Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314013568

Estimating the distribution of Galaxy Morphologies on a continuous space

Giuseppe Vinci[1], Peter Freeman[1], Jeffrey Newman[2], Larry Wasserman[1] and Christopher Genovese[1]

[1] Dept. of Statistics, Baker Hall, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
email: gvinci@andrew.cmu.edu

[2]Dept. of Physics & Astronomy, University of Pittsburgh, 310 Allen Hall 3941 O'Hara St., Pittsburgh, PA 15260, USA

Abstract. The incredible variety of galaxy shapes cannot be summarized by human defined discrete classes of shapes without causing a possibly large loss of information. Dictionary learning and sparse coding allow us to reduce the high dimensional space of shapes into a manageable low dimensional continuous vector space. Statistical inference can be done in the reduced space via probability distribution estimation and manifold estimation.

Keywords. dictionary learning, manifold estimation, Radon transform, redshift, sparse coding, galaxies: statistics

1. Introduction

The evolution of the Universe has led to the formation of complex objects apparently without any regular shape, which our mind would just classify as *irregular*. Thus, the incredible variety of galaxy shapes cannot be summarized by human defined discrete classes of shapes (e.g. "Hubble sequence") without causing a possibly large loss of information. Our human concept of shape could limit the complete understanding of the complex structure of the galaxies. Estimating the distribution of galaxy morphologies is one means to test theories of the formation and the evolution of the Universe. We estimate the distribution of morphologies on a *continuous* Euclidean space, such that a particular shape will be viewed as a point in a continuous space. This task must be performed in an *unsupervised* way, i.e. free from any human judgement. Galaxy images are intrinsically high-dimensional data, and we use *dictionary learning* and *sparse coding* [Mairal *et al.* (2010)] to reduce the high dimensional space of shapes into a manageable low dimensional one. Essentially, galaxy images will be approximated by sparse linear combinations of basis pictures, which are *learned* from the data. Statistical inference on the reduced space can be performed via probability distribution estimation. We propose a testing procedure and analyse a dataset of galaxy images† to show some examples.

2. Dictionary Learning and Sparse Coding - Radon Transform

The general idea of dictionary learning and sparse coding is to approximate images by *sparse* linear combinations of a fixed number of *basis images*, which are not predefined,

† GOODS-South Early Release Science Field dataset observed in the near-infrared regime by the Wide Field Camera 3 on-board the Hubble Space Telescope [see Windhorst *et al.* (2011), Freeman *et al.* (2013)].

but are *learned* from the data. Let $x_i \in \mathbb{R}^{a\times b}$ be an image, which has $a \times b$ dimensions. For $m << a \times b$, we want to approximate x_i as:

$$x_i \approx \sum_{j=1}^{m} \alpha_{ij} B_j \tag{2.1}$$

where $\alpha_i = (\alpha_{i1}, ..., \alpha_{im}) \in \mathbb{R}^m$ is a sparse vector of coefficients, and $\{B_j\}_{j=1}^m$ is a collection of basis images $B_j \in \mathbb{R}^{a\times b}$. Notice that the basis images will not be imposed to be orthogonal such that the dictionary can easily adapt to the structure of the data [Mairal *et al.* (2010)]. Moreover, learning the bases from the data was shown to perform better in signal reconstruction with respect to using predefined bases [Elad *et al.* (2006)].

2.1. *Optimization problem*

From a dataset of galaxy images $\{x_i\}_{i=1}^n$, we can estimate the dictionary $D = \{B_j\}_{j=1}^m$ and the vectors of coefficients $A = \{\alpha_i\}_{i=1}^n$ by solving the following optimization problem:

$$\begin{cases} \min\limits_{\{\alpha_i\}_{i=1}^n, \{B_j\}_{j=1}^m} \sum\limits_{i=1}^{n} \left[\frac{1}{2} \left\| x_i - \sum\limits_{j=1}^{m} \alpha_{ij} B_j \right\|_2^2 + \underbrace{\lambda \|\alpha_i\|_1}_{\text{Sparsity}} \right] \\ \textbf{s.t.} \quad \|B_j\|_2^2 \leqslant 1, \forall j = 1, ..., m \end{cases} \tag{2.2}$$

where $\lambda \geqslant 0$ is a sparsity parameter and $\|*\|_2^2$ is the Frobenius norm [Mairal *et al.* (2010); R package "spams"]. We suggest to choose m and λ via *cross validation*. See Mairal *et al.* (2010) for other configurations of problem (2.2).

2.2. *Standardization of the images. Radon transform*

Before solving problem (2.2), images must be standardized to eliminate any spurious dimensionality and improve the quality of the approximations (2.1). We are talking about: *centring*, *resizing* and *rotation orientation*. While the first one can be easy to perform, the two others are not. Images can be rotated and resized by using Radon Transform (RT) and Inverse RT (IRT). The RT of a function f is $\mathcal{R}_f(t,\theta) = \int_{-\infty}^{\infty} f(t\cos\theta - u\sin\theta, t\sin\theta + u\cos\theta)du$, where $(t,\theta) \in \mathbb{R}^2$. An image can be viewed as the discrete evaluation of a function. The orientation of the texture of an image can be estimated by $\theta^* = \arg\min_\theta \frac{\partial^2 \sigma_\theta^2}{\partial \theta^2}$, where σ_θ^2 is the variance of $\mathcal{R}_f(t,\theta)$ at angle θ [Jafari-Khouzani *et al.* (2005), Arodź (2012); R package "PET"]. Rotating images by angle $-\theta^*$ essentially makes all the pictures *horizontally oriented*. To rotate an image we need to: 1) evaluate its RT on a discrete grid, say $\hat{R}_{M\times(\omega 180+1)} = \{\mathcal{R}_f(t,\theta)\}$ with $t \in \{t_1, ..., t_M\}$, $\theta \in \{\frac{j}{\omega 180}\pi\}_{j=0}^{\omega 180}$, and $\omega \in \mathbb{N}^+$; 2) find θ^* and move the first $k^* = \theta^* \frac{\omega 180}{\pi}$ columns of $\hat{R}$ as described in Figure 1 to get $\tilde{R}$ ("rotation" in the Radon domain); 3) computing the IRT of $\tilde{R}$ on a grid of desired resolution ("resizing"). In Figure 2 we show some effects of images standardization.

3. Statistical inference on the reduced space

In this section we propose a method to estimate the distribution of galaxy morphologies on a low-dimensional space, and we use the GOODS-S dataset to perform a simulation.

3.1. *Probability distribution of galaxy morphologies.*

For a dataset of n images $\{x_i\}_{i=1}^n$, where x_i is a matrix of nonnegative light intensity:

(*a*) Standardize all the images as described in paragraph 2.2;

(*b*) Obtain the dictionary D and the vectors $A = \{\alpha_i\}_{i=1}^n$ according to paragraph 2.1;

https://doi.org/10.1017/S1743921314013568 Published online by Cambridge University Press

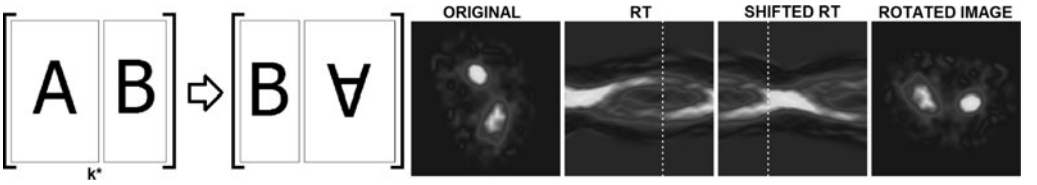

Figure 1. Left: vectors A are moved after vectors B with values moved up and down. Right: starting from an original image, we compute its Radon transform on a discrete grid, then by shifting the vectors of this matrix according to the orientation θ^*, we can obtain a standardized rotated version of the image as the IRT of the shifted RT.

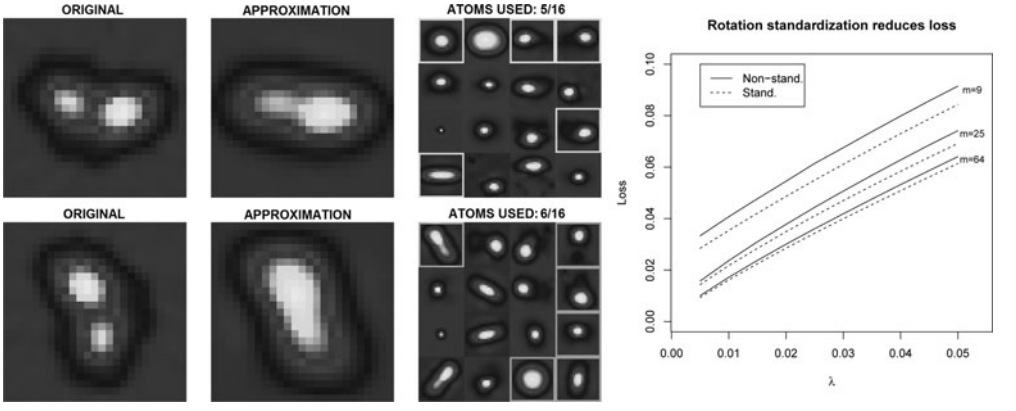

Figure 2. Rotation standardization improves the fit. Left: an image approximated using a dictionary learned with rotation standardization (top) and not (bottom). Spurious dimensionality negatively affects the dictionary at the bottom, while rotation standardization may lead to more refined approximations. Right: for different numbers of atoms ($m = 9, 25, 64$), the minimum loss (2.2) is smaller when using standardized images. Images are from the GOODS-S dataset, H-band.

(c) Estimate the joint distribution of vector $\alpha_i \in \mathbb{R}^m$. Call it $\hat{P}_\alpha$.

Given the fitted dictionary D, estimate $\hat{P}_\alpha$ can be viewed as an approximation of the distribution of galaxy morphologies.

3.2. *Comparing populations of shapes*

In this section we propose a method to compare the distributions of two collections of images. Let X, Y be two collections of images. Suppose we want to test hypothesis $X \overset{\mathcal{D}}{=} Y$, i.e. a distribution test. We propose the following method:

(a) Pool X and Y into a unique dataset $Z = [X, Y]$

(b) From Z, fit dictionary D and vectors of coefficients $\{\alpha_{Z,k}\} = [\{\alpha_{X,i}\}, \{\alpha_{Y,j}\}]$.

(c) Implement a distribution test $\alpha_X \overset{\mathcal{D}}{=} \alpha_Y$.

For step (c), we suggest to use the nonparametric test based on the Maximum Mean Discrepancy (MMD) statistic (Gretton *et al.* (2012); R package "kernlab"). We can call this testing procedure "DSM test" (Dictionary Learning - Sparse Coding - MMD).

3.2.1. *Simulation*

We selected two subsets of images of the GOODS-S dataset in the H-band (see Figure 3): X_1 with 25 images of non-mergers, and X_2 with 25 images of mergers. To generate n images of non-mergers and n images of non-mergers we: 1) randomly sample with replacement n images from X_1 and n images from X_2, respectively; 2) randomly rotate them by angles $\theta \sim$Unif$(0, 2\pi)$, i.i.d.; 3) add heteroscedastic noise: $\epsilon_{jk} \overset{\text{indep}}{\sim} N(0, \beta^2 \times I_{jk})$,

https://doi.org/10.1017/S1743921314013568 Published online by Cambridge University Press

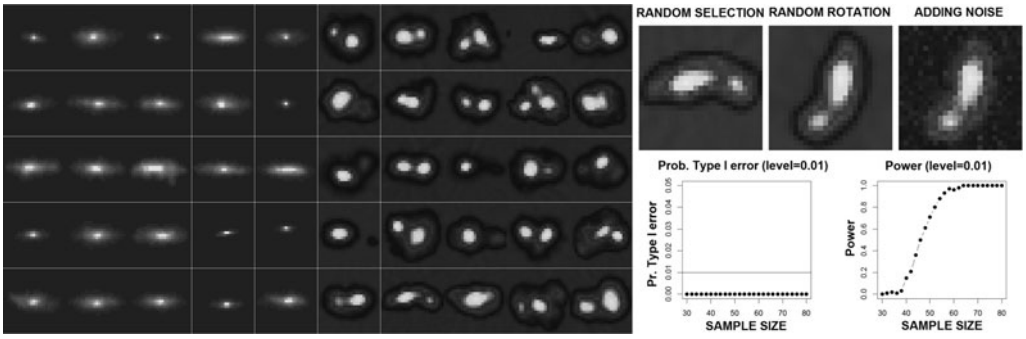

Figure 3. Left: selected non-mergers (X_1) and mergers (X_2) from the GOODS-S dataset, H-band. Top right: procedure to simulate an image from X_i. An image is randomly selected from the subset, randomly rotated and heteroscedastic Gaussian noise is added to each pixel. Bottom right: the DSM test helps to distinguish different shapes. The probability of Type I error of the DSM test is always smaller than the level of the test; the power of the test is increasing in the sample size. The shape of the power function depends on the original sets X_1, X_2.

where $I_{jk} \geqslant 0$ is the light intensity at position jk in a matrix. We repeat comparisons (via DSM test) of samples of the same kind (Mer Vs Mer, NMer Vs NMer) and different one (Mer Vs NMer) to estimate the probability of Type I error and the power of the test as functions of the sample size (see Figure 3). We chose $m = 4$ and $\lambda = 0.05$ via 10-CV.

4. Conclusions and future work

An unsupervised analysis based on dictionary learning and sparse coding allows us to approximate the distribution of galaxy morphologies by a multivariate distribution defined on a subset of $\mathbb{R}^m$, where dimension m is much smaller than the dimension of a galaxy image. Hypothesis testing on the reduced space can help to distinguish the distributions of two sets of images. Current and future work is: using dictionary learning and sparse coding to put constraints on the parameters of cosmological models; comparing the distribution of galaxy shapes at different redshift ranges; manifold estimation: some clusters may correspond to some human defined shapes (e.g. spiral, elliptical) and filaments [see Chen *et al.* (2013)] may describe the transition from a shape to another one; analysing images of other astronomical objects and 3D images.

References

Arodź, T. 2012, *Computing and Informatics*, 24 no. 2 (2012): 183-199.

Chen, Y.-C., Genovese, C. R., & Wasserman, L. 2013, *arXiv preprint* arXiv:1312.2098 (2013).

Elad, M. & Aharon, M. 2006, *Image Processing, IEEE Transactions on* 15, no. 12 (2006): 3736-3745.

Freeman, P. E., R. Izbicki, A. B. Lee, J. A. Newman *et al.* 2013, *MNRAS* (2013): stt1016.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., & Sriperumbudur, B. K. 2012, *Advances in neural information processing systems*, pp. 1205-1213. 2012.

Jafari-Khouzani, K. & Soltanian-Zadeh , H. 2005, *Pattern Analysis and Machine Intelligence*, IEEE Transactions on 27, no. 6 (2005): 1004-1008.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. 2010, *The Journal of Machine Learning Research* 11 (2010): 19-60.

Windhorst, R. A., Cohen, S. H., Hathi, N. P., McCarthy, P. J. *et al.* 2011, *ApJS* 193, no. 2 (2011): 27.

https://doi.org/10.1017/S1743921314013568 Published online by Cambridge University Press