient with respect to the mean parameters (Hoffman et al. (2013)). Following Hensman, Fusi and Lawrence (2013), the natural gradient for variational parameters $\boldsymbol{\lambda} = \mathbf{m}_{\boldsymbol{\lambda}}, \mathbf{S}_{\boldsymbol{\lambda}}$ is straightforward to express using the natural parameterization of the multivariate normal $q_{\boldsymbol{\lambda}}$; see Appendix C in the Supplementary Material (Miller et al. (2022)) for more details.

Whitened parameterization. Draws of $\mathbf{u} \sim p(\mathbf{u}|\bar{\mathbf{x}}, \boldsymbol{\theta})$ are highly dependent on $\boldsymbol{\theta}$. Consider a draw $\mathbf{u}$ conditioned on a fixed length scale parameter of ℓ. This same draw will be highly improbable under the prior with a different length scale parameter, for example, 2ℓ. The variational approximation $\boldsymbol{\lambda}$ targets the posterior distribution $p(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta})$ which is highly dependent on the structured Gaussian process prior. This dependence frustrates the joint inference of $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$; small changes in $\boldsymbol{\theta}$ can make the approximate distribution over $\mathbf{u}$ suboptimal. This dependence forces gradient methods to take small steps which leads to extremely slow joint inference.

An effective strategy for coping with this dependence is to do posterior inference in an alternative parameterization of the same model that decouples variables $\mathbf{u}$ and $\boldsymbol{\theta}$. For structured latent Gaussian models (e.g., Gaussian processes), this alternative model uses the *whitened* prior (Murray and Adams (2010)) or *noncentered* parameterization (Bernardo et al. (2003)). For the dust map model we whiten the prior by altering the target of posterior inference; instead of approximating the posterior over inducing point values $p(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta})$, we target the posterior over inducing point *noise* values $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$, where the relationship between $\mathbf{z}$ and $\mathbf{u}$ is constructed to produce an equivalent model. For notational clarity we define $\mathbf{K} \triangleq \mathbf{K}_{\mathbf{u},\mathbf{u}}^{(\boldsymbol{\theta})}$ for a fixed value of $\boldsymbol{\theta}$ and $\mathbf{L}$ such that $\mathbf{K} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$ is a general matrix square root. The whitened model can be written

$$\mathbf{z} \sim \mathcal{N}(0, I_M), \qquad \mathbf{u} \triangleq \mathbf{L}\mathbf{z} \sim \mathcal{N}(0, \mathbf{K}). \tag{23}$$

We now target the posterior distribution $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ by fitting an approximation $q_{\boldsymbol{\lambda}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\tilde{\mathbf{m}}, \tilde{\mathbf{S}})$. The linear relationship between $\mathbf{z}$ and $\mathbf{u}$ implies that $\mathbf{m} = \mathbf{L}\tilde{\mathbf{m}}$ and $\mathbf{S} = \mathbf{L}\tilde{\mathbf{S}}\mathbf{L}^\top$. We describe additional details of this parameterization in Appendix D in the Supplement (Miller et al. (2022)).

4.2. *Adapting to integrated observations.* The previous section detailed a scalable approximate inference algorithm for Gaussian processes in the typical noisy pointwise observation setting. Here, we adapt this framework to integrated observations. The challenge is now efficiently computing the appropriate semi- and doubly-integrated covariance values to be used within each minibatch update.

The natural gradient described in equation (39) (Appendix C) consists of two types of covariance matrices, $\mathbf{K}_{\mathbf{u},\mathbf{u}}$ and $\mathbf{K}_{\mathbf{u},\boldsymbol{\rho}}$. The former describes the prior covariance between inducing point values and remains exactly the same in the integrated observation setting. The latter describes the cross covariances between process locations and inducing point values; this cross covariance must be adjusted to reflect the covariance between the *integrated process* and the inducing point values. This is precisely what is defined by the semi-integrated covariance function. The entries of $\mathbf{K}_{\boldsymbol{\rho},\mathbf{u}}$ have to simply be switched to the semi-integrated covariance defined in Claim B1

$$(\mathbf{K}_{\boldsymbol{\rho},\mathbf{u}})_{n,m} = \int_{x \in R_n} k(x_m, x)\, dx. \tag{24}$$

In addition to the semi-integrated cross covariances, ZIGGY will need to calculate doubly-integrated covariance terms. First, note that the gradient defined in equation (39) (Appendix C) does not require doubly-integrated terms. This is convenient; the inducing point method does not require cross covariances between integrated observations, and these cross covariances also happen to be the most expensive to approximate with numerical methods. However, computing the variational objective (equation (18)) and making predictions (equation (78), Appendix E) do require the doubly-integrated diagonal terms, $\mathbf{K}_{n,n}$, for each observation.

Given the above dependence on integrated covariance terms, the full adaptation of the SVGP method to integrated observations runs into two technical hurdles. First, the semi-integrated covariance is not available in closed form beyond the simple squared exponential which is considered too smooth (and thus of limited capacity) in many settings. This limitation will make it difficult to use custom covariance kernels constructed with astronomical theory (Leike and Enßlin (2019)). Further, it will make it difficult to compare different models and choose the best predictor among them. Second, even when the semi-integrated covariance kernel is available in closed form, the doubly-integrated covariance is not. The SVGP (and any inducing point) algorithm only relies on *the diagonal* of the covariance kernel, so we only need to approximate this diagonal, scalar-valued function. We show that this scalar-valued function can be easily approximated using a small grid of interpolated values, where the "true" values are approximated using numerical integration.

We introduce two computationally efficient approximations to these operations on the covariance function to overcome these technical issues: (i) Monte Carlo estimates for semi-integrated covariance functions and (ii) linear interpolation for the doubly-integrated diagonal covariance function.

Monte Carlo estimators for semi-integrated kernels. The natural gradient updates in equation (39) (Appendix C) require computing the covariance between observations and inducing point values, $\mathbf{K}_{\boldsymbol{\rho},\mathbf{u}}$. For integrated observations, the covariance between observations and inducing points is given by the semi-integrated kernel in equation (24). For kernels with a closed form semi-integrated kernel (e.g., the squared exponential), we can simply substitute $k^{(\text{semi})}(\bar{x}_m, x_n)$ for $k(\bar{x}_m, x_n)$ in the loss and gradients and proceed within the typical SVGP framework.

For kernels without a closed form semi-integrated kernel, we run into a computational issue. How do we compute the appropriate cross covariance values? One approach is to use numerical integration—the semi-integrated covariance is a one-dimensional integral that can be approximated with quadrature techniques. However, each batch requires computing $M \times |B|$ semi-integrated covariance values; this can be computationally expensive, which would force us to use small batches, and this leads to slower learning.

Stochastic natural gradient updates are already an estimate of the true gradient. How *good* does this estimate have to be to effectively find the optimal solution? Numeric integration approximations to $k^{(\text{semi})}(\cdot, \cdot)$ are precise to nearly machine precision but are computationally expensive. Analogously, we can ask, how *good* do the semi-integrated covariance approximations have to be to effectively find the optimal solution?

Pursuing this idea, we propose using a Monte Carlo approximation to the semi-integrated covariance values; we sample uniformly along the integral path and average the original covariance values. More formally, we introduce a uniform random variable v_{R_n} that takes values along the ray R_n from the origin to x_n. We can now write the semi-integrated covariance as

$$k^{(\text{semi})}(x_m, x_n) = \int_{x \in R_n} k(x_m, x)\, dx = |R_n| \int_{v \in R_n} \underbrace{\frac{1}{|R_n|}}_{=p(v)} k(x_m, x)\, dx = |R_n| \mathbb{E}_v\big[k(x_m, v)\big].$$

This form admits a simple Monte Carlo estimator

$$v^{(1)}, \ldots, v^{(L)} \sim \text{Unif}(R_n), \qquad \hat{k}^{(\text{semi})}(x_m, x_n) = \frac{|R_n|}{L} \sum_{\ell} k\big(x_m, v^{(\ell)}\big). \tag{25}$$

It is straightforward to show that equation (25) is an unbiased and consistent estimator for the true semi-integrated covariance value evaluated at x_m and x_n. While plugging this directly into the gradient formula results in a biased estimator for the natural gradient, we find that with a modest number of samples we can closely match the optimization performance of the true natural gradient estimator. We investigate the relationship between samples L and the resulting optimization in Figure 9 (Appendix E). Appendix E.1 contains additional details and experimental results validating this approach.

Although conceptually similar, note that this Monte Carlo estimate of the semi-integrated covariance is not equivalent to discretizing along each line of site as in (Rezaei Kh. et al. (2017)). The estimator can use a small number of samples along the ray and still be useful within stochastic gradient optimization.

Integrated observation variance. The SVGP algorithm requires an additional covariance calculation—the marginal variance of an integrated observation at point x_n, $k^{(\text{doubly})}(x_n, x_n)$. Note that we do not need to compute the nondiagonal terms of the doubly-integrated kernel, a consequence of the separability of equation (21). Similar to the semi-integrated kernel function, the doubly-integrated function can be approximated with numerical methods—a double quadrature routine for each observation. But, similar to the semi-integrated case, this double quadrature call becomes a computational bottleneck within each batch. Furthermore, computing the gradient of the ELBO with respect to the covariance parameters with autodifferentiation tools would require differentiating through a double quadrature call—prohibitive in our setting.

We solve this problem by observing that a rotationally invariant (and stationary) kernel admits a doubly-integrated version that is a function only of the distances to each of the two integral endpoints. That is, $k^{(\text{doubly})}(x_i, x_j)$ can be written as a function $f(|x_i|, |x_j|)$ for a fixed setting of covariance function parameters. Further, we can write the diagonal of the doubly-integrated kernel as a function of only the distance to the single point, $k^{(\text{doubly})}(x_n, x_n) = f(|x_n|)$.

Given that this is a one-dimensional function shared by all N of our observations, we approximate the doubly-integrated kernel diagonal with linear interpolation. The interpolation scheme allows us to cheaply compute a highly accurate approximation to the doubly-integrated diagonal. It also enables us to easily backpropagate gradients to the kernel function parameters which enables efficient tuning during the variational inference optimization routine. Appendix E.2 describes the interpolation scheme for $k^{(\text{doubly})}(x_n, x_n)$ in further detail.

4.3. *Method summary and remarks*. To summarize, ZIGGY scales Gaussian process inference with integrated observations by fusing several strategies: (i) the explicit representation of the inducing point posterior within the SVGP framework, (ii) a whitened parameterization to decouple $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ in the variational objective, (iii) on-the-fly Monte Carlo estimates of the semi-integrated covariance function, (iv) and fast kernel interpolation of the doubly-integrated diagonal covariance function that leverages the stationary and isotropic properties of the covariance functions considered.

Algorithm 2 (in the Supplementary Material) describes the main loop for updating based on stochastic gradients of the variational parameters and the covariance function parameters θ. Maximizing the variational lower bound (equation (18)) with respect to prior parameters θ is a similar strategy to empirical Bayes (Efron (2008, 2019)), a method used to reduce the average error of posterior estimates.

We find that the inducing point approach dovetails nicely with integrated observations; most evaluations of the covariance function are between inducing points and observations. This only requires computing the (easier) semi-integrated covariance function. Further, the doubly-integrated covariance values between any two distinct observations never needs to be computed. Rather, only the *diagonal* of this covariance function needs to be computed, but this conveniently reduces to a one-dimensional function that can be efficiently approximated. ZIGGY circumvents computing the (prohibitive) $N \times N$ doubly-integrated covariance.

We note that the SVGP framework for scaling Gaussian process is one approach to build upon for our application; other frameworks for scaling kernel methods could have been employed. One example is random (or carefully selected) features for approximating the kernel function (Le, Sarlós and Smola (2013), Rahimi and Recht (2008), Yang et al. (2012)). While these approximate kernel methods could yield superior performance, one reason we reach for the SVGP framework is extensibility. The variational inference framework can incorporate more complex probabilistic graphical model structure. For this application, additional types of observations can be incorporated into a more sophisticated likelihood model to help identify the spatial dust distribution and incorporate additional sources of uncertainty.

5. Validation with a domain simulation. We developed ZIGGY with the goal of generating a three-dimensional dust map of the Milky Way using two billion stars in the Gaia dataset (Brown et al. (2018)). As a step toward that goal and to test ZIGGY in an applied environment, we infer the dust distribution of a mechanistic and physically motivated domain simulation.[6] The domain dataset is a synthetic Gaia survey of a high resolution Milky Way-like galaxy simulation.

The domain dust density $\rho(x)$ was generated in a Milky Way-like galaxy simulation, one of many in the Latte suite of simulations of Milky Way-mass galaxies (Wetzel et al. (2016), Hopkins et al. (2018)). It was run as part of the Feedback In Realistic Environments (FIRE) simulation project which selfconsistently models extinction and star formation in a cosmological context at high resolution.

[6]Code is available in the Supplementary Material (Miller et al. (2022)).

In astronomy, most observations prefer a model in which our universe is predominantly made of cold dark matter (Aghanim et al. (2018)). On small scales, however, there are significant challenges to this model (Klypin et al. (1999)), which this suite of simulations helps resolve (Wetzel et al. (2016)). This Latte simulation overcomes the computational challenge of including gas, dust, and stars with high enough resolution to resolve these tensions. It includes more complex physics of "normal" matter, as apposed to just dark matter, including a high-resolution disk of gas and dust. Because of the results from these simulations, which resolved these tensions, astronomers are more confident in a universal model of dark matter.

The simulation has enough resolution to resolve the formation of structures in the dense gas that forms dust and stars, creating a realistic environment for inference. The positions of the extinction observations, or the positions of stars within $\rho(\cdot)$, was generated by the Ananke framework. The Ananke framework generated a realistic, mock star catalog from the FIRE simulation and kept intact important observational relationships between gas, extinction, and stellar populations (Sanderson et al. (2018)). This catalog was specifically designed to resemble a Data Release 2 Gaia astrometric survey (Brown et al. (2018)), with a similar resolution of stellar density as Gaia. We integrated $\rho(x)$ along lines of sight to one million stars in the Ananke dataset within a 0.5 kpc $\times$ 0.5 kpc $\times$ 0.1 kpc region of the synthetic sun. So, we generate a set of N noisy integrated observations

$$x_n \sim \text{Ananke}(\mathcal{X}) \quad \text{stellar locations,} \tag{26}$$

$$e_n = \int_{x \in R_n} \rho(x)\, dx \quad \text{domain extinctions,} \tag{27}$$

$$a_n \sim \mathcal{N}\big(e_n, \sigma_n^2\big) \quad \text{noisy integrated observation,} \tag{28}$$

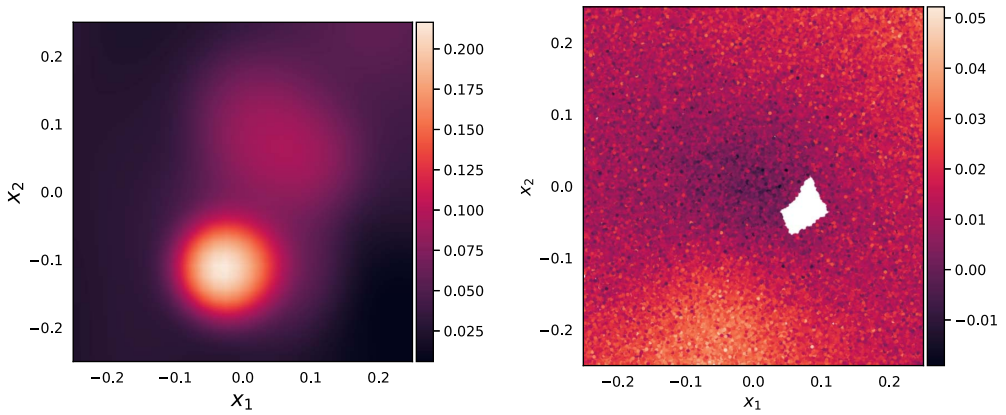
where the domain $\mathcal{X} = [(-0.25, 0.25), (-0.25, 0.25), (-0.05, 0.05)]$, the median extinction value is 0.01, and noise variance is chosen to be $\sigma_n = 0.005$. So our median signal-to-noise value is 3. The true extinctions, e_n, are computed to high precision using numerical quadrature, integrating from the origin to the Ananke stellar locations x_n. The integrated observations are depicted in Figure 2b. We estimate this domain specific dust density $\rho(\cdot)$ from noisy integrated observations using ZIGGY.

In Section 5.1 we compare posterior estimates of $\rho(\cdot)$ as a function of training data set size. This highlights the importance of scaling inference to more observations in order to obtain a more accurate statistical estimator. In Section 5.2 we compare posterior estimates of $\rho(\cdot)$ as a function of kernel choice with a fixed data set size.

5.1. *The quality of the estimate.* We study the quality of the estimate of $\rho(x)$ as a function of data set size N. We fit the dust model to the Ananke dataset with data set sizes $N \in \{10^3, 10^4, 10^5, 10^6\}$ for $10^5, 10^4, 10^3, 100$ epochs, respectively. We trained each model using minibatches of size $|B| = 2000$, an initial step size of 0.01, reducing it every epoch by multiplying by a decay factor of 0.99. We measure model quality by computing root mean squared error (RMSE) and log-likelihood (LL) on a set of $N_{\text{test}} = 2000$ held out extinction values. Each model uses $M = 16 \times 16 \times 4$ inducing points, evenly spaced in a grid in the input space.

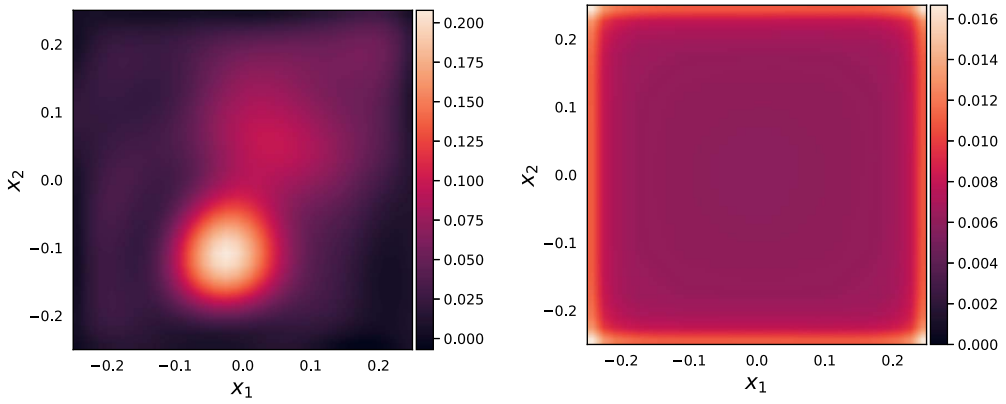
Figure 2 summarizes the model's fit using one million observations, the most we tried in this example. Figure 2c displays the posterior mean for $\rho(x)$, given the one million observations depicted in 2b. Figure 2d shows the posterior uncertainty (one standard deviation) about the estimated mean. We see that the million-observation model recovers the true latent function particularly well.

Figure 3 quantifies the improvement in the RMSE and log-likelihood on a held out sample of test stars as a function of dataset size N. As the dataset size increases, both statistics



(a) True $\rho(x)$

(b) Integrated observations a_n



(c) Inferred $\mathbb{E}[\rho \mid \mathcal{D}]$.

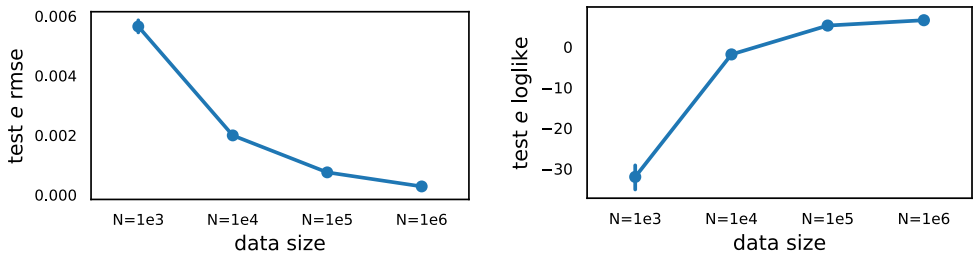
(d) Inferred posterior scale $\mathbb{V}[\rho(x) \mid \mathcal{D}]^{1/2}$.

FIG. 2. *Integrated observation GPs can reveal spatial structure from a severely limited vantage point. This figure summarizes domain simulated data posterior inference using* $N = 1{,}000{,}000$ *examples in a three-dimensional space. Panel* 2*a depicts the true (unobserved)* $\rho(x)$ *that generates the data. Panel* 2*b depicts the observed data, noisy integrated measurements of* ρ *from the origin to stellar positions in the domain simulation.* 2*c depicts the inferred* ρ *given these observations using* ZIGGY *with a squared exponential kernel.* 2*d depicts the marginal posterior standard deviation at each location in* $\mathcal{X}$*-space. All Panels are a slice at* $z = 0$.

drastically improve. Figure 4 visualizes a direct comparison of the posterior estimate of the latent function, as we increase data set size N. We can clearly see that, as we incorporate more data into the nonparametric model, the form of the true underlying function emerges. To accurately visualize and interpret large scale features of the latent dust distribution, we will want to incorporate as many stellar observations as possible. This is particularly true in the low signal to noise regime of extinction estimates.

5.2. *Comparing kernels.* Similar to the synthetic dataset, we compare multiple models by fitting the variational approximation and tuning the covariance function parameters for five different kernel choices.

In this domain setup we ran each model for 100 epochs, saving the the model with the best ELBO value. We used minibatches of size $|B| = 2000$ and started the step size at 0.01, reducing it every epoch by multiplying by a decay factor of 0.99. For kernels that do not have a closed form semi-integrated version, we used $L = 50$ uniform grid Monte Carlo samples to estimate the integrated covariance.

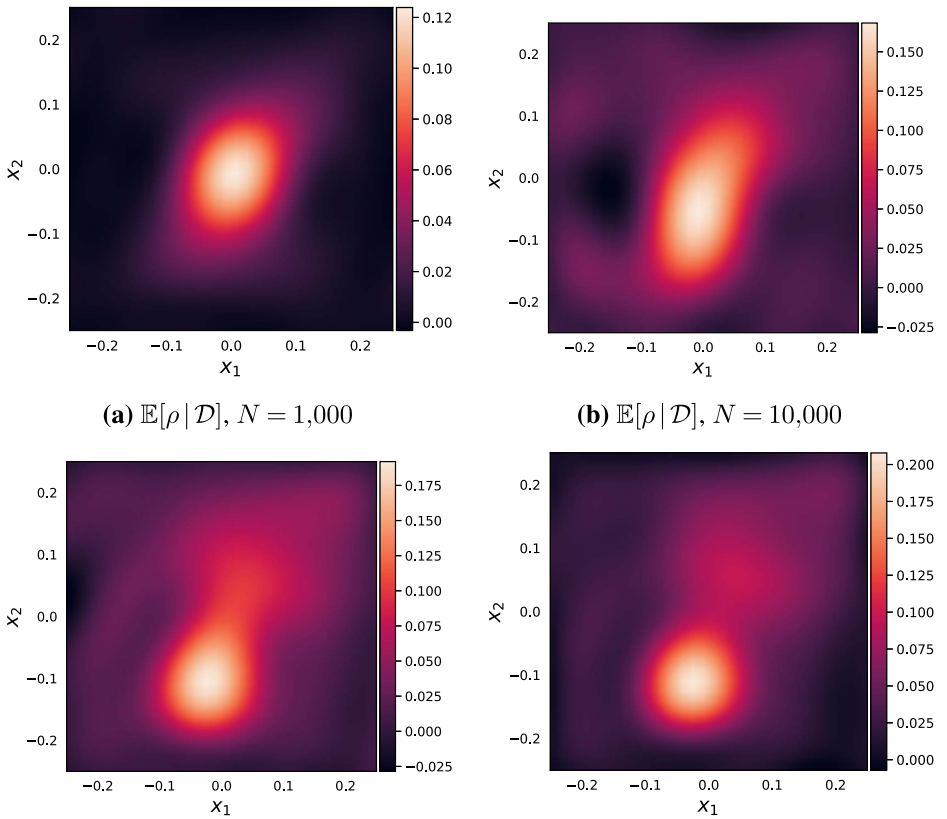


(a) Test e RMSE **(b)** Test e log likelihood

FIG. 3. *Scaling to massive N improves estimator performance. We compare predictive RMSE* (*left*) *and log–likelihood* (*right*) *as a function of data set size N on a held out sample of test stars.*

We validate the inferences against a set of 2000 held-out test extinctions e_*. We quantify the quality of the inferences with mean squared error, log-likelihood, and the χ^2-statistic on the test data. We also visually validate ZIGGY with a QQ-plot and coverage comparison, also on test data. We also visualize the behavior of ZIGGY as a function of distance to the origin; specifically, we visualize how accurate and well-calibrated test inferences are, as synthetic observations are made farther away.

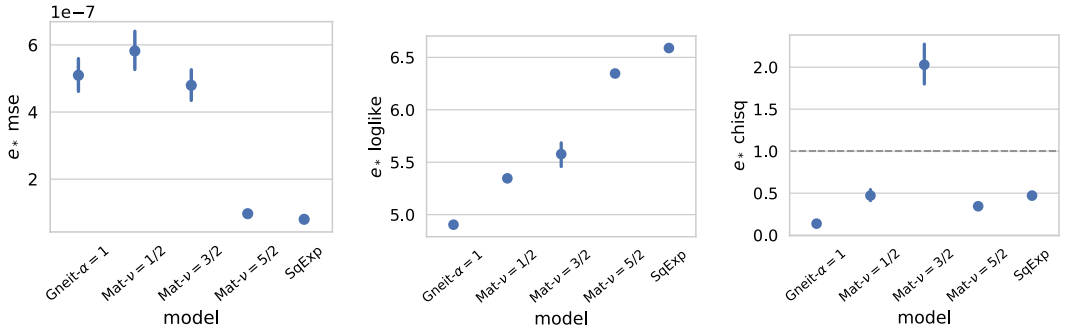
In Figure 5 we compare the mean squared error (MSE), the log-likelihood, and the χ^2 statistic on the held-out test data. We compare five different kernels: (i) Gneiting $\alpha = 1$ from



(a) $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 1{,}000$ **(b)** $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 10{,}000$

(c) $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 100{,}000$ **(d)** $\mathbb{E}[\rho \,|\, \mathcal{D}]$, $N = 1{,}000{,}000$

FIG. 4. *Slices of the posterior mean function at* $z = 0$ *for various dataset sizes. More data leads to more accurate predictions of* $\rho(\cdot)$.



(a) Test e MSE **(b)** Test e log likelihood **(c)** Test e χ^2

FIG. 5. *Predictive summaries across different models and inference schemes. The smoother models have better predictive summary statistics which is consistent with the underlying density field being very smooth.*

Rezaei Kh. et al., 2018, (ii)–(iv) Matérn, $\nu \in \{1/2, 3/2, 5/2\}$, and (v) the squared exponential kernel. The results match our expectations. The true $\rho(\cdot)$ is quite smooth, and we see that the smoother kernels tend to have lower predictive error and higher log-likelihood. All kernels have similarly calibrated χ^2 statistic, except for Matérn, $\nu = 3/2$, which underpredicts it's posterior variances. This can also be seen in Figure 6. The test statistic comparison has chosen a good match in the squared exponential kernel.

To test the calibration of posterior uncertainties, we inspect statistics of test-sample z-scores for e_*. We test that the statistics of predictive z-scores resemble a standard normal distribution using a QQ-plot. For a more detailed description of the QQ-plot, see Appendix F.2. To compare different models, we visualize QQ-plots for the different covariance functions in Figure 6. As an additional summary of calibration, we compare the fraction of predicted examples covered by $1/2$, *one*, *two*, and three posterior standard deviations, summarized in Table 1. We see that there is room for improvement in the posterior variances. The z-scores and empirical coverage show posterior variances that are well calibrated for the squared exponential and Matérn, $\nu = 5/2$ kernels but are too large for the other kernels. The models for the less smooth kernels are compute bound, requiring thousands of inducing points to tile the domain. We ran the models for 100 epochs which maxed out the memory and compute time. Miscalibration could be the result of misspecification of the model itself, for example,

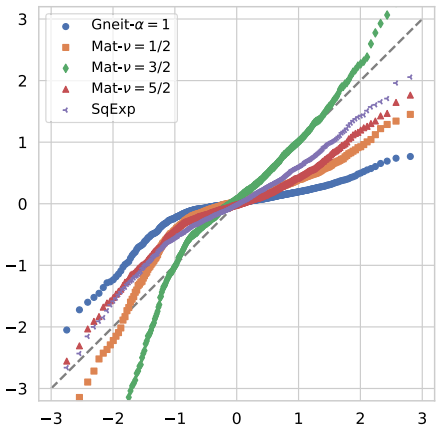


FIG. 6. *QQ-plot for predicted distributions on* 2000 *held out integrated-ρ test points e_*. Theoretical normal quantiles are on the horizontal axis, and predictive z-scored quantiles are on the vertical axis. Posterior for extinctions e are moderately well calibrated, with the squared exponential kernel having the best calibration.*

TABLE 1

Coverage fractions at various levels of z-score standard deviation σ for test extinctions e_*. Compared with theoretical coverage fractions, we see that our inferences generate predictions that are well calibrated for the smoothest kernels

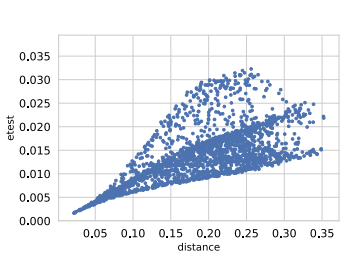
σ	Gneit-$\alpha = 1$	Mat-$\nu = 1/2$	Mat-$\nu = 3/2$	Mat-$\nu = 5/2$	SqExp	$\mathcal{N}(0, 1)$
0.5	0.894	0.760	0.451	0.730	0.625	0.383
1.0	0.965	0.900	0.677	0.900	0.869	0.683
2.0	0.997	0.969	0.878	0.991	0.986	0.955
3.0	1.000	0.994	0.947	1.000	0.999	0.997

mismatch between the covariance function k and the true underlying ρ or the inflexibility of the variational approximation. To diagnose this shortcoming, methods that allow us to scale the expressivity of the variational approximation, for example, larger M would be necessary.

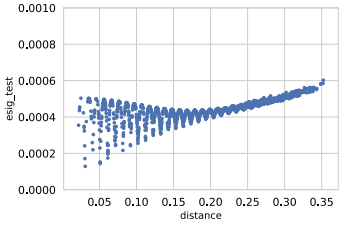
We also depict model fit statistics as a function of distance from the origin in Figure 7. Figure 7a visualizes the test extinction values. ZIGGY is able to accurately reconstruct the dust map. Figure 7b shows the posterior standard deviation by distance which shows a noise reduction of about $10\times$ for most stars. In Figure 7c we see that the predictions have posterior variances that are slightly too high, causing z-scores that lie mostly within 2σ independent of distance, with only the most nearby stars extending to 3σ. Decent calibration, including at far distances, is encouraging, as the goal is to form good estimates of this density far away from the observer location.

6. Discussion. Mapping the three-dimensional distribution of dust in the Milky Way is fundamental to astronomy. Star dust obscures our observations of star light, and so an accurate dust map would help us more clearly observe the universe. Such a dust map would be a new lens into the dynamics of the Milky Way, providing a trace of the process by which it formed.

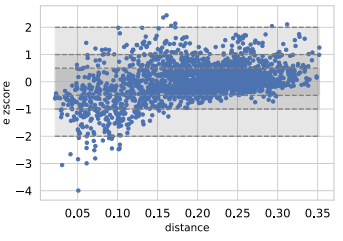
In this paper we developed ZIGGY, a method to estimate the dust map from millions of astronomical observations. ZIGGY incorporates noisy information into a single, coherent spatial model. We developed technical innovations to accommodate integrated observations into the Gaussian process framework and to scale such a model to millions of observations without spatial discretization. We validated these algorithmic innovations with numerical studies, and measured the performance of ZIGGY in both a synthetic setting and a realistic, high-fidelity dataset used in astronomical study. We showed that ZIGGY is able to incorporate millions of observations and recover accurate and well-calibrated estimates of both stellar extinctions and pointwise values of the three-dimensional dust map.



(a) Test e values by distance



(b) Test e posterior standard deviation



(c) Test e z scores by distance

FIG. 7. *Posterior summaries as a function of distance to the origin (observation point). The noise variance for the extinction estimates is* $\sigma_n = 0.005$, *so Figure (a) demonstrates our low signal to noise regime, Figure (b) demonstrates our shrinkage by a factor of* 10, *and Figure (c) demonstrates our decent calibration in this very low signal to noise regime.*

There are several avenues for future research and improvement. ZIGGY assumes that distances are known, when, in reality, they are noisy. In practical applications we will only include stars with precise distance measurements, lowering the spatial density of observations, which may degrade estimates. We view ZIGGY as a step toward a probabilistic model of photometric measurements that jointly infers brightnesses, colors, distance, and spatial dust. Jointly modeling brightnesses, distances, and a spatial dust map using a Bayesian hierarchical model will allow us to leverage the coherence of Bayesian probabilistic modeling to form tighter estimates of all modeled quantities.

Additionally, a limitation of our treatment is the Gaussian error assumption. Non-Gaussian observation errors could degrade extinction estimates, for example, forcing ρ to be more complex (i.e., with a smaller length scale) than it truly is. A flexible hierarchical Bayesian model could incorporate a non-Gaussian likelihood that could be more robust to heavier tailed observation error.

A full map that spans the entirety of the Milky Way will require increasing the capacity of the approximation. To reconstruct both global and local features using an inducing point method, we need to ensure that small length scales (relative to the input domain) can be resolved. When inducing points are spatially distant from one another, the inducing point approximation will not have the capacity to represent sharp changes in $\rho(x)$ at small scales. This limitation can be overcome by including more, and closer, inducing points. However, introducing more inducing points butts up against the $O(M^3)$ computational limitations. In this work we assume we are able to populate the space with enough inducing points, but the scaling concern is an important avenue of future work. Incorporating more inducing points will be crucial to resolving both global features and fine local features within the Milky Way.

The ultimate goal of this line of work is to produce an accurate catalog of the properties of stars, such as probabilistic brightnesses and distances, and the focus of scale is motivated by the size of modern photometric catalogs. PAN-STARRS 1 catalog includes 2.4 million detected stars (Flewelling (2016)), the fifteenth data release of the Sloan Digital Sky Survey includes photometry for over 260 million detected stars (Aguado et al. (2019)), and the second GAIA data release has 1.3 billion stars with parallax measurements that can be used to estimate the interstellar dust distribution (Brown et al. (2018)). As a final avenue of research, we will scale ZIGGY to incorporate billions of observations in a much larger spatial domain.

SUPPLEMENTARY MATERIAL

Appendices (DOI: 10.1214/22-AOAS1608SUPPA; .pdf). Appendices to the main text.

Code (DOI: 10.1214/22-AOAS1608SUPPB; .zip). Implementation of the method presented.

REFERENCES

AGHANIM, N. et al. (2018). Planck 2018 results. VI. Cosmological parameters. E-prints. Available at arXiv:1807.06209.

AGUADO, D. S., AHUMADA, R., ALMEIDA, A., ANDERSON, S. F., ANDREWS, B. H., ANGUIANO, B., ORTÍZ, E. A., ARAGÓN-SALAMANCA, A., ARGUDO-FERNÁNDEZ, M. et al. (2019). The fifteenth data release of the Sloan Digital Sky Surveys: First release of MaNGA-derived quantities, data visualization tools, and Stellar Library. *Astrophys. J., Suppl. Ser.* **240** 23.

AMARI, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Comput.* **10** 251–276.

AMBIKASARAN, S., FOREMAN-MACKEY, D., GREENGARD, L., HOGG, D. W. and O'NEIL, M. (2014). Fast direct methods for Gaussian processes. Preprint. Available at arXiv:1403.6015.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184

BERNARDO, J., BAYARRI, M., BERGER, J., DAWID, A., HECKERMAN, D., SMITH, A. and WEST, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In *Bayesian Statistics* 7: *Proceedings of the Seventh Valencia International Meeting* **307**. Oxford Univ. Press, USA.

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BROWN, A., VALLENARI, A., PRUSTI, T., DE BRUIJNE, J., BABUSIAUX, C., BAILER-JONES, C., BIERMANN, M., EVANS, D. W., EYER, L. et al. (2018). Gaia data release 2-summary of the contents and survey properties. *Astron. Astrophys.* **616** A1.

BURT, D. R., RASMUSSEN, C. E. and VAN DER WILK, M. (2019). Rates of convergence for sparse variational Gaussian process regression. Preprint. Available at arXiv:1903.03571.

CHEN, B. Q., HUANG, Y., HOU, L. G., TIAN, H., LI, G. X., YUAN, H. B., WANG, H. F., WANG, C., TIAN, Z. J. et al. (2019). The galactic spiral structure as revealed by O- and early B-type stars. *Mon. Not. R. Astron. Soc.* **487** 1400–1409. https://doi.org/10.1093/mnras/stz1357

CRESSIE, N. (1990). The origins of Kriging. *Math. Geol.* **22** 239–252. MR1047810 https://doi.org/10.1007/BF00889887

CRESSIE, N. (1992). Statistics for spatial data. *Terra Nova* **4** 613–617.

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706 https://doi.org/10.1080/01621459.2015.1044091

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866 https://doi.org/10.1214/07-STS236

EFRON, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Statist. Sci.* **34** 177–201. MR3983318 https://doi.org/10.1214/18-STS674

FLEWELLING, H. (2016). The pan-STARRS 1 medium deep field variable star catalog. In *American Astronomical Society Meeting Abstracts #*227. *American Astronomical Society Meeting Abstracts* **227** 144.25.

FOREMAN-MACKEY, D., AGOL, E., AMBIKASARAN, S. and ANGUS, R. (2017). Fast and scalable Gaussian process modeling with applications to astronomical time series. *Astron. J.* **154** 220.

GENTON, M. G. (2002). Classes of kernels for machine learning: A statistics perspective. *J. Mach. Learn. Res.* **2** 293–312. MR1904760 https://doi.org/10.1162/15324430260185637

GEORGELIN, Y. M. and GEORGELIN, Y. P. (1976). The spiral structure of our Galaxy determined from H II regions. *Astron. Astrophys.* **49** 57–79.

GNEITING, T. (2002). Compactly supported correlation functions. *J. Multivariate Anal.* **83** 493–508. MR1945966 https://doi.org/10.1006/jmva.2001.2056

GREEN, G. M., SCHLAFLY, E. F., FINKBEINER, D., RIX, H.-W., MARTIN, N., BURGETT, W., DRAPER, P. W., FLEWELLING, H., HODAPP, K. et al. (2018). Galactic reddening in 3D from stellar photometry–an improved map. *Mon. Not. R. Astron. Soc.* **478** 651–666.

GREEN, G. M., SCHLAFLY, E. F., ZUCKER, C., SPEAGLE, J. S. and FINKBEINER, D. P. (2019). A 3D dust map based on gaia, pan-STARRS 1 and 2MASS. Preprint. Available at arXiv:1905.02734.

HEATON, M. J., DATTA, A., FINLEY, A. O. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. MR3996451 https://doi.org/10.1007/s13253-018-00348-w

HENSMAN, J., FUSI, N. and LAWRENCE, N. D. (2013). Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* 282–290. AUAI Press.

HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

HOPKINS, P. F., WETZEL, A., KEREŠ, D., FAUCHER-GIGUÈRE, C.-A., QUATAERT, E., BOYLAN-KOLCHIN, M., MURRAY, N., HAYWARD, C. C., GARRISON-KIMMEL, S. et al. (2018). FIRE-2 simulations: Physics versus numerics in galaxy formation. *Mon. Not. R. Astron. Soc.* **480** 800–863. https://doi.org/10.1093/mnras/sty1690

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KLYPIN, A., KRAVTSOV, A. V., VALENZUELA, O. and PRADA, F. (1999). Where are the missing galactic satellites? *Astrophys. J.* **522** 82–92. https://doi.org/10.1086/307643

KRIGE, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. S. Afr. Inst. Min. Metall.* **52** 119–139.

LE, Q., SARLÓS, T. and SMOLA, A. (2013). Fastfood-computing Hilbert space expansions in loglinear time. In *International Conference on Machine Learning* 244–252.

LEIKE, R. and ENSSLIN, T. (2019). Charting nearby dust clouds using Gaia data only. Preprint. Available at arXiv:1901.05971.

MATÉRN, B. (1960). Spatial variation: Stochastic models and their applications to some problems in forest surveys and other sampling investigations. *Meddelanden från Statens Skogsforskningsinstitut* **49** 1–144.

MATHERON, G. (1963). Principles of geostatistics. *Econ. Geol.* **58** 1246–1266.

MATHERON, G. (1973). The intrinsic random functions and their applications. *Adv. in Appl. Probab.* **5** 439–468. MR0356209 https://doi.org/10.2307/1425829

MATHIS, J. S. (1990). Interstellar dust and extinction. *Annu. Rev. Astron. Astrophys.* **28** 37–70.

MILLER, A. C, ANDERSON, L., LEISTEDT, B., CUNNINGHAM, J. P, HOGG, D. W and BLEI, D. M (2022). Supplement to "Mapping interstellar dust with Gaussian processes." https://doi.org/10.1214/22-AOAS1608SUPPA, https://doi.org/10.1214/22-AOAS1608SUPPB

MURRAY, I. and ADAMS, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems* 1732–1740.

QUIÑONERO-CANDELA, J. and RASMUSSEN, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* **6** 1939–1959. MR2249877

RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* 1177–1184.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

REZAEI KH., S., BAILER-JONES, C., HANSON, R. and FOUESNEAU, M. (2017). Inferring the three-dimensional distribution of dust in the Galaxy with a non-parametric method-preparing for Gaia. *Astron. Astrophys.* **598** A125.

REZAEI KH., S., BAILER-JONES, C. A. L., HOGG, D. W. and SCHULTHEIS, M. (2018). Detection of the Milky Way spiral arms in dust from 3D mapping. *Astron. Astrophys.* **618** A168. https://doi.org/10.1051/0004-6361/201833284

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SANDERSON, R. E., WETZEL, A., LOEBMAN, S., SHARMA, S., HOPKINS, P. F., GARRISON-KIMMEL, S., FAUCHER-GIGUÈRE, C.-A., KEREŠ, D. and QUATAERT, E. (2018). Synthetic Gaia surveys from the FIRE cosmological simulations of Milky-Way-mass galaxies. E-prints. Available at arXiv:1806.10564.

SCHLEGEL, D. J., FINKBEINER, D. P. and DAVIS, M. (1998). Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500** 525.

SNELSON, E. and GHAHRAMANI, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems* 1257–1264.

TITSIAS, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics* 567–574.

WAINWRIGHT, M. J., JORDAN, M. I. et al. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WETZEL, A. R., HOPKINS, P. F., HOON KIM, J., FAUCHER-GIGUÈRE, C.-A., KEREŠ, D. and QUATAERT, E. (2016). Reconciling Dwarf Galaxies with ΛCMD cosmology: Simulating a realistic population of Satellites Around a Milky Way Mass Galaxy. *Astrophys. J.* **827** L23. https://doi.org/10.3847/2041-8205/827/2/l23

YANG, T., LI, Y.-F., MAHDAVI, M., JIN, R. and ZHOU, Z.-H. (2012). Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems* 476–484.