A deep generative model for astronomical images of galaxies

Jeffrey Regier[1], **Jon McAuliffe**[1], **Prabhat**[2]
[1]Department of Statistics, University of California, Berkeley
[2]Lawrence Berkeley National Laboratory

1 Introduction

Galaxies in astronomical images (Figure 1) often resemble galaxy prototypes (Figure 2) and possess shared characteristics like spiral "arms", a "bar", or a "bulge". Even irregular galaxies—typically resulting from the collision of two regular galaxies—have shapes greatly constrained by physics.

Figure 1: The Whirlpool galaxy—a classic spiral galaxy.
credit: ESA Hubble / NASA.

Simple parametric models (e.g. [1, 2, 3, 4, 5]) suffice to describe many idealized galaxy shapes, but severely misfit actual galaxies: they are not sufficiently flexible [5]. The popular program GALFIT copes with the limitations of simple parametric models by allowing users to fit an arbitrary number of mixture components [6, 7]. These mixtures are not learned from actual galaxies, so the models cannot provide meaningful uncertainty estimates. To our knowledge, no existing galaxy models are learned from a training set, which would allow for such uncertainty estimates. Indeed, only [2] attempts a Bayesian treatment of galaxy shapes[1], albeit one based on just a few manually specified parameters. Yet modeling galaxies is an important part of learning about the universe from large-scale astronomical sky surveys [4, 5, 8, 9], and billions of images of galaxies are available for training.

Neural networks—high-dimensional parametric models—have enjoyed great success for classifying images [10]. They have been effective at discriminating between stars and galaxies [11] and for labeling images of galaxies as possessing or lacking certain features, such as a "bar" or spiral "arms" [12]. However, to our knowledge no one has yet reported on meldingmarrying the flexibility of neural networks to a generative probabilistic model for galaxies. Recent advances in variational inference for non-conjugate models [13, 14] make this possible. To our knowledge, this is the first publication to report applying these advances to a problem in the physical sciences.[2]

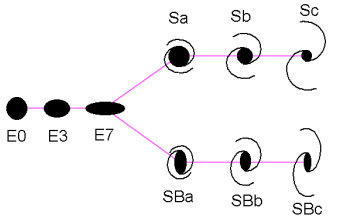

Figure 2: The Hubble "tuning fork" of galaxy morphology.
credit: Todd Thompson.

2 The model

For a particular image x of a galaxy, let z be a low-dimensional latent random vector, distributed as a multivariate standard normal. Given z, we model the observed intensities of the image's pixels $x = (x_1, \ldots, x_m)$, as

$$x|z \sim \mathcal{N}\left(f_\mu(z), f_\Sigma(z)\right). \tag{1}$$

[1]The galaxy models in [4] and [5] are deterministic though they are embedded in probabilistic models.
[2] We are informed by a review of reverse citations and personal correspondence with Diederik Kingma.

We take the deterministic functions f_μ and f_Σ to be neural networks that share some weights. We constrain f_Σ to produce diagonal covariance matrices. As shorthand, let the neural network $f(z) := (f_\mu(z), f_\Sigma(z))$.

2.1 Inference

Given an image's pixel intensities $x = (x_1, \ldots, x_m)$, we aim to infer the posterior distribution of $z = (z_1, \ldots, z_n)$. Unfortunately, integrating z out of the joint distribution (x, z) to compute the marginal likelihood of x is intractable due to the nonlinear form of f. Therefore, we turn to variational inference. In keeping with the approaches in [13] and [14], let the variational approximate posterior take the form

$$q(z|x) = \mathcal{N}(g_\mu(x), g_\Sigma(x)), \tag{2}$$

where g_μ and g_Σ are neural networks that map x to a mean vector and a diagonal covariance matrix, respectively. As shorthand, let $g(x) := (g_\mu(x), g_\Sigma(x))$. By the standard construction of the variational lower bound,

$$\begin{aligned} \log p(x) &\geq \log p(x) - D_{\mathrm{KL}}\left[q(z|x), p(z|x)\right] &(3)\\ &= \mathbb{E}_q\left[\log p(x|z)\right] - D_{\mathrm{KL}}\left[q(z|x), p(z)\right]. &(4)\end{aligned}$$

Therefore, the distribution q that maximizes (4) minimizes $D_{\mathrm{KL}}\left[q(z|x), p(z|x)\right]$: this q is the best approximation of form (2) to the posterior. Let W_f and W_g be the weights of neural networks f and g, respectively. Maximizing over $W = (W_f, W_g)$ simultaneously finds the q that best approximates the posterior and the model p that assigns the highest probability to our data.

The normal-normal KL-divergence $D_{\mathrm{KL}}\left[q(z|x), p(z)\right]$ has a closed form, but $\mathbb{E}_q\left[\log p(x|z)\right]$ is not. We can nonetheless efficiently compute unbiased estimates of its gradient, and therefore maximize (4) by stochastic gradient optimization.

The stochastic gradient described in Example 5.1 of [15] and [13, 14, 16], based on "the reparameterization trick", has the lowest variance among all unbiased estimators. Let $\epsilon \sim \mathcal{N}(0, I)$. Then

$$\begin{aligned} \frac{\partial}{\partial W}\mathbb{E}_q\left[\log p(x|z)\right] &= \frac{\partial}{\partial W}\mathbb{E}_\epsilon\left[\log p\left(x|z = g_\Sigma(x)\epsilon + g_\mu(x)\right)\right] &(5)\\ &= \mathbb{E}_\epsilon\left[\frac{\partial}{\partial W}\log p\left(x|z = g_\Sigma(x)\epsilon + g_\mu(x)\right)\right]. &(6)\end{aligned}$$

Hence, for e sampled from ϵ,

$$\frac{\partial}{\partial W}\log p\left(x|z = g_\Sigma(x)e + g_\mu(x)\right) \tag{7}$$

is an unbiased estimate of the derivative of $\mathbb{E}_q\left[\log p(x|z)\right]$.

3 Experiments

We apply our model to preprocessed 424×424-pixel images of galaxies from the Sloan Digital Sky Survey [17, 18]. In keeping with the approach of [12], we crop each image to surround just the most prominent galaxy and downscale these subimages to 69×69 pixels. Based on a blob detection routine, we exclude images where the most prominent galaxy overlaps with other bright astronomical objects[3], leaving 43,444 images for training.

3.1 Implementation

Fitting our model by stochastic gradient descent involves simultaneously training two neural networks: f, for specifying the generative model $p(x|z)$, and g, for specifying the variational distribution $q(z|x)$.

An alternative perspective is helpful for implementing the fitting procedure: Both f and g are components in a single neural network called a "generalized denoising autoencoder" (GDAE) [19, 20].

[3] "Deblending" astronomical objects is a related problem, likely facilitated by an accurate galaxy model, but beyond the scope of this work.

An image x is input to g, yielding $g_\mu(x)$ and $g_\Sigma(x)$. The next layer in the GDAE corrupts $g_\mu(x)$. Its inputs are $g_\mu(x)$, $g_\Sigma(x)$, and a sample e from $\mathcal{N}(0, I)$. Its output is $g_\mu(x) + g_\Sigma(x)e$. This output z serves as the input to f. The output of f is penalized by the expected negative reconstruction error: $-\log p\left(x|z\right)$. As a form of regularization, the output of g is penalized too, according to $D_{\mathrm{KL}}\left[q(z|x), p(z)\right]$.

This perspective facilitates adapting existing neural network software to learn the proposed generative model. Mocha.jl [21] is a neural network toolkit written in Julia, inspired by Caffe [22]. We reuse the basic framework from Mocha.jl, but augment it with new types of layers to compute the proposed loss function. The parts of the network corresponding to f and g each have two fully connected hidden layers composed of 128 hidden nodes each, with rectified linear units. The parts corresponding to the output layers of f and g use exponential nonlinearities to ensure that variances are strictly positive. We set the dimension of z to eight. On an Nvidia Tesla K20X GPU, our network performs roughly 200 iterations per second. (Each iteration involves forward and backward propagation for one image.) Parameter-specific learning rates are set adaptively [23].

3.2 Results

First, we examine the trained model qualitatively. Figure 3 shows sample input images to the trained autoencoder from a held-out set, and the resulting output. The mean of the reconstruction $f_\mu(z)$, $z \sim \mathcal{N}\left(g_\mu(x), g_\Sigma(x)\right)$, resembles a smoothed version of the input. The variance of the reconstruction $f_\Sigma(z)$ is low for the backgrounds, which by construction is nearly black in the original images. Variance is higher for the foreground, particularly near the borders of each galaxy—presumably z cannot store enough information to represent slight differences in galaxies' sizes. The intensity of the third galaxy's center is particularly uncertain, which may reflect that some but not all galaxies have a prominent "bulge" in the center.

Figure 4 shows a two-dimensional embedding of a held-out set of galaxies, generated by applying t-SNE [24] to the 8-dimensional means $g_\mu(x)$ of the galaxies' variational distributions. At this resolution, galaxies are clearly grouped by their orientations. Some clustering of spiral galaxies is apparent too.

Figure 5 shows $f_\mu(z)$, that is, the mean of $p(x|z)$, for values of z selected by a one-at-a-time experimental design.

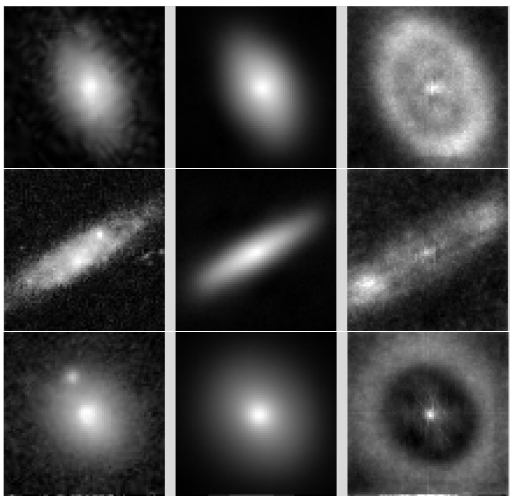

Figure 3: Each row corresponds to a different example from a test set. The left column shows the input x. The center column shows the output $f_\mu(z)$ for a z sampled from $\mathcal{N}\left(g_\mu(x), g_\Sigma(x)\right)$. The right column shows the output $f_\Sigma(z)$ for the same z.

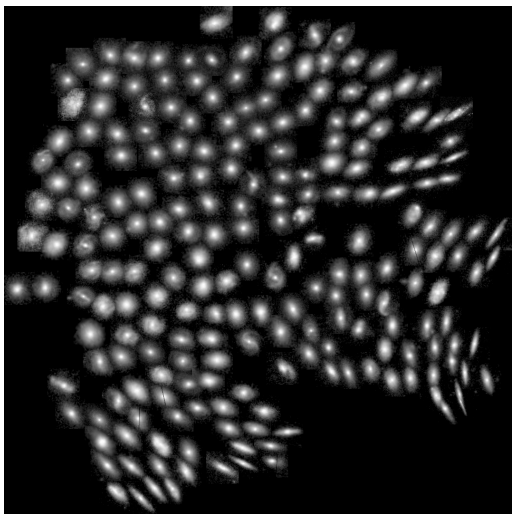

Figure 4: Galaxies embedded in two dimensions based on the means of their variational distributions, $f_\mu(x)$.

Because the model we propose is, to our knowledge, the first galaxy model learned from a training set, comparing it to existing galaxy models is not straightforward. Comparison is also challenging because the most common galaxy models do not explicitly model uncertainty.

We also compare the proposed galaxy model to a current common practice: fitting a scaled bivariate Gaussian density function to each imaged galaxy. On a held-out dataset of 1000 images of galaxies, we compute $f_\mu(g_\mu(x))$. This amounts to running the proposed autoencoder with the layer for sampling z replaced with the mean of z. For each image, we also fit a scaled bivariate Gaussian density to minimize squared error averaged over pixels. The optimization was performed with BFGS over six unconstrained parameters: two for the mean, three for the Cholesky decomposition of the covariance, and one for the scale. For 971 of 1000 images, $f_\mu(g_\mu(x))$ fit x more closely than the best scaled bivariate Gaussian density. In some sense this is not surprising, since only the proposed model uses training data. On the other hand, the parameters of the proposed model are only optimized by a feed-forward recognition model rather than an iterative algorithm, and only the scaled bivariate Gaussian model is explicitly trained to minimize residual sum of squares.

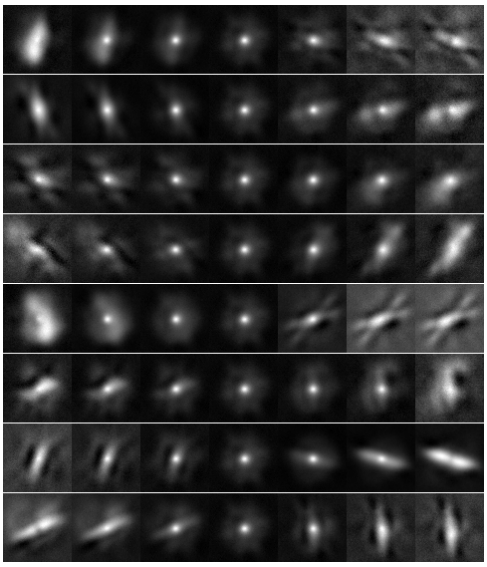

Figure 5: $f_\mu(z)$ for z values sampled according to a one-at-a-time experimental design. In each row, from left to right, one dimension of z is incremented by one standard deviation per column, while the other dimensions are fixed at zero. The center column in each row is $f_\mu(0, \ldots, 0)$. The leftmost and rightmost columns are 3 standard deviations from the mean and thus highly unlikely; we show these extremes to highlight the effect of each dimension of z.

Fitting a function to minimize residual squared error averaged across pixels is analogous to maximizing the likelihood of the data for a model where all pixels have a Gaussian distribution with a common variance. This interpretation lets us compare our proposed model, conditioned on a particular z, to the scaled bivariate Gaussian density function in terms of log likelihood. Now for each image, in addition to fitting a scaled bivariate Gaussian density function to each held-out image, we learn the variance shared by all pixels that assigns the highest likelihood to the image. (The solution has a closed form.) We compare this to the likelihood assigned to the data by the model we propose, for a particular z. For 972 of 1000 images, the model we propose better explains the data. Both models treat each pixel's value as Gaussian, but only the model we propose assigns different variance to each pixel.

4 Future work

The proposed model shows little sign of overfitting our existing training set, and billions of additional images of galaxies are freely available. By increasing the dimension of z and by making our network deeper, we could almost certainly improve held-out accuracy. Augmenting f with intermediate latent layers [14] would also likely improve held-out accuracy and better model uncertainty about the structure of the galaxies, rather than just uncertainty at the level of individual pixels.

We could also exploit the rotational and reflective symmetries of galaxies, either through data augmentation or with a network architecture that explicitly enforces it.

Our immediate focus, however, is on embedding the current galaxy model into the model for raw astronomical images (not cropped around galaxies) described in [5]. Augmenting the broader model with this data-adaptive galaxy model likely will improve its performance across the board.

Acknowledgments

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] Robert Lupton, Zeljko Ivezic, et al. SDSS image processing II: The photo pipelines. Technical report, Princeton University, 2005.

[2] Lance Miller, Thomas Kitching, et al. Bayesian galaxy shape measurement for weak lensing surveys–i. Methodology and a fast-fitting algorithm. *Monthly Notices of the Royal Astronomical Society*, 382(1):315–324, 2007.

[3] Luc Simard, Christopher Willmer, et al. The deep groth strip survey. II. Hubble space telescope structural parameters of galaxies in the groth strip. *The Astrophysical Journal Supplement Series*, 142(1):1, 2002.

[4] Dustin Lang and David Hogg. Tractor: Astronomical source detection, separation, and photometry. `http://thetractor.org/`, 2015. [Online; accessed January 30, 2015].

[5] Jeffrey Regier, Andrew Miller, Jon McAuliffe, et al. Celeste: Variational inference for a generative model of astronomical images. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[6] Chien Y. Peng et al. Detailed structural decomposition of galaxy images. *The Astronomical Journal*, 124(1):266, 2002.

[7] Chien Y. Peng et al. Detailed decomposition of galaxy images. II. Beyond axisymmetric models. *The Astronomical Journal*, 139(6):2097, 2010.

[8] Marco Barden et al. GALAPAGOS: From pixels to parameters. *Monthly Notices of the Royal Astronomical Society*, 422(1):449–468, 2012.

[9] Boris Häußler et al. Megamorph—multiwavelength measurement of galaxy structure: complete sérsic profile information from modern surveys. *Monthly Notices of the Royal Astronomical Society*, 430(1):330–369, 2013.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] Emmanuel Bertin and Stephane Arnouts. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series*, 117(2):393–404, 1996.

[12] Sander Dieleman, Kyle Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

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

[14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[15] James C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2005.

[16] Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1971–1979, 2014.

[17] Chris Stoughton, Robert Lupton, et al. Sloan digital sky survey: early data release. *The Astronomical Journal*, 123(1):485, 2002.

[18] Galaxy zoo—the galaxy challenge. `https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge`, 2013. [Online; accessed October 1, 2015].

[19] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013.

[20] Shakir Mohamed. A statistical view of deep learning. `http://blog.shakirm.com/wp-content/uploads/2015/07/SVDL.pdf`, 2015. [Online; accessed October 1, 2015].

[21] Chiyuan Zhang. `https://github.com/pluskid/Mocha.jl`, 2015. [Online; accessed October 5, 2015].

[22] Yangqing Jia et al. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.