PoPE: A Population-based Approach to Model the Spatial Structure of Astronomical Systems

Arya Farahi\textsuperscript{1,2}, Daisuke Nagai\textsuperscript{3,4}, and Yang Chen\textsuperscript{5}

\textsuperscript{1} The Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, USA; arya@umich.edu
\textsuperscript{2} Department of Physics and Leinweber Center for Theoretical Physics, University of Michigan, Ann Arbor, MI 48109, USA
\textsuperscript{3} Department of Physics, Yale University, New Haven, CT 06520, USA
\textsuperscript{4} Department of Astronomy, Yale University, New Haven, CT 06520, USA
\textsuperscript{5} Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

Received 2020 July 5; revised 2020 October 8; accepted 2020 October 16; published 2020 December 15

Abstract

We present a novel population-based Bayesian inference approach to model the average and population variance of the spatial distribution of a set of observables from ensemble analysis of low signal-to-noise-ratio measurements. The method consists of (1) inferring the average profile using Gaussian processes and (2) computing the covariance of the profile observables given a set of independent variables. Our model is computationally efficient and capable of inferring average profiles of a large population size from noisy measurements, without stacking data or parameterizing the shape of the mean profile. We demonstrate the performance of our method using dark matter, gas, and stellar profiles extracted from hydrodynamical cosmological simulations of galaxy formation. POPULATION PROFILE ESTIMATOR is publicly available in a GitHub repository. Our new method should be useful for measuring the spatial distribution and internal structure of a variety of astrophysical systems using large astronomical surveys.

Unified Astronomy Thesaurus concepts: Astronomy data modeling (1859); Astronomy data analysis (1858); Spatial point processes (1915); Astrostatistics (1882); Open source software (1866)

1. Introduction

The spatial distribution and internal structure of astronomical systems contain vast amounts of information about the underlying physics that governs the formation, evolution, and fate of these systems. While astronomical data collected by large-, medium-, and small-size surveys are becoming more abundant, precise, and accurate, modeling is becoming more challenging. The scale and complexity of these multiwave-length surveys are exceeding the capabilities of traditional data analysis models, hence the need for novel inference models.

One of the key challenges of modeling the empirical data is how to account for the measurement errors of varying magnitudes (i.e., “heteroscedastic” errors; see also Kelly 2007). The low signal-to-noise-ratio (S/N) regime hinders the ability to infer the spatial structure of a population from abundant but noisy measurements, diluting the spatial signals. Typical measurements with S/N below the detection limit are often discarded or stacked to boost the signal above the detection limit (e.g., Bulbul et al. 2014; Greco et al. 2015; Mezcua et al. 2016; McClintock et al. 2019; Okabe et al. 2019; Umetsu & Diemer 2017). Binning and stacking can introduce selection bias (Dietrich et al. 2014) and information loss (Arzner et al. 2007), and smear out the signal component (Kipping 2010). While stacking amplifies the S/N of the population’s average properties, it suppresses the intrinsic scatter of the population under study. In practice, performing a statistical inference on large astronomical data sets has become a bottleneck of traditional population- and likelihood-based approaches.

In this work, we present POPULATION PROFILE ESTIMATOR (POPE), a population-based, Bayesian inference model to analyze a class of problems that are concerned with the spatial distribution or internal spatial structure of a sample of astronomical systems (Diemand & Moore 2011; Conselice 2014). We adopt a multivariate normal distribution to model the spatial profile of a $d$-dimensional noisy observable, with errors of varying magnitude, as a function of radial distance and global properties of the astronomical system. A multivariate normal distribution is fully described with a mean function and a covariance matrix. POPE is a two-step analysis tool that (1) reconstructs the average profile using nonparametric regression with Gaussian processes (GPs) and (2) estimates the property profiles’ covariance given a set of independent variables. Our population-based method is computationally efficient and is capable of inferring the average profiles of a population from noisy measurements, without stacking or parameterizing the shape of the average profile.

The structure of this paper is as follows. In Section 2, we set up our model. In Section 3, we assess the performance of our model in a controlled numerical experiment and show that our model can accurately reconstruct the input average profile and the covariance matrix from low-$S/N$ measurements. In Section 4, we present a novel application of our model in studying the internal structure of dark matter halos. We discuss the modeling of discrete observables in Section 5.2 and broader applications of our model in Section 5.1. Finally, we conclude this work in Section 6.

Definitions: Scalar variables are in italic and vectors are in bold italic. The only matrices in this work are $\mathbf{R}$, $\mathbf{\Sigma}$, and $\mathbf{\Sigma}_{\text{err}}$. We denote logarithm base $e$ by ln and logarithm base 10 by log. Unless otherwise noted, the reported confidence regions and error bars are 68% uncertainty intervals.

2. A Population Statistics Model

Property profiles, denoted by $x$, are the primary observables, which are a function of $x$, a set of independent variables. A property profile refers to the spatial distribution of an observable around a population of astronomical objects or the internal spatial observables of a population of astronomical objects.
systems. For example, $s$ can be the number density profile of galaxies around a set of galaxies or the hot gas temperature profile of dark matter halos. A key element of $x$ is a position vector. In our applications, we will focus on modeling “radial profiles”; hence, the first element of $x$ is the physical or angular distance from the center of the astronomical objects in question, such as galaxies or dark matter halos. Our model, however, is general and can be applied to any geometry and coordinate system. The property profiles can also be a function of other parameters (e.g., local matter density or mass of the host halo, respectively). Our aim is to model the conditional statistics of $s_i$ given $x$: i.e., $P(s_i|x_i)$, $P(s_j|x_i)$ is modeled with a multivariate normal distribution which can be specified with a mean vector $i$ and the covariance matrix $Cov(s_i|x_i)$. The notation and model parameters are defined in Table 1.

We propose a two-step inference model. In the first step, we infer the mean relation per data point (i.e., $\langle s_i | x_i \rangle$) for data point $i$. The second step is to estimate the conditional covariance matrix given the posterior on the mean relation inferred from the first step. In Section 2.1, we setup the notation used throughout this work, then in Sections 2.2 and 2.3, we explain the two steps of our model. The implementation considerations are discussed in Section 2.4.

2.1. Notation Setup

We shall denote the independent variable by $x$ and the dependent variable by $s$. In statistics literature, $x$ and $s$ are also referred to as the “covariate” and the “response,” respectively. $x$ and $s$ are $N$-dimension and $M$-dimension vectors, respectively. We do not observe the actual value of $s$; instead, $s$ is measured by measurement noise. The measured quantities and their noise are denoted by $\tilde{s}$ and $\Sigma_{\text{err}}$, respectively.

In the following, $\langle s \rangle$ denotes the expectation value or the mean of $s$ given $x$. $\langle s | x \rangle$ denotes the random variable $s$ conditioned on $x$. $i$ is the index over data points, and $s_i$ and $\langle s_i \rangle$ imply $\langle s_i | x_i \rangle$ and $\langle s_i | x_i \rangle$, respectively. The property profile vector of an astronomical system is a random variable,

$$\langle s_i | x_i \rangle = \langle s_i \rangle + \epsilon(x_i),$$

where $\epsilon_i \equiv \epsilon(x_i)$ is a random variable described with a multivariate normal distribution with mean zero. $\epsilon$ defines the intrinsic covariance of property profiles. The covariance can also be a function of our independent variable.

As mentioned earlier, only a noisy version of $s_i$ is observed. The observed property profile is another random variable

$$\langle \tilde{s}_i | s_i \rangle = s_i + \epsilon_{\text{err}}(x_i),$$

where $\epsilon_{\text{err}}$ is the measurement noise vector, which is a random variable drawn from a multivariate normal distribution with mean zero. In Section 5.2, we provide examples of observables for which this assumption does not hold and discuss potential solutions.

Because of the multivariate normal assumption in Equation (1), our model can be fully described with two quantities: (1) the conditional average profile $\langle s | x \rangle$ and (2) the conditional covariance $Cov(s|x) = \Sigma$. Our goal is to estimate these two quantities in two steps by employing a data-driven model in which the form of $\langle s | x \rangle$ is not specified, and the constraints on the amplitude, shape, and its dependence on the independent variables are inferred from the data.

2.2. Inferring the Mean Relation Using a Gaussian Process

The key feature of our model is employing a GP prior to reconstruct the average property profile. A large class of smooth functions can be reconstructed from a GP prior without the need to explicitly parameterize the shape of the curves (Williams & Rasmussen 2006). Its computational tractability and analytical features make it a suitable choice for our problem. A primary advantage of GPs is that it can capture nonlinear and nonmonotonic behaviors in the average profiles as a continuous function of the independent variables. As opposed to binning and stacking strategies, GPs enable the investigator to capture the impact of multiple variables simultaneously.

There has been a surge in applications of GPs in astronomy and astrophysics data analysis, including modeling of asteroseismic data (Brewer & Stello 2009), galactic black hole light curves (Kelly et al. 2011), and cosmic microwave background (Wandelt & Hansen 2003; Jewell et al. 2004); reconstructing the Hubble constant (Melia & Yennapureddy 2018); analyzing time domain data (Foreman-Mackey et al. 2017); estimating photometric redshift (Way et al. 2009; Almosallam et al. 2016); and reconstructing the probability distribution (McCintock & Rozo 2019). In this work, we apply GPs to model the average profiles as a function of multiple input variables $x$ in a continuous fashion, and it can capture nonlinear and nonmonotonic trends, which is impractical with binning and stacking strategies.

Specifically, we fit $\langle s_j | x \rangle$ for each observable $j$ independently. Without loss of generality, we drop the index $j$ in the following. We employ a GP to model the mean relation as a function of the distance and other independent observables. A GP is a prior probability distribution whose domain is over the space of a continuous function (Williams & Rasmussen 2006). The advantage of this model is that it does not require a parameterization of the density profile, and thus is nonparametric;

| Table 1: Notations |
|----------------------|
| **Parameter** | **Explanation** | **Category** |
| $x$ | Independent variables vector | Input variable |
| $s$ | Property profiles vector | Random variable |
| $\tilde{s}$ | Observed properties profile vector | Random variable |
| $\langle s | x \rangle$ | Conditional average property profiles vector, conditioned on $x$ | Model parameter |
| $\Sigma$ | Conditional property profiles covariance matrix, conditioned on $x$ | Model parameter |
| $\Sigma_{\text{err}}$ | Measurement error matrix | Constant |
| $l$ | Scale factor in our Gaussian process covariance function | Hyperparameter |
| $\sigma_{\text{gp}}$ | Uncertainty on the mean | Hyperparameter |
| $\sigma$ | Average population scatter for the profile of a property | Hyperparameter |
| $i$ | Index over data points | Index |
| $j$ | Index over the vector of property profiles | Index |
| $k$ | Index over the vector of independent variable | Index |
| $n$ | Number of data points | … |
| $N$ | Dimension of vector $x$ | … |
| $M$ | Dimension of vectors $s$ and $\tilde{s}$ | … |
a binning and stacking strategy is not required; and finally, due to its analytical properties, the posterior has a closed-form solution, under the assumption that the model hyperparameters are known, and thereby is computationally tractable. Additionally, almost any smooth function can be reconstructed from a GP prior without the need to explicitly define the shape of the curve.

A GP prior defined on the space of functions \( s(x) \)

\[
<s(x)> \sim \mathcal{GP}(m(x), k(x, x')).
\]

The function values are modeled as a draw from a multivariate normal distribution that is parameterized by the mean function, \( m(x) \), and the covariance function (also known as the kernel function), \( k(x, x') \). Due to the marginalization and conditioning properties of the multivariate normal distribution, GPs are a convenient choice as priors over functions. Even though the goal of this work is not a prediction for a new set of observations, we note that the marginal distribution over a new point \( x_* \) can be evaluated easily (see Equation (24) in Quiñonero-Candela & Rasmussen 2005). Here, the GP prior is defined to draw the expected \( s \) as a function of input variables, \( x \). We now need to model the property profile scatter about the conditional population average property profile. The property profile is a random variable that follows Equation (1) where \( \epsilon \) is a random variable with mean zero. We assume that \( \epsilon \) is drawn from a normal distribution with a constant variance \( \sigma^2 \) that is not a function of independent variables. We will revisit this assumption in Section 2.3. Thus, \[ s(x) \sim \mathcal{N}(\hat{s}(x), \sigma^2), \] where \( \sigma \) is a free parameter in our model.

The actual value of \( s(x_i) \) is not directly observable. We observe a noisy estimation of \( s(x_i) \), which is denoted by \( \hat{s}(x_i) \). If the measurement noise has a normal distribution, \( \hat{s}(x_i) \) becomes

\[
(\hat{s} | s) \sim \mathcal{N}(s, \sigma^2_{err}),
\]

where the mean value falls at the true density, and the \( \sigma^2_{err} \) is the measurement uncertainty, which can vary and be a function of \( x_i \). We assume that the measurement uncertainties are known a priori. Now, we have all ingredients to build the likelihood of the observed property profile \( \hat{s} \) given \( x \).

The observed profile for data point \( i \) can be described with the following set of equations,

\[
\langle s_i \rangle | x_i \sim \mathcal{GP}(m(x_i), k(x_i, x'_i)),
\]

\[
(s_i | \langle s_i \rangle) \sim \mathcal{N}(\langle s_i \rangle, \sigma^2),
\]

\[
(\hat{s}_i | s_i) \sim \mathcal{N}(s_i, \sigma^2_{err}).
\]

Because the marginalization over \( s_i \) is analytical, the above model simplifies to

\[
\langle s_i \rangle | x_i \sim \mathcal{GP}(m(x_i), k(x_i, x'_i)),
\]

\[
(\hat{s}_i | \langle s_i \rangle) \sim \mathcal{N}(\langle s_i \rangle, \sigma^2_{err} + \sigma^2).
\]

**GP Setup:** We set the mean function in the GP to zero, \( m(x) = 0 \). The property profile can be considered as a stationary process (i.e., there is a constant normalization). The best strategy to model a property profile is to renormalize the measured profiles of a population by subtracting the average profile. We assume that a vector \( s \) is properly normalized, which justifies \( m(x) = 0 \). For the kernel function, we employ the Matérn kernel (Genton 2001). The class of Matérn kernels is a generalization of the radial basis kernel functions and the absolute exponential kernel functions parameterized by an additional parameter \( \nu \), which specifies the smoothness of functions generated by the GP prior. We set \( \nu = 5/2 \), a prior over a set of two times differentiable functions. This is an important feature. In certain applications (see Section 5.1 for examples), we want to be able to compute the second derivative of \( s \) with respect to distance or log distance. Thus, we want our function to be at least twice differentiable. The form of this kernel function is

\[
k(x, x') = \sigma^2_{gp} \times \prod_{k=1}^{N} \left( 1 + \frac{\sqrt{5}(x_k - x'_k)^2}{l_k} + \frac{5(x_k - x'_k)^2}{l_k^2} \right)^{\frac{\nu}{2}} \times \exp \left( -\sqrt{5} \frac{(x_k - x'_k)}{l_k} \right),
\]

where \( N \) is the dimension of independent variable vector, and \( x_k \) is the \( k \)th independent variable. This kernel function has two sets of free parameters, \( \sigma_{gp} \) and \( l_k \), \( \sigma_{gp} \) is the overall amplitude of the GP covariance, and \( l_k \) specifies the correlation scale for the \( k \)th independent variable. We emphasize that even though the GP does not parameterize the space of potential curves, the kernel function and the mean function used to draw curves are parametric.

For illustration, Figure 1 shows several examples of one-dimensional curves randomly drawn from a GP prior for \( \sigma_{gp} = 1 \) and \( l_i \). The scale factor \( l \) sets the smoothness of the function class drawn from a GP prior. A GP prior with small \( l \) produces small-scale fluctuations, and a large \( l \) produces smooth curves. If there are not enough data to constrain the small-scale behaviors, a small-scale factor results in a noisy posterior.

The kernel function \( k(x, x') \) is only used to define the prior distribution and should not be confused with the covariance matrix that describes the property profile correlation in a population. In our examples, the results are insensitive to the small variation in covariance matrix hyperparameters and even the choice of the covariance matrix. The user needs to perform a sensitivity analysis in their application.

At the end of the inference, the only quantity that is important is the average profile at a given \( x \). Thus, we shall compute the posterior over \( \langle s(x) \rangle \) for every single data point \( i \). We perform this inference per property profile \( j \), independently. To compute the posterior distribution, we have to specify the prior for each free parameter.

Under certain weak conditions, the posterior consistency of the GP is guaranteed (Choi 2005), which means that as the sample size increases, the posterior distribution will concentrate around the true value of the parameter (Rousseau 2016).

**Priors Setup:** Our model has in total \( 2 + N \) free parameters, where \( N \) is the dimension of the independent variable vector. There are three categories of free parameters: (1) the scale parameters in the GP kernel function \( l_i \), (2) the uncertainty on the average property profile \( \sigma_{gp} \), and (3) the average population conditional scatter \( \sigma \).

We use a delta function prior for \( l_i \), which controls the smoothness of the GP generated curves. For large sample sizes,
the exact value of \( l_k \) becomes less relevant. However, setting an extreme \( l_k \) (where the true solution is in the tail of the GP prior) results in a biased solution. Large (small) \( l_k \) results in underfitting (overfitting) the model. The smoothing scale can be found by performing experiments. It should be chosen based on the range of data, number of data points, and measurement uncertainties that vary for different applications. In our examples, we set \( l_k = [\max(x_{ij}) - \min(x_{ij})] \). We find that the final results are insensitive to variation in the \( l_k \) scales. But, the user must perform a sensitivity analysis and choose a scale that is appropriate for their application and data.

We set a weakly informative prior on the two most important quantities in this model: \( \sigma_{gp} \) and \( \sigma \), as they are physically relevant quantities. \( \sigma_{gp} \) can be interpreted as the interface uncertainty on the average profile curve given a set of data points, and \( \sigma^2 \) is the population average profile variance at fixed \( x \). We employ a half-normal distribution with large width with respect to the population variance as a prior for these two parameters. This prior is a standard choice for variance in Bayesian inference (e.g., Carpenter et al. 2017).

In Equation (10), \( \sigma \) is independent of \( x \), which is not strictly true, and in most applications such as ours is inaccurate. Even though including a population variance in our model is necessary, the constant assumption does not have an impact on the outcome of the inferred average profile. We model the conditional dependence of the population property profile variance \( \sigma \) on \( x \) in Section 2.3, so we will discard \( \sigma \) in our GP model after the inference.

### 2.3. Inferring the Covariance Matrix

Our second aim is to infer the conditional covariance matrix of a set of property profiles \( s \) at fixed \( x \). In this section, we bin, but not stack, the data and compute a conditional covariance matrix for each bin independent of other bins. In the following, we discard the bin index on \( s \) and \( x \); hence, the index on the bin is implicitly assumed. The binning construction defines on which independent variables the covariance matrix is conditioned on and on which independent variables it is averaged over. For example, if the data are binned in radius, then the estimated covariance is marginalized over halo properties and conditioned on the radial distance. It is further assumed that the covariance matrix is a slowly varying function of \( x \) (i.e., it is almost constant within each bin).

Alternatively, one may parameterize \( \Sigma \) as a function of \( x \) and constrain the model parameters instead of binning and independently estimating the covariance matrix for each bin. A parametric approach, if known, might be preferred over the binning approach. In our applications, we do not know a well-defined parametric form a priori, so we model \( \Sigma \) using the binning approach. The bin size depends on the application and data. We assume that the covariance matrix in a given bin does not rapidly change, and there are enough measurements to obtain meaningful constraints on the covariance matrix.

The data consist of a vector of observed property profiles denoted by \( s \) (a random vector of \( M \) dimensions) in a given bin of \( x \) and their corresponding measurement error covariance. We assume that the conditional distribution of \( s \) given \( x \) is described by a multivariate Gaussian distribution,

\[
(s \mid \langle s \rangle, \Sigma) \sim \mathcal{N}(\langle s \rangle, \Sigma).
\]  

The mean and the covariance of this conditional distribution are denoted by an \( M \) vector \( \langle s \rangle \) and an \( M \times M \) matrix \( \Sigma \), respectively. We do not observe the actual values of the vector \( s \), but instead observe values of \( \hat{s} \), which are measured with error \( \Sigma_{err} \). The measured quantity is assumed to be drawn from a multivariate Gaussian distribution,

\[
(\hat{s} \mid s) \sim \mathcal{N}(s, \Sigma_{err}).
\]  

We marginalize over \( s \), and the result is another multivariate Gaussian distribution. Each data point \( i \) is generated by a multivariate Gaussian distribution,

\[
(\hat{s}_i \mid \langle s \rangle_i, \Sigma) \sim \mathcal{N}(\langle s \rangle_i, \Sigma_{err,i} + \Sigma),
\]  

where \( i \) is the index over the data, \( \Sigma_{err,i} \) is the error covariance for data point \( i \). The expected mean profile \( \langle s \rangle_i \) is a function of independent variables \( x_i \).

Given a set of observations \( \hat{s}_i \), we want to estimate the posterior distribution on the covariance matrix. The next step is to specify prior distributions for \( \langle s \rangle_i \) and \( \Sigma \).

**Setting up Priors:** To set up the priors on the mean property profiles \( \langle s \rangle_i \) per data point, we employ the posteriors estimated in Section 2.2. We estimate the prior on each \( \langle s \rangle_i \) with a normal
distribution,

$$(s)_{ij} \mid x_i \sim \mathcal{N}(\mu_{ij}, \sigma_{\mu,ij}^2),$$

with mean $\mu_{ij}$ and variance $\sigma_{\mu,ij}^2$ of data point $i$ and observable $j$. $\mu_{ij}$ and $\sigma_{\mu,ij}^2$ are computed from the mean and variance of the posterior estimate of the GP model. If we ignore the uncertainty on the hyperparameters of the GP prior, the posterior distribution on $\mu_{ij}$ is a multivariate normal distribution. We employ the “maximum a posteriori” point estimation to estimate the hyperparameters of the GP (see Section 2.4 for more discussions).

We then take the diagonal components of the GP posterior on $\langle s \rangle_{ij}$. This implies that our prior is broader than it needs to be. In our examples and applications, the posterior variance on $\langle s \rangle_{ij}$ is small compared to the measurement uncertainties and the intrinsic variance (see Section 5.3 for more discussions).

Therefore, the contribution of off-diagonal components of the prior to $\mu_{ij}$ is negligible. Implementing the full covariance adds to the computational complexity of our model, while its impact on the posterior distribution of $\Sigma$ is negligible. To keep the model simple and computationally efficient, we will use the diagonal components only in this work. The full covariance is an $n \times n$ matrix, where $n$ is the number of data, which becomes computationally infeasible for large data. Diagonalizing the uncertainty in the covariance matrix reduces the computational burden at the cost of inflating the uncertainty (by a few percent) on the estimated average profiles. In our applications, we show that the uncertainty on the average profiles is significantly smaller than the intrinsic scatter and the measurement errors of the population; therefore, by a few percent overestimation of the uncertainties of the average profile, there will be inconsequential loss in the constraining power.

Constructing a proper prior for the covariance matrix is discussed in great detail in the statistics literature (e.g., Barnard et al. 2000; Lewandowski et al. 2009; Alvarez et al. 2014). Sampling a high-dimensional covariance matrix is a difficult task, especially if the prior is not chosen carefully. We employ Gelman et al.’s (2015) recommended approach, which decomposes $\Sigma$ into a correlation matrix $\mathbf{R}$ and a scale vector $\tau$ (see also Barnard et al. 2000):

$$\Sigma = \text{diag}(\tau) \mathbf{R} \text{diag}(\tau).$$

$\tau$ is a vector of the standard deviations of the hyperparameter $\mu$, which describes the population mean. The prior on $\tau$ is taken to be an inverse-Gamma distribution with shape $\alpha$ and rate $\beta$,

$$\tau \sim \text{inv-Gamma}(\alpha, \beta) \propto \tau^{-2\alpha-2} \exp\left(-\beta\tau^{-2}\right).$$

This prior is chosen so as to prevent divergences in the sampling while allowing large values of the variance.

A Lewandowski–Kurowicka–Joe (LKJ) distribution prior is used on the correlation,

$$P(\mathbf{R} \mid \nu) \propto \det(\mathbf{R})^{\nu-1},$$

where the shape parameter $\nu > 0$. This distribution converges toward the identity matrix as $\nu$ increases (sparse covariance matrix), allowing the control of the correlation strength between the multiple parameters and consequently the variance and covariance of parameters in the population. A flat prior for marginalized elements of the correlation matrix can be imposed by setting $\nu = 1$ and for $0 < \nu < 1$, the density has a trough at the identity matrix, which is not desired. Figure 2 compares the LKJ prior with $\nu = 1$, 2, and 4.

The correlation matrix $\mathbf{R}$ is decomposed into its Cholesky factor $\mathbf{L}_R$ and its transpose $\mathbf{L}_R^T$,

$$\mathbf{R} = \mathbf{L}_R \mathbf{L}_R^T,$$

where an LKJ prior parameterized in terms of the Cholesky decomposition has been imposed.

In our applications, we do not want to force sparsity on the covariance matrix to ensure that physical models can produce large correlations between independently measured properties, hence no large values of $\nu$. Ideally, the shape factor should be close to 1, i.e., marginally noninformative. We find that with $\nu = 2$, the posterior samples convergence faster than $\nu = 1$, and in our toy models, we recovered the input correlation. Thus, we set $\nu = 2$, which keeps the prior on the correlation matrix marginally weakly informative.

Rather than using the Gamma or truncated-normal prior on the scale and the LKJ prior on the correlation, it is more common in these sorts of hierarchical analyses to set the prior on $\Sigma$ to be the scaled inverse Wishart distribution (e.g., Sellentin & Heavens 2016). This choice is usually made for its analytical tractability or conjugacy on Gaussian likelihoods and simplicity within Gibbs sampling. However, this distribution undesirably assumes a prior relationship between the variances and correlations (see Alvarez et al. 2014), which is ill suited for our application. Our model does not have a closed-form solution, and we need to sample the posterior distribution. In a sampling method, conjugate priors are not necessary. It is worth noting that the combined scale and LKJ prior can be sampled more efficiently and gives us control over the diagonal elements of $\Sigma$. In our experiments, we assume the same set of hyperparameters for all bins, but a user of PoPE may choose different hyperparameters for each bin. However, the posterior distribution on the covariance matrix of each bin is determined independently of the other bins.
2.4. Implementation and Computational Considerations

The main computations consist of (1) deriving the mean relation using GP (Sections 2.2) and (2) computing the covariance matrix (Section 2.3). The pseudo-code for these two steps are provided in Algorithms 1 and 2. The entire model is implemented in \texttt{PyMC3} (Patil et al. 2010), which has an implementation of GPs, “Fully Independent Training Conditional” (FITC) sparse approximation, “Maximum a Posteriori” (MAP) estimation, and the No-U-Turn sampler (NUTS) algorithm (Hoffman & Gelman 2014).

The output of the first step is the joint posterior on the average profiles for all observations, \( P(\tilde{y}_i|\beta, \mu, \sigma) \). Sampling the full posterior, for large sample sizes, can become computationally infeasible and in most applications, like ours, is unnecessary. We instead estimate the posterior density employing optimization algorithms. If we set the value of \( \sigma_{gp} \) and \( \sigma \), the posterior on \( \tilde{y}_i \) become analytically tractable. For a fixed \( \sigma_{gp} \) and \( \sigma \), the joint posterior \( P(\tilde{y}_i|\beta, \mu, \sigma) \) is a multivariate normal distribution. We employ the \texttt{find_MAP} function in \texttt{PyMC3} for a point estimation of \( \sigma_{gp} \) and \( \sigma \), which uses the Broyden–Fletcher–Goldfarb–Shanno optimization algorithm (Fletcher 2013), a fast converging, iterative method for solving unconstrained nonlinear optimization problems. We use MAP estimates to compute the GP posteriors of the average property profiles. Fixing the hyperparameters and estimating the marginal posterior on \( \tilde{y}_i \) are sufficient for most applications.

One of the limiting factors of general GPs is their computational cost (Foreman-Mackey et al. 2017). Evaluating a general GP likelihood scales as the cube of the number of data points \( O(n^3) \) times the dimensional of independent variables, which can become intractable even for small-size data. With the current era of deep and wide surveys, we expect as many as \( \sim 10^4 \)–\( 10^9 \) profile measurements for a typical sample of galaxies or clusters of galaxies. For example, there are \( \sim 7 \times 10^3 \) optically selected clusters in the Dark Energy Survey Year 1 data, and there are \( \sim 14 \) radial bins with weak lensing shear measurements (McClintock et al. 2019). As a result, there would be total of \( \sim 10^5 \) measurements. With \( O(n^3) \) scaling in computational costs, evaluating the posterior density is a computationally demanding task.

To address this computational bottleneck, we utilize sparse approximation methods (see Quiñonero-Candela & Rasmussen 2005 for a review of GP sparse approximations). Specifically, we employ the so-called FITC approximation method (Quiñonero-Candela & Rasmussen 2005; Snelson & Ghahramani 2006). This sparse approximation does not form a full covariance matrix over all \( n \) data inputs. Instead, it relies on a set of \( m \)-inducing points, where \( m \ll n \). This sparse approximation reduces the \( O(n^3) \) complexity of GPs down to \( O(mn^2) \)—which makes the MAP estimation and average profile posterior estimation tractable. The inducing points are denoted by \( x_u \). We place these points “uniformly” in each dimension throughout the domain of the independent variables, \( x_u = x_{u,1} \times x_{u,2} \times \cdots x_{u,N} \), where \( x_{u,k} \in [\min(x_k), \cdots \max(x_k)] \) and \( \min(x_k) \), \( \max(x_k) \) are, respectively, the minimum and maximum of the \( k \)th independent variable in our sample. \( m_u \), the number of inducing points on the domain of the independent variable \( x_u \) is another hyperparameter of our model that needs to be fixed. In our application, we choose \( m_u \) to be between 8 and 16 points. We find that the resultant estimated posteriors on the average profiles are consistent with the input models.

We pass the estimated posterior mean and variance to our covariance matrix inference model. Our model does not have a closed-form solution, hence we need to sample the posterior distribution. Sampling the posterior distribution is fast enough that there is no need for an approximation. To sample the posterior distribution, we employ the NUTS algorithm (Hoffman & Gelman 2014), an extension of the Hamiltonian Monte Carlo sampling algorithm that eliminates the need to set a number of steps.

Algorithm 1. The conditional average property profile inference algorithm.

1. **Input**: \( x, \beta, \sigma_{gp} \), a set of independent variables, observed profiles, and uncertainty of the observed profiles.
2. **Output**: Compute the posterior mean and variance of the average profile per \( x \).
3. Set the hyperparameters \( l_x \).
4. Specify the priors on the hyperparameters \( \sigma \) and \( \sigma_{gp} \).
5. Specify the \( x_u \) array.
6. Estimate MAP for the model parameters \( \sigma \) and \( \sigma_{gp} \).
7. Compute the conditional posterior mean and variance of the average profile at \( x \) per data point \( i \).

Algorithm 2. The conditional covariance matrix inference algorithm.

1. **Input**: \( x, \beta, \Sigma_{err} \), a set of bins, and the posterior mean and variance of the average profile per \( x \).
2. **Output**: posterior samples of the covariance matrix per bin.
3. Construct the prior on \( \Sigma(x) \).
4. Initialize the hyperparameters \( \nu, \alpha, \) and \( \beta \).
5. for \( b \) in \( \{1, \cdots, \text{nbins}\} \)
6. Find the subset of data in bin \( b \).
7. Pass the subset to the likelihood.
8. Draw posterior samples using the NUTS algorithm.
9. Check convergence.
10. end for

3. Model Performance: A Controlled Simulation Experiment

This section provides an example application of the model discussed above and assesses its performance on a toy model. Our goal is to illustrate that our inference model can recover an input model.

We start with two exponentially decaying profiles. One profile has a core and the other one is cuspy. The coefficients are arbitrary, but we chose to keep the range of \( \ln(\rho) \) within two orders of magnitude. The input average profiles are

\[
\langle \ln \rho_1(r) \rangle = -\frac{r^2}{5} - 2 \frac{r}{5} + 1,
\]

and

\[
\langle \ln \rho_2(r) \rangle = -\frac{r^3}{5} + 2 \frac{r}{5}.
\]

The independent vector \( x \) consists of only one element \( r \), thereby \( N = 1 \), and the property profiles vector has two elements \( s = \{ \ln(\rho_1), \ln(\rho_2) \} \), thereby \( M = 2 \). We uniformly sample the independent variable \( r \) 4000 times and compute the
average profiles ($\ln \rho_i(x)$). We then add correlated Gaussian intrinsic noise with variance and correlation coefficient of 0.25 and 0.5, respectively. This makes a random realization of true profiles. The observed property profiles are computed by adding Gaussian noise to the true values. The width of the measurement noise is random and drawn from a uniform distribution. The maximum and minimum of the uniform distribution are tuned to confine the S/N between 1 and 3. This allows us to assess the performance of our model in a low-S/N regime where all observations are below a detection threshold of 5.

Figure 3 shows a realization of the true profiles (the top two left panels) and observed profiles (the top two right panels). The input average profiles, specified in Equations (21) and (22), are in black dashed lines. The true profiles are gray points in the left panels, and the observed profile measures with their 68% measurement errors are shown in the right panels of Figure 3. We pass the simulated observed profiles and their errors to our inference models and compare the inferred quantities with the input model.

Figure 4 compares the posterior average profiles (the blue and red lines) for $\ln(\rho_1)$ and $\ln(\rho_2)$ (top and bottom panels, respectively) with the input average profiles (dashed black line). The shaded regions show the 68% posterior intervals. The true profiles, gray points, are shown for reference. If the posterior interval for the average profile is significantly smaller than the intrinsic scatter of the data, the intrinsic scatter and the correlations can be detected with high S/N.

We take the output of our GP model and pass it as a prior to the covariance matrix estimator. The data are binned in 10 evenly spaced bins in $r$. There are on average 400 measured quantities in each bin. Then, we run our inference model and evaluate the posterior distribution in each bin. Figure 5 shows the median and 68% posterior intervals of the correlation coefficients (the blue points) per bin, and the yellow line is the input correlation coefficient. This figure shows that with 400 measurements of property profiles with S/N < 3, we can put reasonable constraints on the correlation coefficient.

4. An Application to Astronomical Data Analysis

One of the major challenges in the era of large multi-wavelength astronomical surveys lies in accurately modeling the spatial structures of astronomical systems for broad mass, redshift, and radial ranges (e.g., Battaglia et al. 2019; Salcedo et al. 2020; Shirasaki et al. 2020). To maximize the scientific returns of these surveys, it is important to develop a novel technique to measure the density profiles of dark matter, gas, and stars and their coupling strength, which in turn contain rich information about the baryonic physics and its impact on the structure of dark matter halos (e.g., Gnedin et al. 2004;
Nagai et al. 2007; Duffy et al. 2010; Schaller et al. 2015; Schneider & Teyssier 2015; Cui & Zhang 2017; Schneider et al. 2019; Huang et al. 2020; Li et al. 2020. Here we employ simulations to perform a forecast. We add a measurement noise to true quantities measured from simulations and show that the correlation between density profiles indeed can be estimated using low-S/N measurements.

**Sample:** We employ gas and stellar density profiles derived from the TNG-100 solution of the IllustrisTNG project. The simulation outputs are provided publicly by the IllustrisTNG team (Nelson et al. 2018, 2019; Pillepich et al. 2018; Springel et al. 2018). TNG is employed because it produces reasonable galaxies and clusters (Barnes et al. 2018; Nelson et al. 2018).

Gas and stellar density profiles measured at $R_{200}$ normalized radius from the center of halos. Our halo sample consists of a subset of halos with mass $10^{12} M_\odot$ and above at redshift $z = 0$. Halos are identified using a “friends-of-friends” percolation algorithm. Our observables are gas and stellar differential density profiles. Gas and stellar density profiles are computed at different spherical shells with respect to the center of a halo. The center of halos are at the minimum of the local gravitational potential. We construct a mock observed sample by adding Gaussian noise to true density profiles (see Figures 6 and 7).

**Benchmark Reference:** We fit a profile curve using a Kernel Localized Regression method motivated by Cleveland (1979). Our approach is to fit a locally linear, but globally nonlinear, relation to a pair of random variables $x$ and $s$ where $s$ is a multidimensional random variable. We note that in our notation $x$ is the independent variable and $s$ is the dependent variable. To perform this regression, we employ the Kernel Localized Linear Regression (KLLR: https://github.com/afarahi/kllr) implementation motivated by Farahi et al. (2018). KLLR employs a weighted least-squares fitting where the weights are assigned a kernel (weight) function. This method allows us to model and identify nonlinear behaviors of the density profile as a function of distance. The width of our Gaussian kernel is 0.15 (dex).

KLLR models $(s|x)$ in a continuous fashion. We further bin our data into several halo mass bins to study the dependence of this conditional distribution on halo mass. We employ 1000 bootstrap realizations of the halo sample to compute the statistical uncertainty intervals. We note that the current implementation of KLLR cannot handle measurement uncertainties. Therefore, we can only pass true quantities to this model. We use the KLLR fit as the reference to compare with the output of the proposed inference model.

A limitation of the current implementation of KLLR is that it does not currently fit density profiles to multivariable input data. To overcome this limitation, we bin our data into three mass bins and report the fit and correlations for halos in each bin. KLLR assumes that all halos in the same bin, regardless of their mass, has the same average profile. Mixing of halos in any given mass bin tends to smear out mass-dependent effects and biases the resulting average profile and correlation coefficients.
The GP model, however, fits a continuous function to multivariable input data. The bias induced by our binning in the average profile is on the order of a few percent (see the mathematical derivation in Evrard et al. 2014), and it suppresses the correlation coefficient.

**Findings:** The blue and red dots in Figure 6 shows the true density profile of our sample; the observed quantities are not shown here, but the distribution of S/Ns of measured quantities is presented in Figure 7. Figure 7 shows the distribution of S/Ns for the mocked observations. We pass the mocked observations and their uncertainties to our inference model. We intentionally keep the S/N small to show that even with noisy measurements, our model is capable of reconstructing the true density profile.

Figure 6 compares the inferred average density profiles as a function of the distance from the center of halo and their host halo mass. We compare these results by the KLLR fits to the true density profiles in three different mass bins. The results are presented in Figure 6. The right panels are the KLLR fits to the true density profiles and the left panels are the GP fits to the noisy mocked density profiles. We note that the average density profile posterior intervals are shown in Figure 6, but in certain cases, the 68\% posterior intervals are smaller than the width of the mean posterior lines.

In Figure 8, we compare the results of the inferred correlation coefficient from our model and the estimated correlation coefficient by the KLLR method. The results from two methods are in agreement. There are a few features that are worth pointing out. For the most massive bin, there are only 168 halos and the S/Ns of the mocked density profiles are pretty small, so our model cannot constrain the correlation coefficient and the posteriors are scattered around zero. For the smallest mass bin that have a large number of halos, there is enough signal to measure the correlation coefficients, and the uncertainty on the correlation coefficients from our model is pretty small. The inferred correlations from our model is consistently larger than the correlation estimated by the KLLR method. To compute the correlations, we need an unbiased estimation of the average profile as a function of mass and distance. Because the KLLR method fit a single average profile to all halos in a given mass bin, there will be additional scatter due to the mass dependence in the density profile per bin. This induced scatter suppresses the value of estimated correlation coefficients. But our GP model infers the average density profile as a function of mass and radii, so the correlations are not suppressed by mixing halos of different masses in a single bin. This example illustrates another advantage of our model over traditional models in which data are binned and all data in a bin treated equally. In our inference model, we still bin the data to infer the correlation, but the key to computing the covariance between two parameters is having an unbiased estimation of the average. In our inference model, the averages are continuous function of mass and radius, so the estimated correlations are not suppressed in each bin.

5. Other Applications and Limitations

In this section, we review additional potential applications and limitations of our population-based method.

5.1. Derived Quantities from Profiles

It is often easier to observe the projected and cumulative profile (e.g., enclosed mass of halos within a sphere of radius \(r\)), while differential quantities are more interesting theoretically. Curves sampled from a GP can be treated like any other curve; it can be integrated, differentiated, or combined with other quantities to derive new properties. With algebraic operations, GP posterior curves, defined on the space of observables, can be transformed to more physically relevant quantities. To compute the derived quantities, the simplest approach would be to draw curves from the posterior densities and for each curve numerically compute the quantity of interest. For instance, a set of posterior mass profiles can be differentiated and divided by the differential volumes to get the posterior on differential densities. It comes with certain limitations. The GP curves are not always differentiable, and the uncertainties grow for higher-order quantities (e.g., the uncertainties on the second-order derivative of a function is larger than uncertainties on the first-order derivative of a function). We provide two examples.

**Example 1:** One of the interesting quantities of the density profiles are their logarithmic slope, \(\frac{\partial \log(\rho)}{\partial \log(r)}\). In Section 4, we inferred the average stellar and gas log-density profiles for TNG-100 halos. From the GP posteriors, 400 posterior average log-density profile curves are drawn. Then, for each profile, the logarithmic slope is numerically computed. Figure 9 shows the inferred logarithmic slope of the gas and stellar density profiles. The logarithmic slopes are noisier than average profiles, but we can identify interesting features. For example, the slope of the stellar mass density profiles within the virial radius of halos is nearly constant, and the slope for the high-mass systems is shallower (\(\sim -3\)) than for the low-mass end (\(\sim -4\)). The advantage of our GP model is that it can capture nonlinear, nonmonotonic trends such as the logarithmic slope of the gas density profiles (see Figure 9).

**Example 2:** A similar approach can be applied to dwarf galaxies in order to model the logarithmic slope of their total mass density profiles. At small scales, the cold-dark matter model predicts that the inner slope of the density profile follows \(\rho_{\text{dm}} \propto r^{-\gamma}\) (Navarro et al. 1997). Empirical measurements of the density profile of dwarf spheroidal galaxies suggest shallower slopes and are often consistent with a constant-density core at the
center. This disagreement between observations and simulations has become known as the core–cusp problem (de Blok 2010; Bullock & Boylan-Kolchin 2017). With the proposed model, one can model the average density profile of a population of galaxies and study how the inner slope changes as a function of galaxy’s observables.

5.2 Modeling Discrete Observables

In some applications in astronomy and astrophysics, the profile observables are discrete (e.g., the number density of halo satellite galaxies). In these applications, our noise model requires modification. Specifically, to model the measurement noise of discrete quantities, Equation (5) must be modified. For example, if the discrete counting measurement is well represented by the Poisson process, then the Gaussian distribution assumed earlier can be substituted with a Poisson distribution,

\[(\hat{\sigma} \mid x) \sim \text{Poisson}(\lambda = s). \tag{23}\]

While high-energy observables are promising probes of the nature of dark matter, these observables are often photon deficient (e.g., Drlica-Wagner et al. 2015; Lee et al. 2016). A proper model of these photon-deficient observations should account for noise due to small numbers in counts of photons.

Similarly, measuring galaxies number counts in groups and clusters of galaxies is affected by discrete, small number statistics. The spatial distribution of satellite galaxies in dark matter halos is one of the key ingredients for modeling the galaxy–dark matter connection and interpreting data from large galaxy surveys (e.g., Gao et al. 2004; Nagai & Kravtsov 2005; Sales et al. 2007; Piscione et al. 2015; Ágústsson & Brainerd 2018). Measuring the spatial distribution of satellite galaxies per halo and inferring their density profile lying in discrete, small number statistics regime require extra care in modeling.

Another application of discrete spatial profile measurement is inferring the dark matter distribution via the distribution of stars in star clusters or nearby dwarf galaxies (e.g., Merritt & Tremblay 1994; Nilakshi et al. 2002; Jurić et al. 2008; Seleznev 2016; Moskowitz & Walker 2020).

5.3 Limitations

There are several additional sources of systematic and measurement uncertainties that must be taken into account when analyzing real data sets.

Our GP model is physics blind; i.e., we have not incorporated any physical principles into our model. While this allows new discoveries to be made, it can also be a limiting factor in certain applications. For instance, inferring the dark matter density distribution can be reconstructed from the velocity distribution of stars using the Jeans equations (Walker & Peñarrubia 2011). One can incorporate physics by adding additional constraints on the relation between the GP curves for different observables. Constructing such a physical model is application dependent and is beyond the scope of this work.

Truncated data, where the selection of the data set depends on one of the dependent variables or is correlated with the dependent variables, and the data are consequently an incomplete and biased subset of a larger population (Mantz 2019). While this case can in principle be handled by modeling the selection function or imputing the missing data, we have not explicitly addressed this limitation in this work.

Another key assumption is the fidelity of measurement uncertainties. Our probabilistic model assumes that the measurement errors are Gaussian with zero mean and known variance. Miscalibration of the measurement uncertainties will result in biases in both inferred intrinsic scatter and correlations. When analyzing real data, such systematic biases must be understood and controlled by developing realistic synthetic light-cone simulations for any given survey and applying the same data reduction techniques to these mock observations.

In this work, we assumed that the measurement uncertainties follow a multivariate Gaussian distribution and may vary in magnitude. This ensures that the posterior average profiles have a closed-form solution for a given set of GP hyperparameters computing via MAP and significantly speed-up computation of the posteriors. However, if measurement uncertainties have a non-Gaussian contribution, the posterior distribution must be sampled or estimated via approximation methods, which may slow down the computation.

In computing the average profile posterior, we employed only the diagonal components of the GP posterior and discarded the nondiagonal contributions. In our examples, the uncertainties on the average profiles are significantly smaller than the intrinsic scatter and the measurement errors of the population; thus, such simplification is justified. However, this simplifying assumption would break down when the sample size is small, and the uncertainties on the average profiles are comparable to the intrinsic scatter. In such limit, it would be important to take the full covariance matrix into account.
6. Summary

In this work, we present a population-based approach to model the average and population variance of spatial distribution of a set of observables from low-S/N measurements. The first step infers the average profile given a set of independent variables and the second step models the covariance of the profile observables. Computations consist of (1) deriving the average relation using Gaussian Process (Section 2.2) and (2) computing the covariance matrix (Section 2.3) given a set of independent variables. We illustrate that the true model can be recovered even in the low-S/N limit, if enough measurements are provided. The effectiveness of our algorithm is illustrated in a controlled simulation setting as well as cosmological simulations of galaxy formation. All of the codes used to produce results in this work are publicly available in a GitHub repository (https://github.com/afarahi/PoPE). Our new population-based method should be useful for modeling a variety of astronomical data.

The authors thank the IllustrisTNG team for making their data and catalogs publicly available and David Barnes for sharing the profile data used in this work. The authors also acknowledge August Evrard for useful comments on the early draft. This work is supported by a Michigan Institute for Data Science (MIDAS) Fellowship and Yale University.

Software: PyMC3 (Salvatier et al. 2016), Matplotlib (Hunter 2007), KLLR.

ORCID iDs

Arya Farahi © https://orcid.org/0000-0003-0777-4618
Daisuke Nagai © https://orcid.org/0000-0002-6766-5942
Yang Chen © https://orcid.org/0000-0002-9516-8134

References

Ágústsson, I., & Brainerd, T. G. 2018, ApJ, 862, 169
Almosallam, I. A., Lindsey, S. N., Jarvis, M. J., & Roberts, S. J. 2016, MNRAS, 455, 2387
Alvarez, I., Niemi, J., & Simpson, M. 2014, arXiv:1408.4050
Arzner, K., Güdel, M., Briggs, K., et al. 2007, A&A, 468, 501
Barnard, J., McCulloch, R., & Meng, X.-L. 2000, Statistica Sinica, 10, 1281
Barnes, D. J., Vogelsberger, M., Kannan, R., et al. 2018, MNRAS, 481, 1809
Barnard, J., McCulloch, R., & Meng, X.-L. 2000, Statistica Sinica, 10, 1281
Brewer, B. J., & Stello, D. 2009, MNRAS, 395, 2226
Bullock, J. S., & Boylan-Kolchin, M. 2017, ARA&A, 55, 343
Carpenter, B., Gelman, A., Hoffman, M. D., et al. 2017, Journal of Statistical Software, 76, 1
Choi, T. 2005, PhD Thesis, Carnegie Mellon University
Cleveland, W. S. 1979, J. Am. Stat. Assoc., 74, 829
Conselice, C. J. 2014, ARA&A, 52, 291
Cui, W., & Zhang, Y. 2017, in Trends in Modern Cosmology, ed. A. J. Capistrano de Souza (London: InTech), 137
de Blok, W. J. G. 2010, AdAst, 2010, 789293
Dietrich, J. P., Zhang, Y., Song, J., et al. 2014, MNRAS, 443, 1713
Farahi, A., Evrard, A. E., McCarthy, I., Barnes, D. J., & Kay, S. T. 2018, MNRAS, 478, 2618
Fletcher, R. 2013, Practical Methods of Optimization (New York: Wiley)
Foreman-Mackey, D., Agol, E., Ambikasaran, S., & Angus, R. 2017, AJ, 154, 220
Gao, L., De Lucia, G., White, S. D. M., & Jenkins, A. 2004, MNRAS, 352, L1
Gelman, A., Lee, D., & Guo, J. 2015, Journal of Educational and Behavioral Statistics, 40, 530
Genton, M. G. 2001, JMLR, 2, 299
Gnedin, O. Y., Kravtsov, A. V., Klypin, A. A., & Nagai, D. 2004, ApJ, 616, 16
Greco, J. P., Hill, J. C., Spergel, D. N., & Battaglia, N. 2015, ApJ, 808, 151
Hoffman, M. D., & Gelman, A. 2014, JMLR, 15, 1593, https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf
Huang, S., Leauthaud, A., Hearin, A., et al. 2020, MNRAS, 492, 3685
Hunter, J. D. 2007, CSE, 9, 90
Jewell, J., Levin, S., & Anderson, C. H. 2004, ApJ, 609, 1
Kelly, B. C. 2007, ApJ, 665, 1409
Kelly, B. C., Sobolewska, M., & Sniemirowska, A. 2011, ApJ, 730, 52
Kipping, D. M. 2010, MNRAS, 408, 1758
Lee, S. K., Lisanti, M., Safdi, B. R., Slattery, T. R., & Xue, W. 2016, PhRvL, 116, 051103
Lewandowski, D., Kurowicka, D., & Joe, H. 2009, Journal of Multivariate Analysis, 100, 199
Li, Q., Cui, W., Yang, X., et al. 2020, MNRAS, 495, 2930
Mantz, A. B. 2019, MNRAS, 485, 4863
McClintock, T., & Rozo, E. 2019, MNRAS, 489, 4155
McClintock, T., Varga, T. N., Gruen, D., et al. 2019, MNRAS, 482, 1352
Melia, F., & Yennapureddy, M. K. 2018, JCAP, 2018, 034
Morris, D., & Tremblay, B. 1994, AJ, 108, 514
Munoz, M., Civan, F., Fabbiano, G., Miyaji, T., & Marchesi, S. 2016, ApJ, 817, 20
Moskovitz, A. G., & Walker, M. G. 2020, ApJ, 892, 27
Nagai, D., & Kravtsov, A. V. 2005, ApJ, 618, 557
Nagai, D., Kravtsov, A. V., & Vikhlinin, A. 2007, ApJ, 668, 1
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 403
Nelson, D., Pillepich, A., Springel, V., et al. 2018, MNRAS, 475, 624
Nelson, D., Springel, V., Pillepich, A., et al. 2019, ComAC, 6, 2
Nilakshi, S. R., Pandey, A. K., & Mohan, V. 2002, A&A, 383, 153
Okabe, N., Oguri, M., Akamatsu, H., et al. 2019, PASJ, 71, 79
Patil, A., Haard, D., & Fosnedge, C. J. 2010, Journal of Statistical Software, 35, 1
Pillepich, A., Springel, V., Nelson, D., et al. 2018, MNRAS, 473, 4077
Piscicchere, J. A., Berlind, A. A., McBride, C. K., & Scoccimarro, R. 2015, ApJ, 806, 125
Quinóñero-Candela, J., & Rasmussen, C. E. 2005, JMLR, 6, 1939
Roussweu, J. 2016, AnRKA, 3, 211
Salcedo, A. N., Wilking, B. D., Weinberg, D. H., et al. 2020, MNRAS, 491, 3061
Salvatier, J., Wiecki, T. V., & Fosnedge, C. 2016, PeerJ Computer Science, 2, e55
Schaller, M., Frenk, C. S., Bower, R. G., et al. 2015, MNRAS, 451, 1247
Schneider, A., & Teyssier, R. 2015, JCAP, 2015, 049
Schneider, A., Teyssier, R., Stadel, J., et al. 2019, JCAP, 2019, 020
Seleznev, A. G., & Walker, M. G. 2020, ApJ, 892, 27
Sales, L. V., Navarro, J. F., Lambas, D. G., White, S. D. M., & Croton, D. J. 2007, MNRAS, 382, 1901
Salvatier, J., Wiecki, T. V., & Fosnedge, C. 2016, PeerJ Computer Science, 2, e55
Schnitzer, P. S., & Weinberg, D. H. 2009, MNRAS, 399, 1774
Walker, M. G., & Peñarrubia, J. 2011, ApJ, 742, 20
Way, M. J., Foster, L. V., Gaziz, P. R., & Srivastava, A. N. 2009, ApJ, 706, 623
Williams, C. K., & Rasmussen, C. E. 2006, Gaussian Processes for Machine Learning, Vol. 2 (Cambridge, MA: MIT Press)