# Likelihood-free Inference of Fornax Dark Matter Density Profile

**Mao-Sheng Liu**
maoshenl@andrew.cmu.edu

Committee:
**Katerina Fragkiadaki**
**Matthew Walker**

## Abstract

The standard model of cosmology $\Lambda$CDM predicts that dark matter (DM) density profile should diverge as $r^{-1}$ at the center of dwarf galaxies (cusp), while the observations tend to suggest a flatter profile (core). The discrepancy is known as the core-cusp problem, and it remains one of the unresolved controversy in small-scale cosmology. Bayesian inferences can be used to distinguish these two cases based on observed data. However, traditional methods like MCMC that relies on likelihood calculations is intractable because of the integration needed to marginalize the the unobserved phase space coordinates of stars. In this study we use conditional Variational Autoencoder (cVAE) and Mixture Density Network (MDN) to make likelihood-free estimation of the model parameter posteriors for Fornax dwarf Spheroidal (dSph) galaxy. Synthetic training sets are generated from a very flexible Strigari-Frenk-White (SFW) spheroidal galaxy model to train the deep networks. Our study shows that because of the need of cVAE to tune the relative importance of latent and reconstruction losses, and possible introduction of bias, we found MDN to be a more suitable method for this problem. Although most of the parameters are unconstrained, we show that the MDN has sensitivity to distinguish between core and cusp profile, and the result can be further improved with larger training set.

## 1 Background

The standard model of cosmology is known as $\Lambda$CDM. The model is successful in predicting the observable universe at the large scale, such as the temperature anisotropies of the cosmic microwave background radiation and the power spectrum of the large scale distribution of galaxies. However, at the sub-galaxy level, some controversies remain. One of which is the core-cusp problem. A particular robust prediction of the standard theory is that, in the absence of baryonic effects and external influences, a dwarf galaxy's dark matter (DM) density distribution should follow the NFW profile [1], where the central profile diverges as $r^{-1}$ (cusp). However, the observation of dwarf satellite galaxies around the Milky Way suggests a less dense and flatter central density profile (core) (e.g. [2]). The problem is particularly interesting because not only will it refine our cosmological model, it can also help to constrain plausible DM candidates (e.g. [3, 4]) and assist in our search for these elusive particles.

Our galaxy Milky Way is surrounded by many satellite dwarf galaxies, we have discovered dozens of them so far. These galaxies are the focal point of the controversy because their mass composed almost entirely of DM. Further more, their small mass and lack of star formation activities remove the possibilities that the central density profile could be disturbed by the baryronic effects. Dwarf spheroidal (dSph) galaxies, named because of their apparent spherical shape, in particularly, also exhibit little external influence from Milky Way. These conditions make dSph galaxies ideal candidates in shedding light on this issue, and Fornax dSph is one such galaxy.

A direct way of inferring the central density profile from a dwarf galaxy is by fitting a stellar dynamical model to a large sample of stars observed in the galaxy. However, because of the distance we can only observe the line-of-sight (los) velocities and angular positions of the stars in a dwarf galaxy. The traditional methods based on Jean's analysis produce ambiguous results due to parameter degeneracy [5]. With high quality data accumulating over recent years, challenges remain at finding the best ways to make inferences from such data. Ideally, we would calculate the likelihood directly to compare different models, but this is computationally intractable for complicated stellar models due to limited observations. In this project we propose to use recent development in deep learning to make likelihood-free inference from the data of Fornax dSph galaxy.

## 2 Related Work

It's a rather common situation that a complicated model can be simulated forward but has intractable likelihood. However, in Bayesian inference likelihood is needed to estimate the model posterior. One method developed to address this problem is Approximate Bayesian Computation (ABC) [6, 7]. ABC and its variants[8] bypasses the likelihood evaluation by simulating artificial data from the model under certain model parameters, which is compared with the real data. The set of parameters that's within $\epsilon$-distance away from the observation is kept to estimate the posterior. The ABC methods are implemented (e.g. [9, 10, 11]) and are used widely in cosmology (e.g. [12, 13, 14, 15]). However, there are generally two challenges involved in applying this method. One is to devise an effective distance measurement or summary statistics to compare the generated-data with the observed data, such that the distance is sensitive to the parameters of interest and is efficient to compute. Another problem is determination of appropriate size of $\epsilon$, which could critically slow down the algorithm if too small, or the estimated posterior becomes too broad if it's too large. Our methods using neural networks don't have the second problem, while the first problem is greatly alleviated because the actual features from data can be extracted directly from the user input, although the specific form of the user input also needs to be chosen with care.

As the amount of observation data increases exponential in recent years, application of deep learning methods are rapidly expanding in cosmology. For example, convolution neural network is used to classify gravitational lensing images [16] and estimate cosmological parameters [17]. Autoencoder is applied to extract features from large multi-weavelength galaxy surveys [18]. Generative models, such as variational auto encoder (VAE) and generative adversarial network (GAN), are used to emulated computationally intensive cosmological simulations[19] and to generate realistic galaxy and astronomical images[20, 21]. Here we use deep learning methods to make inference of model parameters and their uncertainties.

## 3 Deep Learning Methods

We use two approaches in deep learning to make likelihood-free inferences. One is based on Variational Bayesian method [22, 23], specifically we use the conditional Variational Autoencoder (cVAE). The second approach is to estimate the the posterior directly with mixtures of Gaussians, or the Mixture Density Network (MDN).

### 3.1 Conditional Variational Autoencoder

Variational autoencoder (VAE) is traditionally used to learn a low dimensional representation, latent variables $z$, of high dimensional data $x$, such as images or audios. It is a generative model, therefore it is possible to sample $x$ thought $z$ from the learned distribution $p(x|z)$. The mathematics behind the technique is Variational Inference (VI), where the posterior is approximated from the result of the optimization.

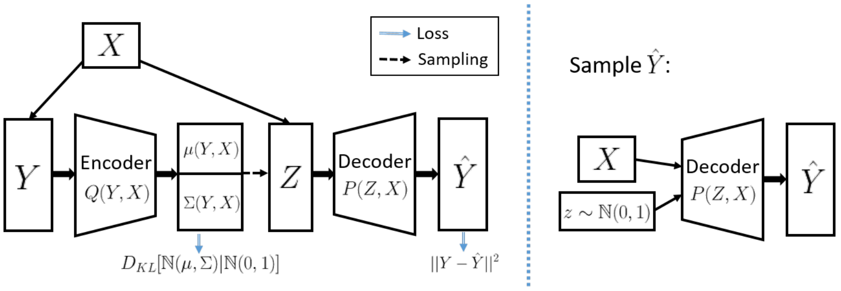

Figure 1: On the left is the conditional Variational Autoencoder network. For a given model parameter $Y$ and the corresponding galaxy X, $Q$ encodes the input into the latent space $Z$. The decoder network decodes $Z$, condition on the same X, to predict the model parameter $\hat{Y}$. We assume a $\mathbb{N}(0,1)$ prior for $Z$, and square error reconstruction loss. One the right, to sample possible values of $Y$ for a given galaxy $X$, we draw $z \sim \mathbb{N}(0,1)$ to decode the possible $\hat{Y}$.

In VI, posterior distribution over latent variables $p(z|x)$ is approximated by $q(z|x)$. Assume that we want to compute the log-probability of observations $x$, which is govern by latent variables $z$:

$$\begin{aligned} log(p(x)) &= \sum_z q(z|x)\log(p(x)) \\ &= \sum_z q(z|x)\log\left(\frac{p(z,x)}{q(z|x)}\right) + \sum_z q(z|x)\log\left(\frac{q(z|x)}{p(z|x)}\right) \\ &= \hat{L} + D_{KL}(q(z|x)||p(z|x)) \geq \hat{L} \end{aligned} \tag{1}$$

where $\hat{L}$ is the variational lower bound, and $D_{KL}(q(z|x)||p(z|x)) \geq 0$ is the Kullback–Leibler (KL) divergence. Since $p(x)$ doesn't depend on the choice of $q(z|x)$, one can optimize $\hat{L}$, and in the process the $D_{KL}$ term will approach 0 (i.e. $q(z|x)$ approaches $p(z|x)$). Using $p(z,x) = p(x|z)p(z)$, we can further rewrite the variational lower bound to arrive at the objective function of the VAE:

$$\begin{aligned} \hat{L} &= \sum_z q(z|x)\log\left(\frac{p(z)}{q(z|x)}\right) + \sum_z q(z|x)\log\left(p(x|z)\right) \\ &= -D_{KL}(q(z|x)||p(z)) + \mathbb{E}_{q(z|x)}\log\left(p(x|z)\right) \end{aligned} \tag{2}$$

The first is the regularization term, where $p(z)$ could be a simple uniform or Gaussian prior. The second term is the reconstruction quality (i.e. if a $z$ can perfectly reproduce a $x$, the quality would be high).

The cVAE is a straightforward extension of VAE, where the probability distributions in Eq. 1 and Eq. 2 are conditioned on an input. In this study we seek to infer the probability distribution of the model parameters given an input galaxy. Specifically for a galaxy $X$ and parameter $Y$, we want to approximate the distribution $P(Y|X)$. Rewrite the expressions in Eq. 1 and 2, condition on the galaxy $X$, we arrive at the objective:

$$\begin{aligned} \log P(Y|X) - D_{KL}[Q(z|Y,X)||P(z|Y,X)] = \; & \mathbb{E}_{z\sim Q(\cdot|Y,X)}[\log P(Y|z,X)] \\ & - D_{KL}[Q(z|Y,X)||P(z|X)] \end{aligned} \tag{3}$$

Figure 1 illustrate our cVAE network. The encoder net $Q(z|Y,X)$ assumes a Gaussian distribution. It encodes the input into a multi-dimensional mean and variance of the Gaussian distribution. The latent variable $Z$ is sampled from the $Q$, and is subsequently decoded condition on the same input galaxy $X$ by the decoder network $P(Y|z,X)$. The regularization loss $D_{KL}(Q(Z|Y,X)|P(Z))$ enforces the $Q$ to be a standard normal distribution. During the testing time, latent $z$ can be randomly drawn from a standard normal distribution to sample the possible parameters $\hat{Y}$ for a given galaxy $X$. Note that because of the random sampling of $Z$ in the network, the optimization of the network is done with backpropagation using the "reparametrization trick" proposed by Kingma and Welling [22].

### 3.2 Mixture Density Network

The second approach is the MDN first proposed by Bishop [24]. The true posterior density $P(Y|X)$ is approximated by neural network representing a mixture of $m$ Gaussian densities, $Q(Y|X) = \sum_{i=0}^{m} w_i(X)\mathbb{N}(Y|\mu_i(X), \Sigma_i(X))$, as show in Figure 2. It is a simple feedforward neural network where the outputs are the means, covariance matrices, and weights of the the mixture models. The objective is to maximize the log-probability log$[Q(Y|X)]$. During the test time, the parameters $\hat{Y}$ of a galaxy $X$ can be drawn from the mixture of Gaussian $Q(Y|X) = Q(Y|W(X), \mu(X), \Sigma(X))$ directly.

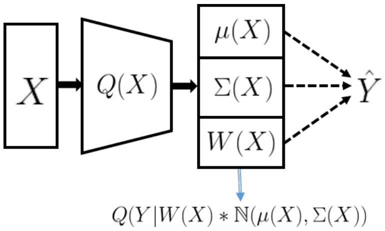

Figure 2: Mixture Density Network. Given a galaxy $X$, a feedforward neural network estimate directly the means $\mu$, covariance matrices $\Sigma$, and weights $W$ of Gaussian components that are used to approximate the truth posterior $P(Y|X)$.

In theory, MDN is capable of representing any conditional distribution arbitrarily accurately, provided that the number of mixture components $m$ and the size of the neural network is sufficiently large, in addition to a sufficiently large training set to fit the network parameters.

## 4 Fornax Data

The dataset that we analyze for this project comes from the Magellan (or MMFS) survey [25], which includes an observation of 2483 individual stars in the Fornax dwarf spheroidal galaxy. For each star the projected x, y positions on the sky, and the line-of-sight (los) velocities $v_z$ were measured. The los velocity measurements rely on the Doppler shift of light: if a star is moving toward (away) from the observer, the star's light spectrum would be red-shifted (blue-shifted). The measurement requires observation of light spectra of individual stars, therefore not all of the stars' los velocities are observed. In addition, stars are observed non-uniformly across the galaxy.

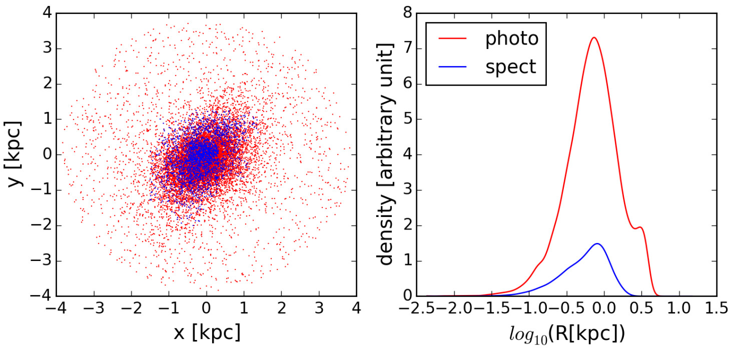

Figure 3: Left panel shows the scatter plot of the positions of the Fornax stars, the photometric dataset is in red, and the spectroscopic dataset is in blue. The right panel shows the density of projected radius $R$ for these two dataset.

The left panel of Figure 3 shows all the observed stars in Fornax. The red dots (a total of 14068 stars) has only the photometric measurement (i.e. only the x, y position of these stars are known)[26]. The stars in blue is our dataset that include the spectroscopic los velocity measurements. Since the galaxy is spheroidal, the position can be represented by projected radius $R = \sqrt{x^2 + y^2}$. The right panel shows the distribution of the projected radius of both photometric $D_{\text{photo}}(R)$ and spectroscopic datasets $D_{\text{spect}}(R)$. The relative density of stars at each $R$ is important because it represent the sampling bias of our data. When generating the synthetic dataset for training, we also need to apply the same bias/mask to make a correct inference. Specifically, once a synthetic galaxy is generated, based on the position R of each star, the star only has a probability of $D_{\text{spect}}(R)/D_{\text{photo}}(R)$ to be kept in the galaxy. The synthetic stars beyond the maximum range of Fornax photometric data is rejected directly.

## 5 Galaxy Model and Data Simulation

The neural network models introduced in section 3 are trained with synthetic data generated from a spheroidal galaxy model introduced by Strigari, Frenk, and White [27]. Since these models typically only accept a fixed array of values in a fixed ordering as inputs, we experiment with different input methods to find an optimal approach.

### 5.1 Strigari-Frenk-White (SFW) model

Strigari, Frenk, and White introduce a flexible density function, SFW model [27], of distribution of stars in an equilibrium spheroidal galaxy. It is a separate function of specific energy $E$ and specific angular momentum $J$: $f(E, J) = g(J)h(E)$.

The distribution of angular momentum is defined as,

$$g(J) = \begin{cases} [1 + (J/J_\beta)^-b]^{-1} & \text{for } b \leq 0 \\ 1 + (J/J_\beta)^b & \text{for } b > 0 \end{cases}. \tag{4}$$

It describes a transition, at scale angular momentum $J_\beta$, from a flat angular momentum core to a power law with index b. The energy distribution function is,

$$h(E) = \begin{cases} NE^a(E^q + E_c^q)^{d/q}(\Phi_{lim} - E)^e & \text{for } E < \Phi_{lim} \\ 0 & \text{for } E \geq \Phi_{lim} \end{cases}. \tag{5}$$

Where N is the normalization constant. The function also follows a power law, with different slopes at different energy level, set by characteristic energy scale $E_c$. The distribution is truncated at the point where the star has enough energy to escape the galaxy, set by the potential at the limiting radius $\Phi(r_{\text{lim}}) = \Phi_{\text{lim}}$.

The energy of a star can be written as $E = v^2/2 + \Phi(r) = (v_r^2 + v_t^2)/2 + \Phi(r)$ and the angular momentum is $J = rv_t$. The radius $r$ is the distance of the star from the center of the galaxy, $v_r$ and $v_t$ are the radial and the tangential velocities of the star with respect to the galaxy center. $\Phi(r)$ is the potential energy which can be calculated directly from the mass density of the galaxy, which is approximately the DM density $\rho_{DM}(r)$ because the system we are interested in is DM dominated,

$$\Phi(r) = 4\pi G \left[ \int_0^r r'^2 \rho_{DM}(r')dr' + \int_r^\infty r'\rho_{DM}(r')dr' \right] + \Phi_0$$

Where $G$ is the gravitational constant and $\Phi_0$ is a constant offset that makes $\Phi(0) = 0$. We consider a DM distribution described by a generalized Hernquist profile [28, 29],

$$\rho_{DM}(r) = \rho_s \left(\frac{r}{r_s}\right)^{-\gamma} \left[1 + \left(\frac{r}{r_s}\right)^\alpha\right]^{(\gamma-\beta)/\alpha} \tag{6}$$

To simply the model we set $\alpha = 1$ and $\beta = 3$, which makes it a generalized NFW[1] profile with a flexible central slope $\gamma$. With $\gamma = 0$ being a core model and $\gamma = 1$ being a cusp model. Therefore we have a total of 11 free parameters that we need to infer from this SFW model: $Y$ = [a, d, e, Ec, $r_{\text{lim}}$, b, q, $J_\beta$, $\rho_s$, $r_s$, $\gamma$]. Note that $\rho_s$ has units $M_\odot/kpc^3$, $r_{\text{lim}}$ and $r_s$ have unit of kpc, $J_\beta$ is in units of $r_s\sqrt{\Phi_s}$, and $E_c$ has a unit of $\Phi_s$, where $\Phi_s = 4\pi G\rho_s r_s^2$.

Here we also reparametrize $\rho_s$ as mass enclosed within the half-light radius $R_h$, $M_{\text{enc}}$. Given the half light radius of the Fornax data ($R_h$ = .72kpc, the median value in the photometric dataset), one can reformulate the $\rho_s$ as $M_{\text{enc}}$, using $M_{\text{enc}} = M(R_h) = 4\pi \int_0^{R_h} r^2\rho_{DM}(r)dr$.

| a | d | e | Ec | $\log(r_{\text{lim}})$ | b | q | $J_b$ | $\log(M_{\text{enc}})$ | r$_s$ | $\gamma$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $[0, .71]$ | $[-10, 1]$ | $[0, 5]$ | $[.05, .51]$ | $[-.5, 1.5]$ | $[-5, 5]$ | $[.5, 10.5]$ | $[.01, .5]$ | $[6, 9]$ | $[-1, 0.5]$ | $[0, 1.5]$ |

Table 1: Prior range of model parameters. Parameters are drawn randomly within the ranges to generate galaxies for training.

### 5.2 Training Data Generation and Transformation

#### 5.2.1 Synthetic data generation

We use the rejection sampling method to draw training samples. Assume that we want to sample from a density $p(x)$, let $q(x)$ be a simple proposal distribution (e.g., a uniform or normal distribution), and given that there is a constant $C$: $1 \leq C < \infty$, such that $p(x) \leq Cq(x)$ for all $x$. The rejection sampling steps are:

1. Draw sample $x^i \sim q(x)$
2. Draw $u^i \sim$ Uniform(0,1)
3. If $u^i < \frac{p(x^i)}{Cq(x^i)}$, accept $x^i$; otherwise, reject.

For this model we have $x^i = (r^i, v_r^i, v_t^i)$, and we set $q(x) = \max(f(r, v_r, v_t|Y))$, the maximum of the density function, and we let $C = 1$. Typically a galaxy of 10,000 stars can be drawn in a few seconds. The sampling efficiency or speed varies depends on the model parameter $Y$.

Based on the spherical symmetry, the set of $\{r, v_r, v_t\}$ is transformed into Cartesian coordinates $\{x, y, z, v_x, v_y, v_z\}$. And because in practice we only observe $x$, $y$, and los velocity $v_z$, we only keep $\{R, v_z\}$ of each galaxy for the training, where $R = \sqrt{x^2 + y^2}$ is the projected radii of the stars.

The prior range of each parameter is listed in Table 1. Each training sample is generated from the parameters that are drawn randomly from these priors. Note that during training, each parameter in the "label" $Y$ is rescaled to have a range between $[0, 1]$ so that each parameter is treated equally by the neural network.

#### 5.2.2 Data input methods

Since $\{R, v_z\}$ is a set of continuous random points without specific ordering, it can't be directly feed into an neural network. Here we try a few different methods. First is try to bin the data into 2-D histogram. An alternative approach is to use Gaussian kernel density estimate to obtain a 2D density estimate of the data, then calculate the density across a 2D mesh-grid of $(R, v_z)$ coordinates. The third method we try is DeepSets [30]. The technique allows an input of a set of data (like ours) to the neural network, and the objective function is invariant with respect to the permutation of elements in the set. Here we apply a version of DeepSets, where each stars in the galaxy $(R^i, v_z^i)$ is transformed to a higher dimension representation by a neural network, and subsequently an average pooling is applied to this set of representations, thus achieves permutation invariance.

The performance of each representation of galaxy is evaluated by running each input through a feedforward neural network to make regression of the galaxy parameters with $L_2$ loss. In this experiment, we use 100,000 galaxies as the training set and 20,000 galaxies as the validation set. Each galaxy has 2500 stars. Only four parameters are varying for this test: $Y = [M_{\text{enc}}, a, r_{\text{lim}}, \gamma]$, and other parameters are held fixed. The loss on the validation dataset result is show in Figure 4. The red curve shows the validation loss if the input is a binned 2D histogram. Specifically each galaxy is binned into 128x128 bins, again in 64x64 bins, 32x32 bins, ..., and so on until 2x2bins, then they are normalized by total number of stars and number of bins, and the results are combined to form a single input. The combination is to account for the fact that we aren't sure what's the optimal choice of the bin number. Here we see that this approach has the worst performance.

Green and blue curves on Figure 4 are the results by using Gaussian kernel density to first estimate the 2D density of $(R, v_z)$, and 30x30 (blue) and 100x100 (green) mesh-grid values are calculated to represent the galaxy. We see that they have the best performance, with 30 grids performs better

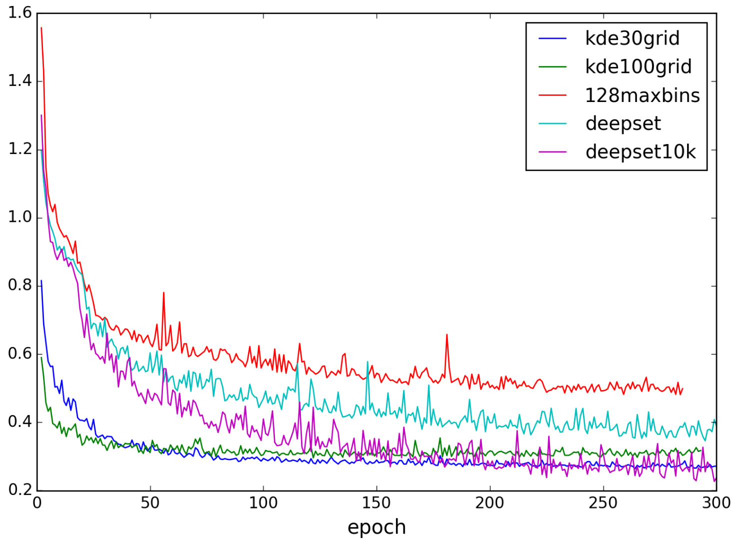

Figure 4: The plot shows the validation loss of different methods of feeding galaxy data $\{R, v_z\}$ into the neural network. In red, stars are progressively binned (128x128+64x64+32x32+....+2x2) using 2D histogram and combined as a single input. In blue and green, joint density of $(R, v_z)$ is obtained by Gaussian kernel density estimate, and density at 30x30 and 100x100 mesh-grid points are evaluated as representation of the galaxy. Teal and purple runs are the results using DeepSets, with purple curve uses 4 times as much (10,000) stars as input.

than 100 grids. The reason could be that since most of these 100x100=10,000 values are zero, adding more zeros makes each galaxy less distinguishable from each other. The purple and teal curves shows the results of DeepSets, where $\{R, v_z\}$ are put directly into the network. The result in teal has an input of 2500 stars, while the purple run has 10,000 stars. It shows that with larger number of stars the performance improves, as expected. The method is promising, especially for problems where the input has higher dimension and kernel density estimate can't be successfully applied. Overall performing kernel density estimate and takes the density at a 30x30 mesh-grids is currently the best option, and will be used for final runs. It is also faster to train because of smaller input size.

## 6 Results

Synthetic data are generated as described in the section 5.1, each galaxy is first generated with 14086 stars, the same number of stars as the Fornax photometric data. Each star is then rejected with probability based on the relative density of the spectroscopic data and the photometric data, as described in section 4. The remaining stars is kernel smoothed, and the density at the 30x30 mesh-grid point are evaluated to represent the galaxy. Note that the mesh-grid boundary is set to encompass the whole range of possible $(R, v_z)$ in the training set. We generated two sets of training data, the smaller one has 100,000 galaxies, and the larger set contains 1 million galaxies. Each parameter is rescaled to be between 0 and 1 based on the prior range in Table 1.

To test the accuracy of our estimations we infer the parameters of two specific mock galaxies. One galaxy has a central cusp profile and the other one has a central core, and they both have similar mass enclosed within half-light radius. The inferred parameters for the mock core and cusp datasets are shown in Figure 5 and 6, respectively.

The blue histograms are the sampled posteriors from the trained cVAE, while the red distributions are the results from MDN. Solid and dash lines are trained on large and small training sets, respectively. On results of cVAE, we see that there is no constrain on most of the parameters, except $\mathrm{r_{lim}}$ and $\mathrm{M_{enc}}$. The model makes a good prediction on these two parameters particularly on the cusp galaxy in Figure 6. The method makes virtually the same estimation of $\gamma$ posterior for both core and cusp galaxies. The increase in the training set size doesn't seem to improve the cVAE inference significantly. Only

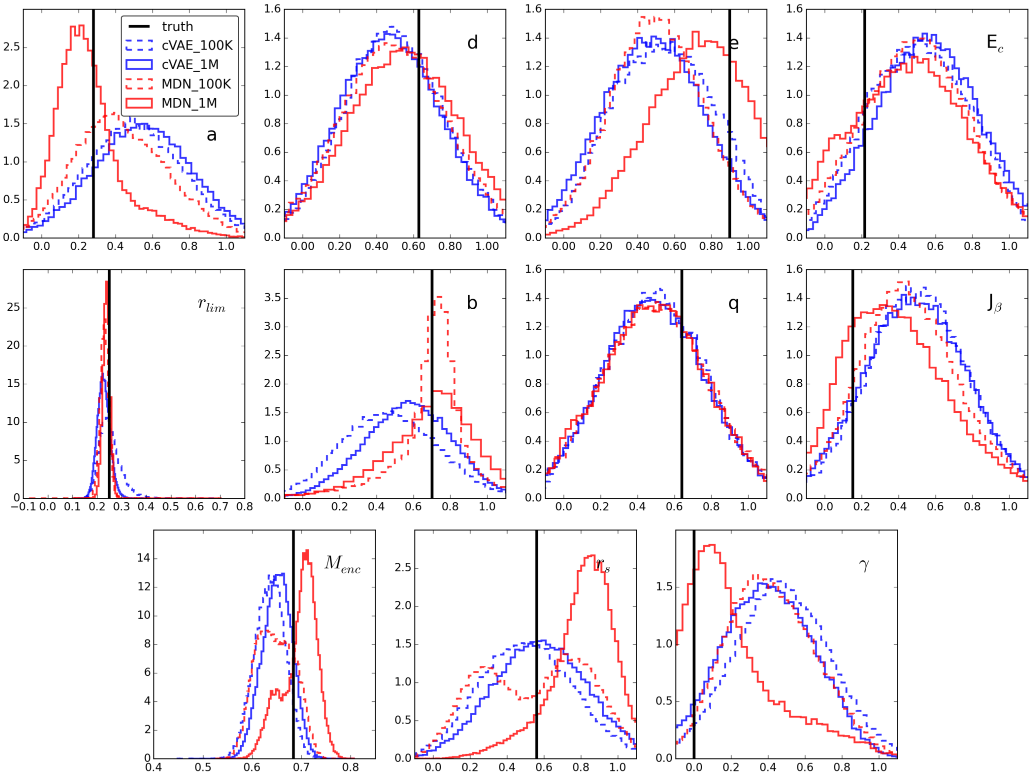

Figure 5: Inference on a central core mock galaxy. Blue distributions are the posteriors learned from cVAE using $10^5$ (dash) and $10^6$ galaxies training datasets, respectively. Similar, red dash and read solid histograms are the results from MDN model with $1x10^5$ and $1x10^6$ galaxies training sets, respectively. The truths are the vertical lines in black.

for parameters r$_{\rm lim}$ and $b$ does the larger training set seem to nudge the distributions toward the truths. We will discuss the difficulties in training the cVAE model in section 7.

On the other hand, the posteriors based on MDN looks much more encouraging. Unlike cVAE, we see that in red curves there is significant improvement when the training set size is increased. Specifically for core model in Figure 5, we see the distribution of parameters $a$, M$_{\rm enc}$, $e$, and $\gamma$ shift toward truth with larger training set. Even the MDN trained on smaller training set seems to outperforms the cVAE. Note that in some cases we see multi-modality in MDN inference, as shown in parameter $r_s$ and possibly in parameter M$_{\rm enc}$. This is most likely due to not having enough number of mixture components. Increase the number of mixture densities will likely removes the bi-modality. Although MDN doesn't strictly constrain the central density slope, the posterior for $\gamma$ does show that the network is sensitive to the truth. For the core galaxy, the peak shifts toward $\gamma = 0$, while for the cusp model the peak of the distribution is toward 0.6 (which means $\gamma = 0.9$ without scaling by the prior), corresponds to a cusp central profile. The result strongly suggest that a even larger net and training dataset can improve the constrain on $\gamma$ parameter.

To further check the accuracy of MDN train on the large training set, we estimate the posterior of 300 randomly generated galaxies. The results on $\gamma$, r$_{\rm lim}$, and M$_{\rm enc}$ are plotted in Figure 7. On each panel the blue scatter points are the means of the predicted posteriors, the corresponding standard deviations are plotted as the error bars. The green line is the linear fit of the means, and the red line indicates the truth. The middle and the right panels shows that the MDN can make a very accurate prediction of r$_{\rm lim}$ and M$_{\rm enc}$, but the predictions of $\gamma$ parameter have large error bars.

Finally we make inference on the Fornax data set using the same cVAE and MDN models. Figure 8 shows the estimated posterior with same color and line schemes as in Figures 5 and 6. As discussed earlier, the result from cVAE is mostly uninformative, except for r$_{\rm lim}$ and M$_{\rm enc}$. The posterior that we can trust the most is the solid red curves, corresponds to MDN trained on large dataset. The scaled $\gamma$

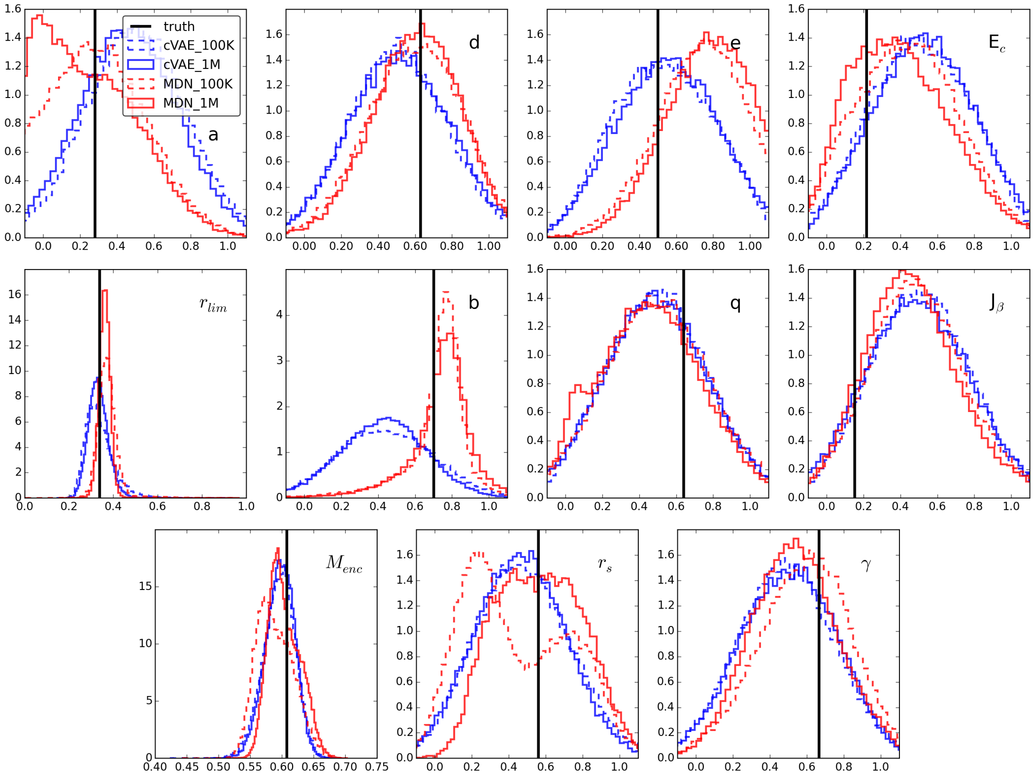

Figure 6: Same as Figure 5, except that the inference is done on a cuspy mock galaxy.

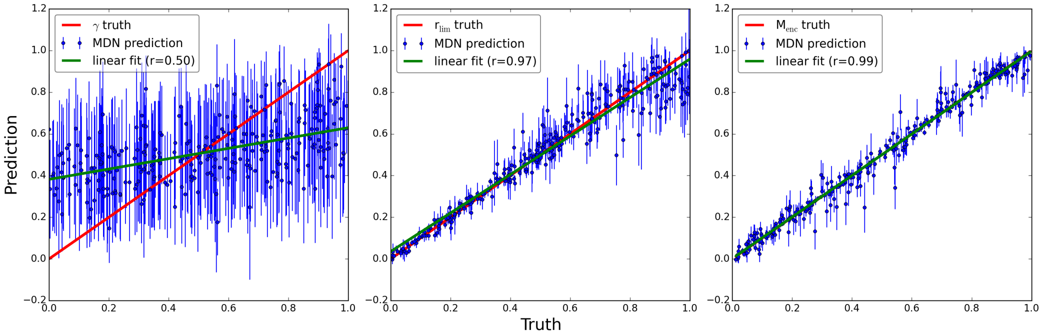

Figure 7: Means and variances of predicted $\gamma$, r$_{\rm lim}$, and M$_{\rm enc}$ for 300 randomly generated mock galaxies using MDN with large training set. Red line is the truth and the green line is the linear fit of the mean points. MDN shows sensitivity on $\gamma$ as the shown by the positive correlation (r=0.50), but the scatter is too large to be conclusive.

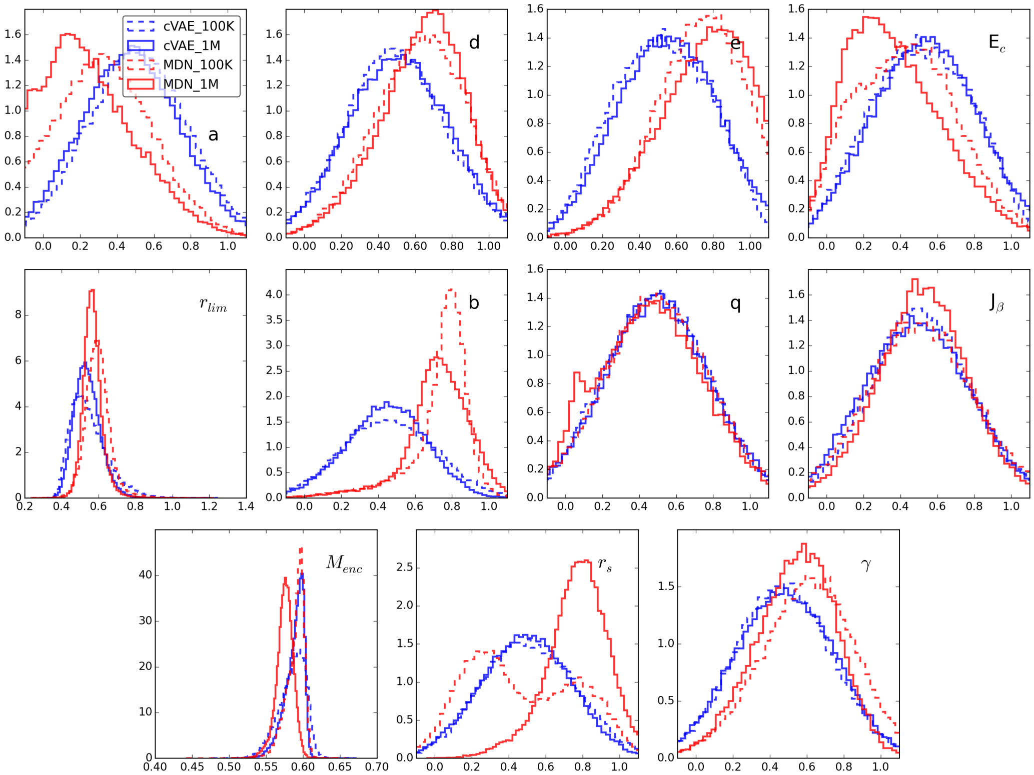

Figure 8: Inference of Fornax data, using the same learned cVAE and MDN models as in Figure 5 and 6.

parameter is peaked at around $0.6$, which mean $\gamma = 0.6 \times 1.5 = 0.9$, corresponds to a cusp central density profile. However, the result is inconclusive as we see in Figure 7, the error bar is still too large.

## 7 Discussion and Future Work

The performance of cVAE doesn't improve as much with the size of the training set. I think it's because of the fundamental limitation of the network architecture as shown in Figure 1. During the training time, regularization loss falls first and quickly compared to the reconstruction loss, as a result decoder relies only on input X to learn $\hat{Y}$. Since Z becomes essentially a pure noise, and decoder tries to ignores its input. As a result, the sampled $\hat{Y}$ has very little variance (Z plays little roles in predicting $\hat{Y}$). Since the reconstruction loss is also high, often time the network will learn a posterior $P(Y|X)$ that rejects the truth. Alternatively one can put extra weight on the reconstruction loss, so the network will first try to lower the reconstruction loss, however having a large weight makes the network ignore the regularization term, therefore Y is directly pass through the network and is used to predict $\hat{Y}$ itself. The result is that decoder couldn't learn to predict $\hat{Y}$ from X and Z alone. During the sampling time and without Y in the network, sampled posterior could be biased and/or has large variance (i.e. no constrain on the parameters). The cVAE results show in the previous figures are first trained on the cVAE network without extra weight on the reconstruction loss. This allows decoder to have chance to learn to predict Y, after a certain epoch extra weight to the reconstruction loss is introduced which increases the variance in the posterior. However this strategy certainly doesn't solve the fundamental problem of this architecture for our application, and the larger training set doesn't help much to improve the outcome either.

The result from last section shows that with larger training data the performance of MDN can be improved. It also shows that the network is sensitive to the central density profile, or the $\gamma$ parameter.

The next step is to further increase the size of the training set, and if the trend continues, we might be able to robustly constrain the $\gamma$ parameter of the Fornax dSph galaxy.

**Acknowledgments**

I would like to thank François Lanusse for helpful discussions and his insights on the working of cVAE and MDN models, which prompted me to use the MDN method. I would also like to thank Manzil Zaheer for helpful discussions on how to use DeepSets as input to our neural networks.

## References

[1] J. F. Navarro, C. S. Frenk, and S. D. M. White. The Structure of Cold Dark Matter Halos. , 462:563, May 1996.

[2] M. G. Walker and J. Peñarrubia. A Method for Measuring (Slopes of) the Mass Profiles of Dwarf Spheroidal Galaxies. , 742:20, November 2011.

[3] M. R. Lovell, V. Eke, C. S. Frenk, L. Gao, A. Jenkins, T. Theuns, J. Wang, S. D. M. White, A. Boyarsky, and O. Ruchayskiy. The haloes of bright satellite galaxies in a warm dark matter universe. , 420:2318–2324, March 2012.

[4] M. Rocha, A. H. G. Peter, J. S. Bullock, M. Kaplinghat, S. Garrison-Kimmel, J. Oñorbe, and L. A. Moustakas. Cosmological simulations with self-interacting dark matter - I. Constant-density cores and substructure. , 430:81–104, March 2013.

[5] J. Binney and S. Tremaine. *Galactic Dynamics: Second Edition*. Princeton University Press, 2008.

[6] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[7] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[8] Brandon M Turner and Trisha Van Zandt. A tutorial on approximate bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012.

[9] Elise Jennings and Maeve Madigan. astroabc: An approximate bayesian computation sequential monte carlo sampler for cosmological parameter estimation. *Astronomy and computing*, 19:16–22, 2017.

[10] Joël Akeret, Alexandre Refregier, Adam Amara, Sebastian Seehars, and Caspar Hasner. Approximate bayesian computation for forward modeling in cosmology. *Journal of Cosmology and Astroparticle Physics*, 2015(08):043, 2015.

[11] Justin Alsing, Benjamin Wandelt, and Stephen Feeney. Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. *arXiv preprint arXiv:1801.01497*, 2018.

[12] ChangHoon Hahn, Mohammadjavad Vakili, Kilian Walsh, Andrew P Hearin, David W Hogg, and Duncan Campbell. Approximate bayesian computation in large-scale structure: constraining the galaxy–halo connection. *Monthly Notices of the Royal Astronomical Society*, 469(3):2791–2805, 2017.

[13] Ewan Cameron and AN Pettitt. Approximate bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 425(1):44–65, 2012.

[14] Tomasz Kacprzak, Jörg Herbel, Adam Amara, and Alexandre Réfrégier. Accelerating approximate bayesian computation with quantile regression: application to cosmological redshift distributions. *Journal of Cosmology and Astroparticle Physics*, 2018(02):042, 2018.

[15] Anja Weyant, Chad Schafer, and W Michael Wood-Vasey. Likelihood-free cosmological inference with type ia supernovae: approximate bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal*, 764(2):116, 2013.

[16] François Lanusse, Quanbin Ma, Nan Li, Thomas E Collett, Chun-Liang Li, Siamak Ravanbakhsh, Rachel Mandelbaum, and Barnabás Póczos. Cmu deeplens: deep learning for automatic image-based galaxy–galaxy strong lens finding. *Monthly Notices of the Royal Astronomical Society*, 473(3):3895–3906, 2017.

[17] Siamak Ravanbakhsh, Junier B Oliva, Sebastian Fromenteau, Layne Price, Shirley Ho, Jeff G Schneider, and Barnabás Póczos. Estimating cosmological parameters from the dark matter distribution. In *ICML*, pages 2407–2416, 2016.

[18] Joana Frontera-Pons, Florent Sureau, Jerome Bobin, and Emeric Le Floc'h. Unsupervised feature-learning for galaxy seds with denoising autoencoders. *Astronomy & Astrophysics*, 603:A60, 2017.

[19] Mustafa Mustafa, Deborah Bard, Wahid Bhimji, Rami Al-Rfou, and Zarija Lukić. Creating virtual universes using generative adversarial networks. *arXiv preprint arXiv:1706.02390*, 2017.

[20] Siamak Ravanbakhsh, Francois Lanusse, Rachel Mandelbaum, Jeff G Schneider, and Barnabas Poczos. Enabling dark energy science with deep generative models of galaxy images. In *AAAI*, pages 1488–1494, 2017.

[21] Jeffrey Regier, Andrew Miller, Jon McAuliffe, Ryan Adams, Matt Hoffman, Dustin Lang, David Schlegel, and Mr Prabhat. Celeste: Variational inference for a generative model of astronomical images. In *International Conference on Machine Learning*, pages 2095–2103, 2015.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[23] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *ArXiv e-prints*, January 2016.

[24] Christopher M Bishop. Mixture density networks. 1994.

[25] M. G. Walker, M. Mateo, and E. W. Olszewski. Stellar Velocities in the Carina, Fornax, Sculptor, and Sextans dSph Galaxies: Data From the Magellan/MMFS Survey. , 137:3100–3108, February 2009.

[26] TMC Abbott, FB Abdalla, S Allam, A Amara, J Annis, J Asorey, S Avila, O Ballester, M Banerji, W Barkhouse, et al. The dark energy survey data release 1. *arXiv preprint arXiv:1801.03181*, 2018.

[27] L. E. Strigari, C. S. Frenk, and S. D. M. White. Dynamical Models for the Sculptor Dwarf Spheroidal in a ΛCDM Universe. , 838:123, April 2017.

[28] L. Hernquist. An analytical model for spherical galaxies and bulges. , 356:359–364, June 1990.

[29] H. Zhao. Analytical models for galactic nuclei. , 278:488–496, January 1996.

[30] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, pages 3394–3404, 2017.

[31] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[32] George Papamakarios and Iain Murray. Fast $\varepsilon$-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, pages 1028–1036, 2016.

[33] EEO Ishida, SDP Vitenti, M Penna-Lima, J Cisewski, RS de Souza, AMM Trindade, E Cameron, and VC Busti. Cosmoabc: Likelihood-free inference via population monte carlo approximate bayesian computation, astronomy and computing 13 (2015) 1–11. *arXiv preprint arXiv:1504.06129*.

[34] Laurence Perreault Levasseur, Yashar D Hezaveh, and Risa H Wechsler. Uncertainties in parameters estimated with neural networks: Application to strong gravitational lensing. *The Astrophysical Journal Letters*, 850(1):L7, 2017.

[35] Anirudh Jain, PK Srijith, and Shantanu Desai. Variational inference as an alternative to mcmc for parameter estimation and model selection. *arXiv preprint arXiv:1803.06473*, 2018.