ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #35086

Distance Measure Concepts for Bayesian Inference of Chaotic Dynamical System Parameters

Mr. Colin Michael Burdine, Baylor University (Student)

Colin Burdine is a fourth-year undergraduate student at Baylor University, majoring in Mathematics and Computer Science. He is working on his B.S. in Computing with plans to attend graduate school upon graduating in May 2021. He is interested in computer science theory, with applications in randomized algorithms and automated theorem proving.

Session 35086

Distance Measure Concepts for Bayesian Inference of Chaotic Dynamical System Parameters

Colin Burdine
colin_burdine1@baylor.edu

Abstract

The problem of determining the parameters of dynamical systems often reduces to developing some notion of a "distance" between observed and simulated system trajectory data. The best parameter fit can then be found by adjusting the parameters that generate the simulated trajectory until the distance is minimized. However, in the case of chaotic dynamical systems, traditional distance measures such as the *Mean Square Error* (MSE) often fail to produce good results, a consequence of these systems' inherent sensitivity to changes in parameters and initial conditions. In this paper, we adopt the perspective that more robust distance concepts can be formulated when the trajectories of these chaotic systems are treated as samples from *probability distributions*, rather than as time series data. Within this framework, we evaluate the efficacy of three candidate distance measure concepts and their parameter likelihood functions: The *correlation integral likelihood* (proposed by *Haario et al.*), the *Wasserstein metric*, and finally a novel family of information-theoretic metrics based on the *Kullback-Leibler divergence*. We give particular emphasis to the performance of these methods on the Lorenz63 system, a canonical chaotic system with applications in modeling atmospheric convection.

1 INTRODUCTION

In order to reliably estimate the parameters of dynamical systems that produce a given system trajectory, one must first determine a robust measure of "similarity" or "dissimilarity" between the data generated by any two systems. With such a distance measure, the parameters of a model that best fit a set of observed data can be inferred by minimizing the distance between the observed data and the data generated by simulating that model. However, in the case of chaotic systems, traditional distance measures such as the mean square error fail to provide reliable parameter estimations, especially when the initial conditions of the model are unknown [2, 1].

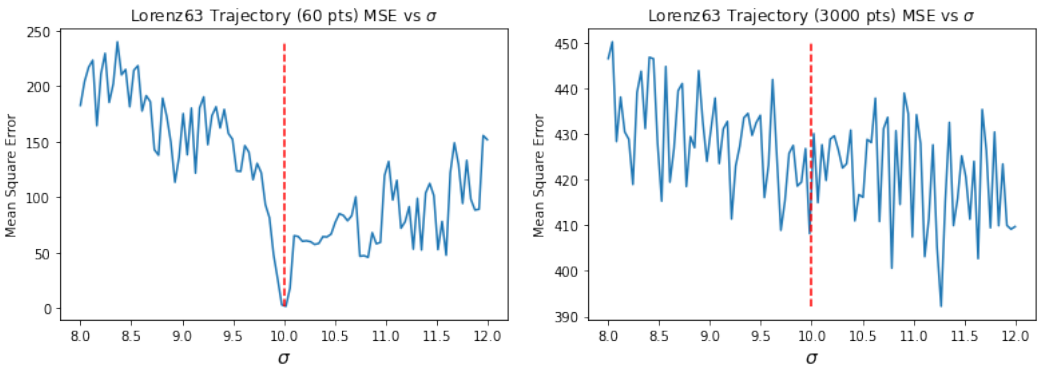


Figure 1: An example of the instability of the mean square error (MSE) metric as the time interval of a chaotic system trajectory increases. The MSE of a 60-point trajectory (left) is compared to a 3000-point trajectory (right)

For well-behaved dynamical systems, it has become standard procedure to use some notion of a residual between a time series data set and a time series model's response to variation in parameters. If one can quantify

the error statistics of the data set, then one can formulate a likelihood function and generate samples from that likelihood using Markov Chain Monte Carlo (MCMC) or Approximate Bayesian Computation (ABC) methods to estimate statistics such as mean, mode, and variance. These methods are often effective, so long as one can show that the trajectories produced by the model are insensitive to slight variations in parameters and initial conditions. However, chaotic dynamical systems are very sensitive to changes in parameters and initial conditions [2]. As these systems evolve, the trajectories produced by similar parameters and initial conditions often diverge from one another, making time-based inference methods ineffective.

In this paper, we disregard the time axis altogether and instead examine the time-invariant attracting sets of these systems and how their shape varies with respect to their parameters. In the case where the attracting sets are isolated points, say a set $\mathbf{X} = \{x_0, x_1, ..., x_n\}$, one can compare the limit points of simulated trajectories with those of a data set using a standard Euclidean distance metric, i.e: $d(\mathbf{X}, \mathbf{X}') = \sum_{i=1}^{N} \|x_i - x_i'\|$. However, not all attracting sets of chaotic systems can be described as the union of discrete mass points; rather, there exist many well-studied examples of multidimensional attractors. One canonical example is the Lorenz63 system, given by (1), with nominal parameters $(\sigma, \rho, \beta) = (10, 28, 8/3)$.

$$\frac{dx_1}{dt} = \sigma(x_2 - x_1), \qquad \frac{dx_2}{dt} = x_1(\rho - x_3) - x_2, \qquad \frac{dx_3}{dt} = x_1 x_2 - \beta x_3, \tag{1}$$

Under initial conditions perturbed slightly from the origin, we observe that integrating this system produces an attractor with two disk-shaped "wings" as seen in Figure 2.

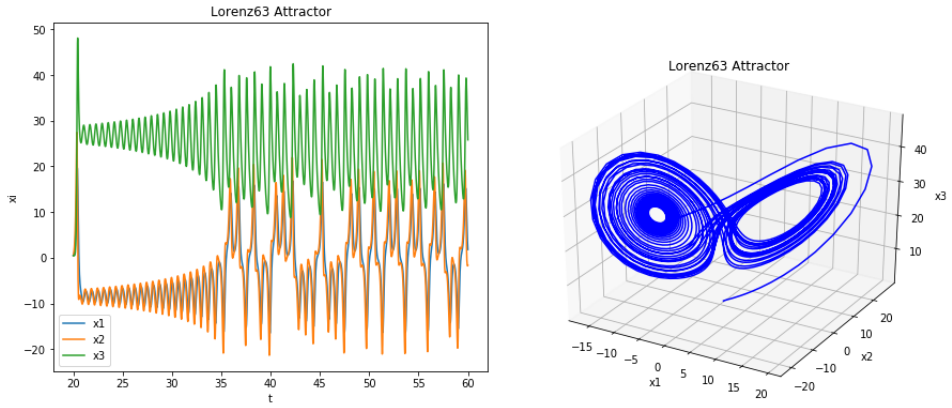


Figure 2: A flattened view (left) and 3D view (right) of the Lorenz63 attractor structure.

Because the trajectory orbits along the surface of one of the two wings seemingly at random, we can see that any attempt to use a time-dependent notion of "distance" between any two isochronous trajectories is hopeless. This provides the motivation for interpreting the state space data as samples from a time-independent probability distribution rather than as time series data. In this paper we will analyze the efficacy three candidate distance measures that use this time-independent approach. The first we will examine is one proposed by Haario et al, which assumes that attractors produced by similar sets of parameters exhibit similar Gaussian behavior in the correlation integral domain [2]. We will also examine a method that approximates the Wasserstein distance metric between distributions. Finally, we will propose new family of metrics based on the Kullback-Liebler divergence between distributions.

2 METHODS

2.1 The Correlation Integral Likelihood

The first method we investigate employs some fractal dimension concepts in statistical physics, as proposed by Haario et al. [2]. In particular, it uses the correlation integral given by (2), which is used to estimate the

dimension of some manifold embedded in a high-dimensional space.

$$C(r, \mathbf{X}) = \frac{1}{|\mathbf{X}|(|\mathbf{X}|-1)} \left[\sum_{x_i, x_j \in \mathbf{X};\ x_i \neq x_j} \#\left(\|x_i - x_j\| < r\right) \right] \tag{2}$$

For example, if one were to calculate the correlation integral about a sufficiently large data set $\mathbf{X}$ that is uniformly distributed within the unit d-hypercube (i.e. $[-1,1]^d \subset \mathbb{R}^d$), one would expect that $\lim_{r\to 0} [\log(C(r,\mathbf{X}))/\log(r)] = d$. Likewise, if $\mathbf{X}$ is sufficiently large and distributed uniformly on the surface of the unit d-hypersphere in $\mathbb{R}^d$, one would expect the limit of the same log ratio to be $d-1$. Intuitively, we can interpret the correlation integral as an exponential function of the *global dimensionality* of the data set at a given scale. In order to extract a meaningful measure of "distance" between two datasets, Haario et al. generalize the calculation of (2) into (3):

$$C(r, \mathbf{X}, \mathbf{X}') = \frac{1}{|\mathbf{X}||\mathbf{X}'|} \left[\sum_{(x_i, x_j) \in (\mathbf{X} \times \mathbf{X}')} \#(\|x_i - x_j\| < r) \right] \tag{3}$$

With this pairwise definition of the correlation integral, we can assume that for some nominal data set $\mathbf{X}$ and a candidate data set $\mathbf{X}'$, $C(r,\mathbf{X}) \approx C(r,\mathbf{X},\mathbf{X}')$ for every $r > 0$ if the parameters that produced $\mathbf{X}'$ are a good fit for $\mathbf{X}$. We note, however, that this definition produces a slightly different result than the standard correlation integral applied to the union of the two datasets (i.e. $C(r, \mathbf{X} \cup \mathbf{X}')$). This is not only computationally cheaper, but also produces a lower correlation integral value for data frmtwo manifolds that have the same dimension, yet are orthogonal to each other. Haario et al. also argues that because the correlation integral is bounded (i.e. $0 \leq C(r,\mathbf{X},\mathbf{X}') \leq 1$) the correlation integral of independently observed attractor datasets is asymptotically normal for every value of r under the Central Limit Theorem. More generally, because C is dependent on the radius at which it is evaluated, we can construct a *correlation integral vector* given by (4), where we sample the correlation integral at several different radii:

$$\mathbf{c_r}(\mathbf{X}) = \left[C(r_1,\mathbf{X}) \ \ C(r_2,\mathbf{X}) \ \ ... \ \ C(r_n,\mathbf{X})\right]^T, \text{ where } \mathbf{r} = \left[r_1 \ \ r_2 \ \ ... \ \ r_n\right]^T \tag{4}$$

By the Central Limit Theorem, the distribution of this vector must converge to a multivariate normal. Supposing that the size of the datasets $\mathbf{X}$ and $\mathbf{X}'$ are sufficiently large such that $\mathbf{c_r}(\mathbf{X}) \sim \text{Norm}(\mu_{\mathbf{c_r}}, \mathbf{\Sigma_{c_r}})$, we can construct an empirical likelihood function, given by (5) and (6):

$$\mathcal{L}(\mathbf{X}'|\mathbf{X},\mathbf{r}) \propto \exp\left[-\frac{1}{2}(\mu_{\mathbf{c_r}} - \widetilde{\mathbf{c}}_{\mathbf{r}}(\mathbf{X},\mathbf{X}'))^T {\mathbf{\Sigma_{c_r}}}^{-1} (\mu_{\mathbf{c_r}} - \widetilde{\mathbf{c}}_{\mathbf{r}}(\mathbf{X},\mathbf{X}'))\right], \tag{5}$$

$$\text{where} \quad \widetilde{\mathbf{c}}_{\mathbf{r}}(\mathbf{X},\mathbf{X}') = \left[C(r_1,\mathbf{X},\mathbf{X}') \ \ C(r_2,\mathbf{X},\mathbf{X}') \ \ ... \ \ C(r_n,\mathbf{X},\mathbf{X}')\right]^T \tag{6}$$

In (5), Haario et al. estimate the quantities $\mu_{\mathbf{c_r}}$ and $\mathbf{\Sigma_{c_r}}$, as the mean and covariance respectively of the set of correlation integral vectors $\{\mathbf{c_r}(\mathbf{X}^*_{(\mathbf{i_1})}, \mathbf{X}^*_{(\mathbf{i_2})}) \mid i \in 1,2,...,N\}$, in which each trajectory pair ($\mathbf{X}^*_{(\mathbf{i_1})}$ and $\mathbf{X}^*_{(\mathbf{i_2})}$) is generated with the estimated best fit parameters for the nominal data set $\mathbf{X}$, but with initial conditions selected at random, ideally from a distribution that covers known canonical parameters. For example, in the Lorenz63 system the initial conditions (x_1, x_2, x_3) were selected from the distribution $\text{Norm}(\mu = (0,0,0), \sigma = 0.5)$, excluding the degenerate point $(0,0,0)$.

We remark that although this method makes relatively few assumptions about the data, it does require a careful choice of the sampling radii $\mathbf{r}$ for the correlation integral calculation. If the elements of $\mathbf{r}$ are too high, then the features of the attractor dataset at smaller scales will not be captured; likewise, if they are too low, then the large scale features will not be captured. We recommend the heuristic (7) for a nominal data set $\mathbf{X}$ and a desired number of correlation samples N:

$$\hat{\mathbf{r}}_i = (r_{min})^{1-i/N} (r_{max})^{i/N}, \text{ where } r_{min} = \min_{x_j \neq x_k \in \mathbf{X}} \|x_j - x_k\| \text{ and } r_{max} = \max_{x_j \neq x_k \in \mathbf{X}} \|x_j - x_k\| \tag{7}$$

2.2 The Wasserstein Distance Likelihood

The Wasserstein metric, or "Earth Mover's Distance" (EMD) has been shown to be a robust metric that provides an intuitive notion of distance between probability distributions. Physically it can be interpreted as the minimal amount of "work" required to transform one "pile of dirt" into another, where "work" is defined as an amount of mass moved times the distance it is moved. Suppose we have two piles of dirt given by the functions $p(x)$ and $q(x)$ defined on some set $\mathcal{X}$, where $\int_{\mathcal{X}} p(x)\ dx = \int_{\mathcal{X}} q(x)\ dx > 0$, then we could devise any number of strategies by which we could transform the mass under the curve of $p(x)$ into that of $q(x)$. Let $\Gamma(p,q)$ denote the space of all such strategies. We define each strategy $\gamma \in \Gamma(p,q)$ as a function $\gamma : (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}_{\geq 0}$ such that:

$$q(s) = \int_{\mathcal{X}} \gamma(x,s)\ dx, \quad \text{and conversely:} \quad p(s) = \int_{\mathcal{X}} \gamma(s,x)\ dx.$$

We can interpret the quantity $\gamma(x_1, x_2)$ as the amount of mass moved from x_1 to x_2 under the strategy γ. From this formulation of the function space Γ, the Wasserstein metric arises naturally under the Euclidean distance metric as (8):

$$W(P,Q) = \inf_{\gamma \in \Gamma(p,q)} \iint_{(\mathcal{X}\times\mathcal{X})} \gamma(x_1,x_2)\|x_1 - x_2\|\ dx_1 dx_2 \tag{8}$$

If p and q are both probability distributions, then by definition they must comprise the marginals of every $\gamma \in \Gamma(p,q)$. Thus, we can reinterpret the optimal transport formulation of the Wasserstein distance as the minimal expected distance between two random variables in (9).

$$W(P,Q) = \inf_{\gamma \in \Gamma(p,q)} \mathbb{E}_{\gamma}\left[\ \|P - Q\|\ \right] \tag{9}$$

Unfortunately, this distance metric does not lend itself to an intuitive likelihood function, so we resort to embedding it in a Gaussian kernel to produce an *empirical likelihood function* given by (10).

$$\mathcal{L}(P|Q) \propto \exp\left[\frac{-(W(P,Q))^2}{\alpha}\right] \tag{10}$$

Due to the Central Limit Theorem, we claim that $W(P,Q)$ is asymptotically normal, so long as the random variable P is produced with initial conditions and parameters similar to that of Q. This justifies the use of the Gaussian kernel. We also observe that if $\mathcal{L}(P|Q)$ is normally distributed, then the quantity $2W(P,Q)^2/\alpha$ must follow a chi-squared distribution with an unknown degree of freedom ν. In (11), we reformulate this relationship in terms of just $(W(P,Q))^2$ and see that we obtain the form of a gamma distribution with scale parameter $\theta = \alpha$:

$$(W(P,Q))^2 \quad \sim \quad \frac{\alpha}{2}\mathcal{X}^2_{(\nu)} \quad \equiv \quad \Gamma\left(\kappa = \frac{\nu}{2}, \theta = \alpha\right) \tag{11}$$

In order to estimate the value of α, we use a similar approach as we used to estimate the values of $\mu_{\mathbf{c_r}}$ and $\mathbf{\Sigma_{c_r}}$ in (5). First, we construct a set of trajectories $\{\mathbf{X}^*_{(Qi)} \mid i \in 1,2,...,N\}$, which are generated from the set of parameters that best fit the distribution of Q, but with randomized initial conditions. We then evaluate the Wasserstein distance of these trajectories from our nominal attractor distribution Q. This gives us the set $\mathcal{D}_Q = \{W(\mathbf{X}_{(Qi)}, Q)^2 \mid i \in 1,2,...,N\}$, which we fit to the gamma distribution in (11). Because we only need to determine the scale parameter, however, we can estimate α in terms of the closed form consistent unbiased estimator for θ given by (12):

$$\widetilde{\alpha} = \widetilde{\theta}, \quad \text{where } \widetilde{\theta} = \frac{1}{(N-1)}\left[\sum_{x \in \mathcal{D}_Q} \ln(x)(x - \mu_{\mathcal{D}_Q})\right], \quad \mu_{\mathcal{D}_Q} = \sum_{x \in \mathcal{D}_Q} \frac{x}{N} \tag{12}$$

2.3 KL Divergence Likelihoods

The Kullback-Liebler (KL) divergence is a measure of difference between probability distributions that derives from the *loss in information* between distributions. Suppose we have a nominal random variable P and a candidate random variable Q that is intended to approximate P. The natural log KL divergence of P from Q is given by (13):

$$D_{KL}(P|Q) = \int_{\text{supp}(P \cap Q)} p(x) \ln\left(\frac{p(x)}{q(x)}\right)\ dx \tag{13}$$

From an information theory perspective, we can interpret D_{KL} as measuring *the expected loss of self-information when P is used as a surrogate for Q* (14). Using the notation of Shannon's entropy and cross entropy function (H and H_{cross} respectively), we can also interpret D_{KL} as *the difference of entropy in P and cross entropy of P given Q* (15). Both of these equivalent interpretations of D_{KL} are shown below:

$$\begin{aligned} D_{KL}(P|Q) &= -\int_{\text{supp}(P \cap Q)} p(x) \ln(q(x))\ dx + \int_{\text{supp}(P \cap Q)} p(x) \ln(p(x))\ dx \\ &= \mathbb{E}_P[I(Q)] - \mathbb{E}_P[I(P)] \qquad (14) \\ &= H_{\text{cross}}(P|Q) - H(P) \qquad (15) \end{aligned}$$

D_{KL} also has the desirable property of additivity; we can evaluate it separately on disjoint subregions of the support of its operands and sum the results to obtain the correct divergence. However, D_{KL} only attains finite nonzero values on the *intersection* of the supports of P and Q. If we were to extend the domain of integration in (13) to the *union* of the supports of P and Q, we would obtain a divergence of 0 over the subregions wherever $p(x) = 0$ and an infinite divergence wherever $q(x) = 0$. Furthermore, D_{KL} is not a symmetric quantity and thus fails to conform to an intuitive notion of "distance". We can force D_{KL} to be symmetric as given in (16), yet this still only attains meaningful values on the intersection of the supports of its operands:

$$D_{symKL}(P, Q) = \frac{1}{2} D_{KL}(P|Q) + \frac{1}{2} D_{KL}(Q|P) \tag{16}$$

One way we can resolve the support issue is to calculate the divergence of both distributions from their *mean distribution* M, which is supported over the union of both distributions. This quantity is referred to as the Jensen-Shannon Divergence and is given in (17).

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}(P|M) + \frac{1}{2} D_{KL}(Q|M), \quad \text{where} \quad m(x) = \frac{p(x) + q(x)}{2} \tag{17}$$

Of the three divergence measures we have discussed ($D_{KL}, D_{symKL},$ and $, D_{JS}$), we give preference to the use of the JS divergence due to the fact that it is additive, symmetric, and attains meaningful values over the union of the support of its operands. However, one additional property of D_{JS} that we must keep in mind is that it is *bounded*. Since $D_{KL} \geq 0$, a consequence of Gibb's inequality, we observe that $D_{JS} \geq 0$ also. Using Gibb's inequality again as well as the Shannon entropy bound ($0 \leq H(X) \leq \ln(2)$), we obtain the upper bound on D_{JS} as shown by (18):

$$\begin{aligned} D_{JS}(P, Q) &= \frac{1}{2} \int_{\text{supp}(P \cup Q)} p(x) \ln\left(\frac{p(x)}{m(x)}\right) + q(x) \ln\left(\frac{q(x)}{m(x)}\right)\ dx \\ &= \frac{1}{2} \int_{\text{supp}(P \cup Q)} p(x) \ln(p(x)) + q(x) \ln(q(x)) - 2m(x) \ln(m(x))\ dx \\ &= \frac{1}{2}(H(P) + H(Q) - 2H(M)) \\ &\leq \frac{1}{2}(\ln(2) + \ln(2) + 0) \\ &= \quad \ln(2) \qquad (18) \end{aligned}$$

Because $0 \leq D_{JS} \leq \ln(2)$, constructing an intuitive likelihood function poses some challenges. Foremost, we would like $\mathcal{L}(P|Q)$ to approach 0 as $D_{JS} \to \ln(2)$. We would also like to be able to scale $\mathcal{L}$ so as to prevent the distribution becoming degenerate or uniform as the size of the parameter space increases or decreases respectively. To satisfy these two conditions, we propose the JS divergence empirical likelihood given in (19):

$$\mathcal{L}(P|Q) \propto \left[1 - \frac{D_{JS}(P, Q)}{\ln(2)}\right]^{1/\alpha} \tag{19}$$

In order to select the value of α, we follow the same method as we did in selecting α for (10), observing that we can re-write (19) in the form of a Gaussian kernel:

$$\mathcal{L}(P|Q) \propto \exp\left(-\frac{1}{2}\left[-\frac{2}{\alpha}\ln\left(1-\frac{D_{JS}(P,Q)}{\ln(2)}\right)\right]\right) \tag{20}$$

In a manner analogous to (11), we observe that we can fit $-\ln(1-D_{JS}/\ln(2))$ to a gamma distribution:

$$-\ln\left(1-\frac{D_{JS}(P,Q)}{\ln(2)}\right) \quad \sim \quad \frac{\alpha}{2}\mathcal{X}^2_{(\nu)} \quad \equiv \quad \Gamma\left(\kappa=\frac{\nu}{2},\theta=\alpha\right) \tag{21}$$

Like we did in the Wasserstein likelihood, we can estimate α by constructing a set of trajectories and corresponding distances $\mathcal{D}_Q = \{D_{JS}(X_{(Qi)},Q)|i \in 1,2,...,N\}$, which we fit to the gamma distribution in (21). Using the closed form unbiased estimator $\widetilde{\theta}$ given in (12), we get the same estimator, $\widetilde{\alpha} = \widetilde{\theta}$.

3 Experiments and Results

3.1 Correlation Integral Likelihood

We recall that the Correlation integral method proposed by Haario et al. employs the idea that samples from the same attractor probability distribution exhibit similar statistics in the correlation domain. We assume that the correlation integral vector of an attractor is distributed (asymptotically) according to a Gaussian distribution given by (5), whose mean and covariance are selected empirically according to the observed distribution of the correlation integral vector about the parameters that provide the best fit for the model. If these parameters are known and we only want to produce an estimate of the uncertainty of the parameters, then we simply follow the procedure outlined above. However, if we cannot even supply an educated guess of the best fit parameters, we must use another method that provides some guarantee of of a global minimum at the optimal fit. Because this method provides a measure of *likelihood* about a proposed model fit, not of a *distance* from it, we caution the reader against using this method to "bootstrap" the way to an optimal fitting.

Assuming that we can produce a set of parameters that optimally fit the data, we remark that the Haario method makes few additional assumptions about the data in the process of formulating the parameter space likelihood function. Being a *data-driven* approach, the likelihood function is calculated in terms of the data itself, not in terms of its underlying probability distribution. To better understand the efficacy of this approach, we performed several experiments that show how the likelihood given in (5) varies with the parameters of chaotic systems. In this method as well as the methods below, we focus on the Lorenz63 system (1).

The first experiment we conducted examined the shape of the likelihood function, varying only one parameter at a time from the nominal values of Lorenz63. Both the nominal data set and each candidate data set ($\mathbf{X}$ and $\mathbf{X}'$) were produced by integrating the Lorenz63 system from $t = 0$ to $t = 500$ to produce a trajectory. 3000 points were sampled from each trajectory at uniform time intervals from $t = 20$ to $t = 500$ so that the time interval where the trajectory is converging to the attractor is excluded. The correlation integral mean and covariance ($\mu_{\mathbf{c_r}}$ and $\mathbf{\Sigma_{c_r}}$ in (5)) were estimated as the mean and covariance of 600 independently sampled trajectory datasets. These trajectory datasets were produced under the nominal conditions, but with initial conditions selected from a 3D Gaussian distribution with $\mu = (0,0,0)$ and $\sigma = 0.5$. The results are plotted in Figure 3:

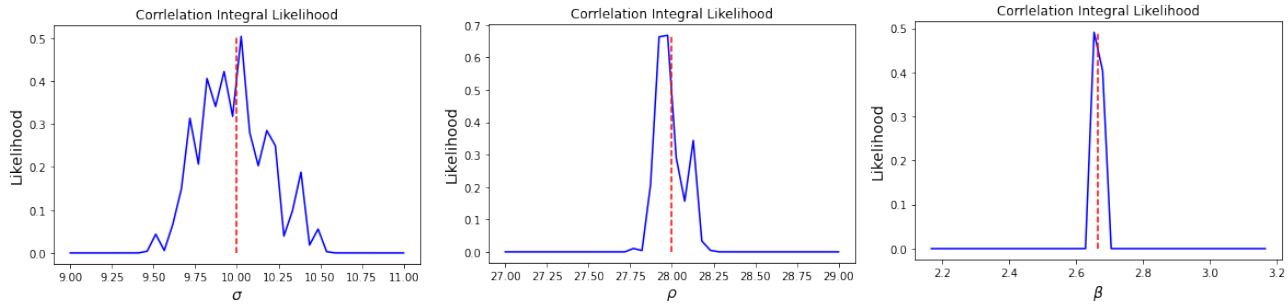


Figure 3: The correlation integral likelihood under variations of each parameter from the nominal values of Lorenz63

Although the likelihood is maximized at the nominal parameters and appears to be unimodal, we can see that it is not entirely smooth. We observe the most noise in the likelihood as we vary the value of σ. In Figure 4, we vary two parameters at once. Although we still see noise in the posterior, the covariances between variables emerges:

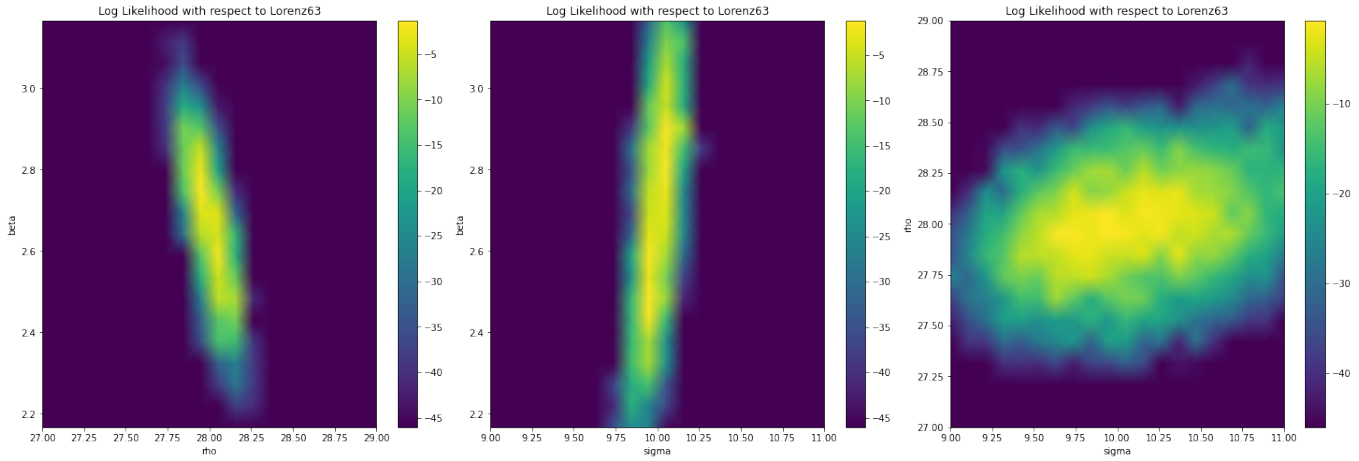


Figure 4: 2D slices of the Lorenz63 correlation integral likelihood function. For each slice, the invariant parameter is fixed at the nominal value.

Finally, we produced samples from this likelihood through Hamiltonian Monte Carlo (HMC) sampling with a uniform prior. Due to the expensiveness of integrating the system to produce each proposal likelihood in the sampling process, we used a 3D linear interpolation surrogate which approximated the value of the correlation integral vector likelihood. To construct this interpolation, the correlation integral likelihood was calculated exactly across a regular 25x25x25 grid centered at the nominal parameters. Values outside this regular grid were treated with likelihood 0 under a uniform prior, which was restricted to the support of the grid. We opted for the linear surrogate because it was computationally inexpensive to fit, and the gradient of the likelihood function could be calculated easily. Also, it allowed us to increase the sampling resolution without needing to recompute the entire surrogate. The marginals of the posterior samples obtained through the sampling procedure are plotted in Figure 5.

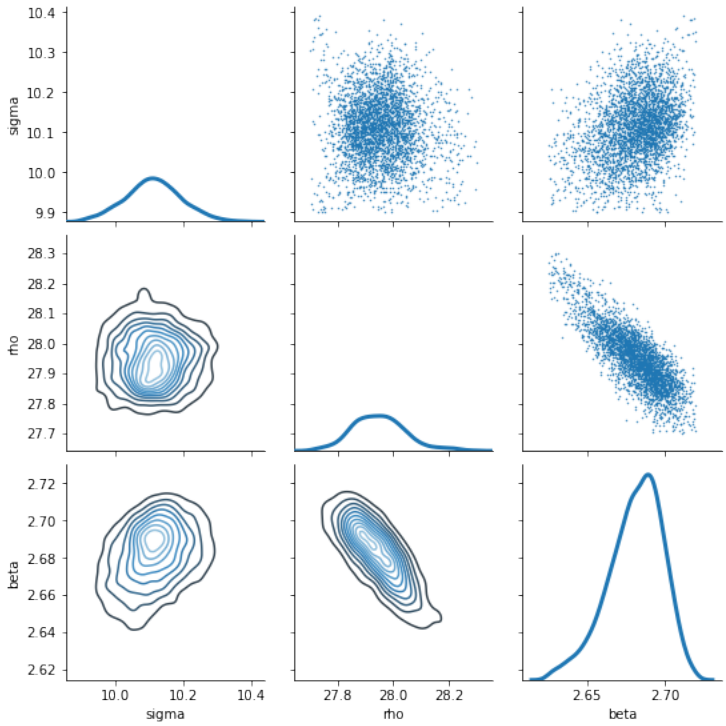


Figure 5: Posterior of Lorenz63 generated from HMC sampling of the Correlation integral likelihood with a uniform prior.

We observe that the posterior is distinctively Gaussian, with the means and modes attained close to the Lorenz63 nominal values $(\sigma, \rho, \beta) = (10, 28, 8/3)$. Of particular note are the covariances, which are positive on the sigma-beta marginal, negative on the rho-beta marginal and slightly negative on the sigma-rho marginal.

3.2 Wasserstein Distance Likelihood Approximation

Recalling the definition of the Wasserstein metric given by (9), we see that attempting to find a strategy function γ with marginals P and Q that minimizes $\mathbb{E}_\gamma[\,\|P-Q\|\,]$ is difficult without making some simplifying assumptions about P and Q. In this paper, we shall assume that the density of both P and Q are continuous and bounded in the state space such that they can be expressed as a uniform limit of the finite union of N disjoint regions of uniform density, as written in (22):

$$P = \lim_{N\to\infty} \sum_{i=1}^{N} (\rho_{\mathcal{C}_N i}) \mathbb{1}_{\{\mathcal{C}_{Ni}\}}, \qquad \text{where:} \quad \text{supp}(P) = \bigsqcup_{i=1}^{N} \mathcal{C}_{Ni}, \quad \sum_{i=1}^{N} \rho_{\mathcal{C}_{Ni}} = 1 \tag{22}$$

While (22) seems to be a straightforward assumption to make, it does not hold under all parameter sets of chaotic systems. In the case of Lorenz63, there exist some parameter configurations that produce limit points and limit cycles, both of which have degenerate distributions in the state space [4]. For the remainder of this paper, we will assume that the parameters being estimated are not degenerate and that (22) holds.

Once we can produce a discretization of both P and Q, we can more easily determine the Wasserstein distance by formulating it as a discrete optimal transport problem. Suppose for some sufficiently large N we decompose the support of P and Q into connected regions $\mathcal{C}_{Pi}$ and $\mathcal{C}_{Q_i}$, with respective means μ_{Pi} and μ_{Qi} and respective densities $\rho_{\mathcal{C}_P i}$ and $\rho_{\mathcal{C}_Q i}$. Let $\mathbf{p}$ and $\mathbf{q}$ be the vectors of all N respective region densities (e.g. $\mathbf{p} = [\,\rho_{\mathcal{C}_P 1}\ \rho_{\mathcal{C}_P 2}\ \ldots\,]^T$). This simplifies the original function space $\Gamma(P, Q)$ to be a set $\mathcal{G}_{P,Q}$ of $N \times N$ matrices with entries on $[0, 1]$, such that for every $\mathbf{G} \in \mathcal{G}_{P,Q}$, the column vectors sum to $\mathbf{p}$ and the row vectors sum to $\mathbf{q}^T$. Finally, let $\mathbf{K}$ represent the kernel distance matrix where $\mathbf{K}_{ij} = \|\mu_{Pi} - \mu_{Qj}\|$. We can now formulate the discrete approximation of

the continuous Wasserstein metric, which we refer to by its more common name as the discrete "Earth Mover's Distance" (EMD) problem given in (23):

$$\text{EMD}(\mathbf{p}, \mathbf{q}; \mathbf{K}) = \min_{\mathbf{G} \in \mathcal{G}_{P,Q}} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{G}_{ij} \mathbf{K}_{ij} \tag{23}$$

This discrete EMD problem can be solved through traditional constrained linear optimization methods, such as *linear programming*, or *constrained gradient descent* [1]. We will not provide a detailed discussion of these different methods here, however we will remark that the linear programming formulation provides an exact solution with an asymptotic run time of $O(N^3)$, while constrained gradient descent provides a close approximation with varying run time [1].

To determine how well the solution to the discrete EMD problem approximates the true Wasserstein distance, we first observe that we can use the triangle inequality to bound the Wasserstein metric from below by the difference in distribution means ($\Delta\mu$) and from above by the discretized EMD plus the sums of the region radii (e.g: $R(\mathcal{C}_{Pi})$), as shown in (24):

$$\begin{aligned} &\Delta\mu \leq W(P,Q) \leq \text{EMD}(\mathbf{p}, \mathbf{q}; \mathbf{K}) + \sum_{i=1}^{N} \left[R(\mathcal{C}_{Pi}) \rho_{\mathcal{C}_{Pi}} + R(\mathcal{C}_{Qi}) \rho_{\mathcal{C}_{Qi}} \right], \\ &\text{where} \quad \Delta\mu = \|\mathbb{E}[P] - \mathbb{E}[Q]\| = \left\| \left(\sum_{i=1}^{N} \mu_{Pi} (\rho_{\mathcal{C}_{Pi}}) \right) - \left(\sum_{i=1}^{N} \mu_{Qi} (\rho_{\mathcal{C}_{Qi}}) \right) \right\| \\ &\text{and } R(\mathcal{C}_{Pi}) = \sup_{\mathbf{x} \in \mathcal{C}_{Pi}} \|\mu_{Pi} - \mathbf{x}\| \end{aligned} \tag{24}$$

From the perspective of a physical transport problem, we can interpret the lower bound of the EMD as being the amount of "work" required to shift the center of mass of P to be that of Q. We can also interpret the upper bound as the discrete EMD plus an upper bound amount of "work" required to convert P and Q into degenerate mass point distributions located at the means of their respective regions. If we let P_δ and Q_δ be the Dirac delta mass point approximations of P and Q respectively, where they discretely attain the mean of each region (e.g. μ_{Pi}) with probability equal to the region's density (e.g. $\rho_{\mathcal{C}_{Pi}}$). Having defined these distributions we can generalize the righthand side of (24) as the pair of inequalities in (25), recalling that the Wasserstein distance is a metric:

$$\begin{aligned} W(P,Q) &\leq W(P,P_\delta) + W(P_\delta, Q_\delta) + W(Q_\delta, Q) \\ \text{and } W(P_\delta, Q_\delta) &\leq W(P_\delta, P) + W(P,Q) + W(Q, Q_\delta) \end{aligned} \tag{25}$$

Since $W(P,P_\delta) + W(Q,Q_\delta) \leq \sum_{i=1}^{N} \left[R(\mathcal{C}_{Pi}) \rho_{\mathcal{C}_{Pi}} + R(\mathcal{C}_{Qi}) \rho_{\mathcal{C}_{Qi}} \right]$, and by definition $W(P_\delta, Q_\delta) = \text{EMD}(\mathbf{p}, \mathbf{q}; \mathbf{K})$, we get (26):

$$|W(P,Q) - \text{EMD}(\mathbf{p}, \mathbf{q}; \mathbf{K})| \leq \sum_{i=1}^{N} \left[R(\mathcal{C}_{Pi}) \rho_{\mathcal{C}_{Pi}} + R(\mathcal{C}_{Qi}) \rho_{\mathcal{C}_{Qi}} \right] \tag{26}$$

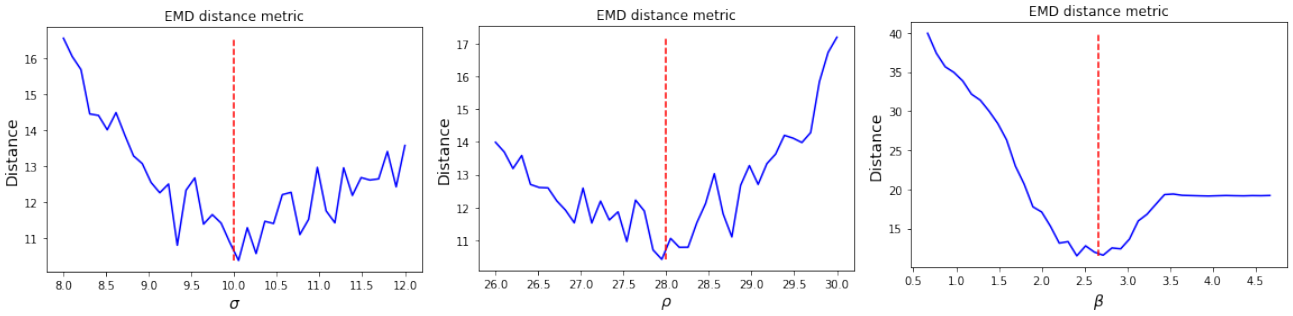
This gives us an approximation error bound of the discrete EMD with respect to the actual Wasserstein distance between P and Q. So long as we can guarantee that the right side of (26) decreases to 0, then our approximation converges to the actual value of the Wasserstein metric. However, in practice we will be working with *samples* from the attractor random variables P and Q instead of the distributions themselves. We will also have to produce the regions $\mathcal{C}_{Pi}$ and $\mathcal{C}_{Qi}$ in such a manner that as $N \to \infty$, the righthand side of (26) vanishes. In order to meet this condition, we must ensure that the radii of the disjoint regions are approximately equal, which we can achieve through *k-means clustering* on the observed trajectories.

Because the k-means algorithm minimizes the mean square distance of each point to its respective cluster mean, it produces clusters with approximately the same maximal radii. Furthermore, when we increase the number of clusters discovered ($k = N$), we decrease the average estimated density of each cluster. Since these two conditions are satisfied through k-means clustering, the righthand side of (26) vanishes and we attain convergence

to the actual Wasserstein distance given sufficiently large attractor samples and a sufficiently large value of N.

We now shift our attention toward employing this approximation as a notion of distance between Lorenz63 attractors. In the first experiment we conducted, we varied one parameter at a time from the nominal values of Lorenz63 and plotted the corresponding EMD metric approximation for $N = 256$ and $N = 1024$ clusters respectively. In order to solve each discrete EMD problem, we used a linear programming formulation to obtain an exact solution. The results of this initial experiment are given in Figure 6:

(a) EMD Metric (256 clusters)



(b) EMD Metric (1024 clusters)

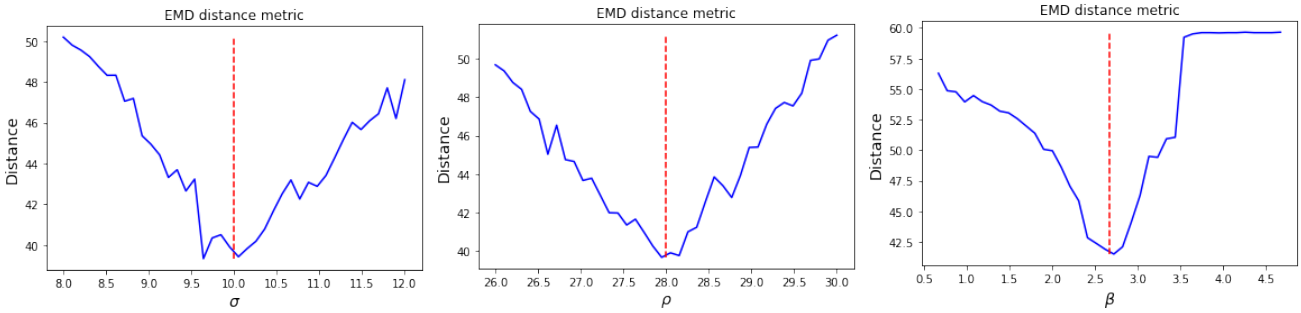


Figure 6: EMD approximation metric under variations of single parameters from the nominal values of Lorenz63.

We can see that there exists a significant amount of noise in the EMD discretization; however, the noise decreases as the number of clusters increases. As we vary two parameters at once from the nominal values, we obtain the likelihood plots shown in Figure 7:

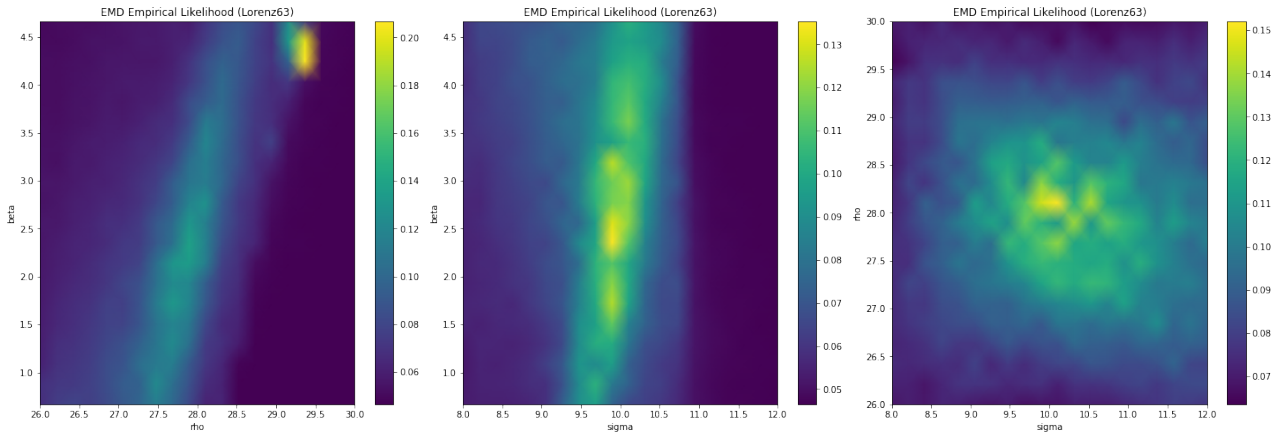
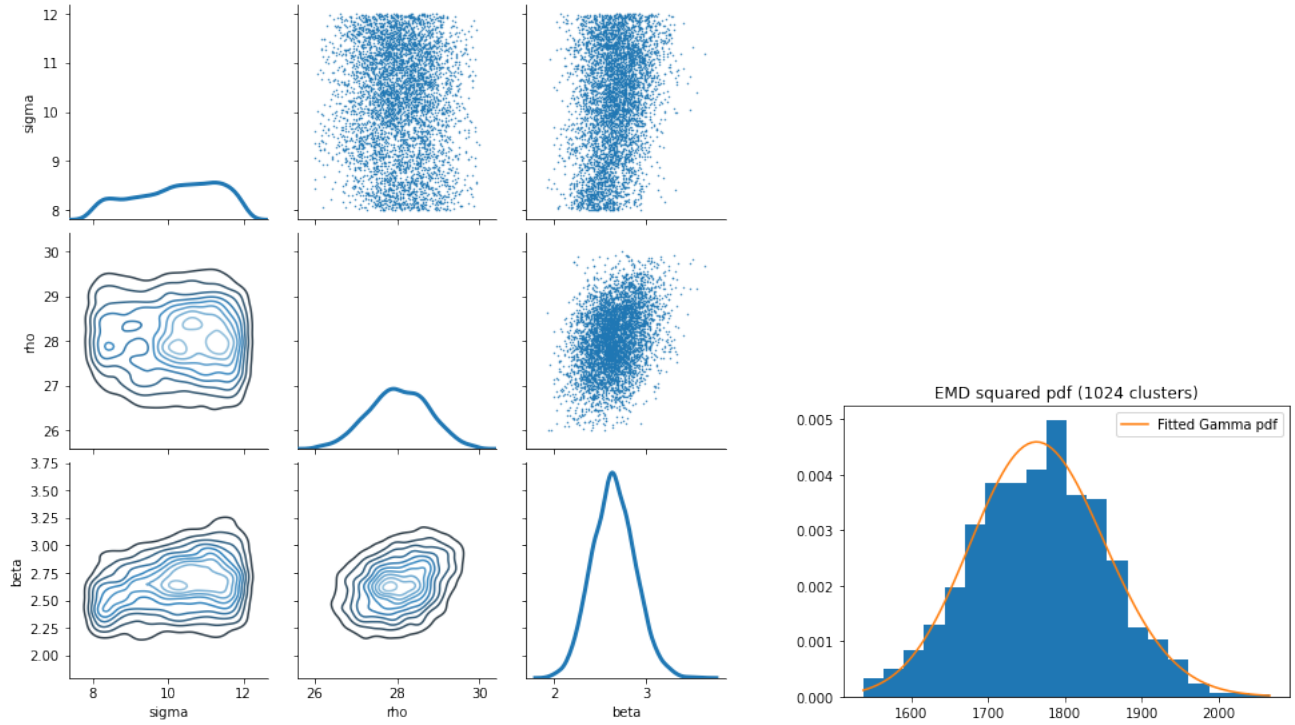


Figure 7: 2D slices of the Lorenz63 Wasserstein likelihood function as given in (10) for fixed $\alpha = 1$. To approximate the actual Wasserstein distance, 1024 clusters were used. In each slice, the invariant parameter is fixed at the nominal value.

Although we observe a distinct correlation among the parameters in the figures above, we note that there is also significant noise the likelihood function, even under the 1024-cluster approximation of the Wasserstein metric. We also observe that the EMD approximation tends to produce small "pockets" of high likelihood in the likelihood function. We can see examples of this close to the bifurcation region where $\beta \approx 4.0$ and $\rho \approx 29.2$.

In order to obtain information about the posterior distribution, we must first estimate the true value of α in (10), which we obtain by producing data sets under the fixed nominal parameters with varying initial conditions. Taking the squares of the observed approximate Wasserstein distances from the nominal data set, we obtain the density plot given in Figure 8b. Using the estimator $\widetilde{\alpha} = \widetilde{\theta}$, where $\widetilde{\theta}$ is given in (12) we obtain our estimate of α:



(a) (b)

Figure 8: Posterior samples of Lorenz63 generated from HMC sampling of the 1024-cluster EMD likelihood with a uniform prior (a) and the gamma distribution used to select the parameter α in the likelihood function (b).

After producing 1000 samples under random initial conditions, we obtained the estimate of $\widetilde{\alpha} = \widetilde{\theta} = 4.425$ for the nominal parameters of Lorenz63. Using this value of α, we produced samples from the likelihood function given in (10) through HMC sampling. We used the same linear likelihood function surrogate as with the correlation integral sampling. In contrast to the correlation integral posterior, we see that the noise in the EMD approximation resulted in a high value of α being selected, which in turn resulted in the high degree of variance observed in the posterior samples in Figure 8a.

3.3 Divergence Likelihood Approximations

We recall that the family of KL divergence distance measures all have the desirable property of additivity. That is, we are able to split up the domain of integration on which the KL divergence in (13) is evaluated. Like the Wasserstein metric, it is difficult to evaluate exactly, so we resort to an approximation that takes advantage of this additivity property. However, we also must be careful about the assumptions that we make in this approximation scheme, as some parameter configurations produce degenerate attractor structures as discussed earlier. Thus, we proceed once again under the assumption that (22) holds. After making this assumption, we can produce a discrete approximation of the KL divergence (D_{KL}) and consequently produce approximations for the family of KL divergence distance measures ($D_{KL}, D_{symKL}, D_{JS}$). Suppose we have two attractor random variables P and Q and the intersection of their support can be decomposed into N disjoint regions, $\mathcal{C}_i$ for $i \in 1, 2, ..., N$. For each region we calculate uniform densities respective to both distributions (ρ_{Pi} and ρ_{Qi}, for $i \in 1, 2, ..., N$) such that they sum to unity. The discretization of the KL divergence is given by (27):

$$D_{KL}(P|Q) \approx \sum_{i=1}^{N} \rho_{Pi} \ln\left(\frac{\rho_{Pi}}{\rho_{Qi}}\right) \tag{27}$$

We observe that because D_{KL} is unbounded it can be sensitive to the selection of the regions C_i, particularly if there are regions of Q with a very small density ρ_{Q_i}. This can be problematic because we are in fact working with *samples* from P and Q ($\mathbf{X}'$ and $\mathbf{X}$ respectively). To remedy this issue, for each region C_i we estimated the density of the attractor distributions using Gaussian *kernel density estimates* (KDEs) with bandwidths selected according to Scott's Rule [3]. For each cluster, we sampled the density of the KDE at points on a regular mesh containing each cluster and computed an average density weighted by the volume of each cluster. Letting $G_{\mathcal{C}_i}$ denote the set of points in each regular mesh, $\rho_{\mathbf{X}i}(\mathbf{g})$ for $\mathbf{g} \in G_{\mathcal{C}_i}$ denote the density at a mesh point $\mathbf{g}$, and $V(G_{\mathcal{C}_i})$ denote the volume of $\mathcal{C}_i$, we can produce a better approximation with (28):

$$D_{KL}(\mathbf{X}'|\mathbf{X}) = \sum_{i=1}^{N} \frac{V(\mathcal{C}_i)}{Z} \left(\sum_{\mathbf{g} \in G_{\mathcal{C}_i}} \rho_{\mathbf{X}'i}(\mathbf{g}) \ln\left(\frac{\rho_{\mathbf{X}'i}(\mathbf{g})}{\rho_{\mathbf{X}i}(\mathbf{g})}\right) \right), \quad \text{where } Z = \sum_{i=1}^{N} V(\mathcal{C}_i) \tag{28}$$

In terms of (28) the JS divergence can be approximated by (29):

$$D_{JS}(\mathbf{X}', \mathbf{X}) = \frac{D_{KL}(\mathbf{X}'|\mathbf{M}) + D_{KL}(\mathbf{X}|\mathbf{M})}{2}, \quad \text{where } \rho_{\mathbf{M}i}(\mathbf{g}) = \frac{\rho_{\mathbf{X}'i}(\mathbf{g}) + \rho_{\mathbf{X}i}(\mathbf{g})}{2} \tag{29}$$

Using this definition of the JS divergence, we performed some simple experiments in which one parameter of the Lorenz63 system was varied at a time. As mentioned above, we gave preference to the use of the JS divergence over the other KL divergence-based metrics, due to its symmetry and nondegeneracy. To approximate the JS divergence, 8 clusters were identified and regular 25x25x25 meshes were used to sample the KDE in each cluster. The results of this experiment are summarized in Figure 9:

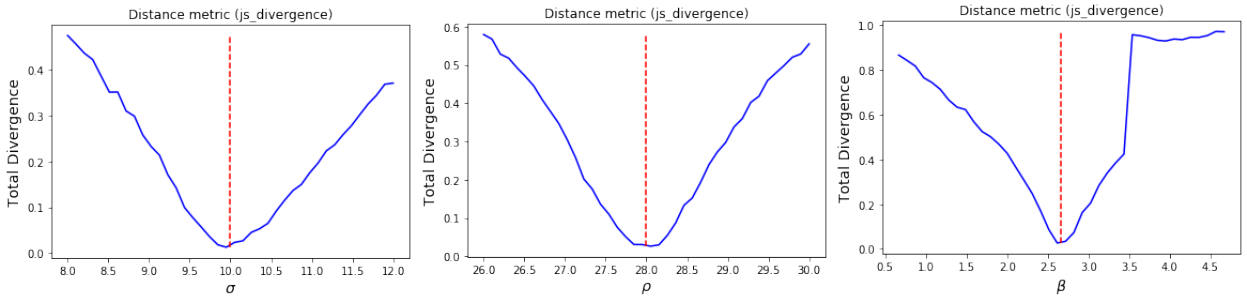


Figure 9: The KDE approximation of the JS divergence under variations of each parameter from the nominal values of Lorenz63.

We note that the JS divergence provides a smooth distance metric that decreases nearly monitonically as the parameters approach the nominal values. In particular, we note that the divergence maintains its monotonicity even at the point $\beta = 3.5$ where the Lorenz63 attractor bifurcates. We see the same smoothness when we embed the JS divergence approximation given by (29) into the empirical likelihood function (19) for a fixed scale parameter α. Some 2D slices of this likelihood function are shown in Figure 9:

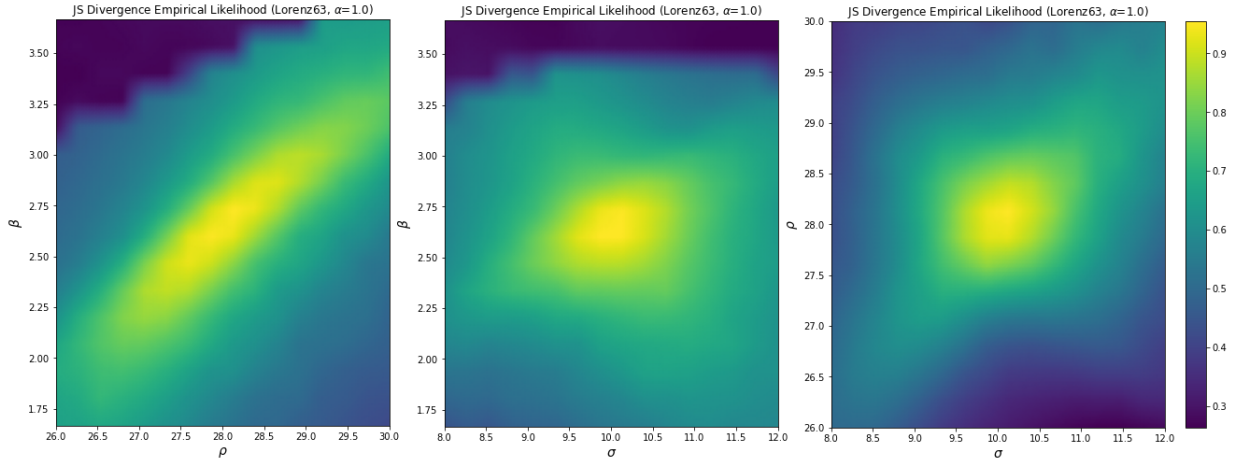
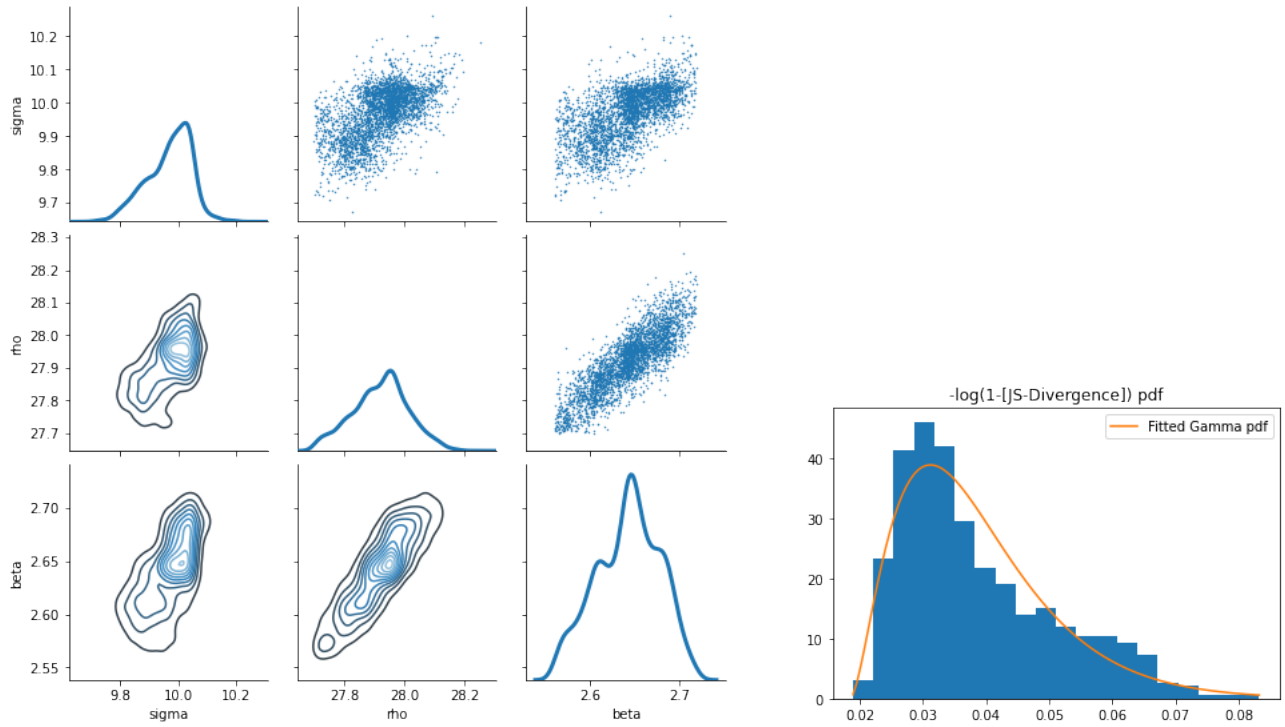


Figure 10: 2D slices of the Lorenz63 empirical likelihood function. For each slice, the invariant parameter is fixed at the nominal value.

We remark that this approximation if the JS divergence produces a distance measure between probability distributions that is much smoother and less computationally expensive to approximate than the Wasserstein metric. We also observe in Figure 10 that while the β-ρ relationship in the likelihood function is approximately Gaussian, the other bivariate relationships appear to be non-Gaussian and significantly differ from those observed in the correlation integral likelihood and the Wasserstein likelihood. We conjecture that this is due to the fact that information-based distance measures are agnostic toward the relative positions of points of high and low likelihood. Indeed, the KL divergence can be shown to be invariant under affine transformations, which eliminates any bias and variance that may be induced from scaling and normalizing the data.

In order to select the value of α in the likelihood function, we use the estimator given in (12). Fixing the parameters at the nominal values, we generated 1000 trajectories by selecting initial conditions close to the origin and approximated their JS divergence from the nominal dataset. Calculating $-\ln(1-D_{JS}(\mathbf{X}', \mathbf{X})/\ln(2))$ for each sample as described in (20), we obtained the pdf given in Figure 11b. Fitting the pdf to a gamma distribution, we obtained $\tilde{\alpha} = \tilde{\theta} = 2.31 \cdot 10^{-3}$. Finally, we used HMC sampling to obtain the posteriors as shown in Figure 11a.



(a) (b)

Figure 11: Posterior samples of Lorenz63 generated from HMC sampling of the JS divergence likelihood (a) and a uniform prior and the gamma distribution used to select the parameter α in the likelihood function (b).

In contrast to the the correlation integral posteriors, the JS divergence likelihood produced samples that are distinctively non-Gaussian, with positive correlations among all pairs of parameters. The means and modes of the posterior were close to the Lorenz63 nominal values $(\sigma, \rho, \beta) = (10, 28, 8/3)$ and were comparable the means and modes of the posterior samples obtained from the correlation integral likelihood.

3.4 Comparison of Methods

In Figure 12, we give some summary statistics of the three distance measures studied in this paper:

Method		**Nominal Value**	**Mean**	**Covariance Matrix**		
				Sigma	Rho	Beta
Correlation Integral	Sigma	10.00	10.13	0.120	0.046	0.001
	Rho	28.00	28.07	0.046	0.063	−0.006
	Beta	2.67	2.66	0.001	−0.005	0.001
Wasserstein Metric	Sigma	10.00	10.22 *	1.237	−0.042	0.065
	Rho	28.00	28.04 *	−0.042	0.523	0.059
	Beta	2.67	2.645 *	0.065	0.059	0.053
JS Divergence	Sigma	10.00	10.08	0.025	−0.012	0.003
	Rho	28.00	27.98	−0.012	0.020	−0.003
	Beta	2.67	2.68	0.003	−0.003	0.001

Figure 12: A table of statistics of the HMC posterior samples.

We remark that all three methods produced means close to the nominal values of the parameters. However, the means produced by the Wasserstein metric (marked with $*$) may be a result of influence from the uniform prior used in the HMC sampling, and thus they may not accurately reflect the true expected value of the distribution. This is a consequence of the noise in the 1024-cluster EMD approximation as seen in Figures 6b and 7, which

resulted in a high value of α being selected. Of the remaining two metrics, we note that the JS divergence method performed slightly better than the correlation integral method, producing a mean closer to the nominal value with slightly less variance.

4 CONCLUSION

In this paper we evaluated the efficacy of three candidate distance measure concepts that can be employed in determining the parameters of Lorenz63. We first investigated the use of the correlation integral method (5) as proposed by Haario et al. We then proposed an approximation scheme for both the Wasserstein distance metric and the Kullback-Leibler divergence. We adapted the KL divergence approximation to estimate the Jensen-Shannon divergence. These two distance measures were converted into empirical likelihood functions by embedding them in a Gaussian kernel (see (10), (19)) and selecting the scale parameter α according to (12). Each of these three methods were evaluated on the canonical Lorenz63 System given by (1). Although the Wasserstein metric proved difficult to approximate without significant error, both the JS divergence method and correlation integral method produced good estimates with low variance. In comparison, the JS divergence method produced a slightly better parameter estimate with slightly less variance in all of the parameters.

Future directions of this work include the analysis of these methods to other canonical chaotic systems and eventually to high-dimensional chaotic systems. The use of information-based metrics in the KL divergence family seems to be a promising avenue of exploration for modeling chaotic dynamical systems.

5 REFERENCES

References

[1] C. M. Greve, K. Hara, R. S. Martin, D. Q. Eckhardt, and J. W. Koo, *A data-driven approach to model calibration for nonlinear dynamical systems*, Journal of Applied Physics, 125 (2019), p. 244901.

[2] H. Haario, L. Kalachev, and J. Hakkarainen, *Generalized correlation integral vectors: A distance concept for chaotic dynamical systems*, Chaos: An Interdisciplinary Journal of Nonlinear Science, 25 (2015), p. 063102.

[3] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*, Wiley series in probability and mathematical statistics, Wiley, New York, 1992.

[4] Q. Wang, W. Huang, and J. Feng, *Multiple limit cycles and centers on center manifolds for lorenz system*, Applied Mathematics and Computation, 238 (2014), pp. 281 – 288.