# Probabilistic Model-Based Detection of Bent-Double Radio Galaxies

Technical Report UCI-ICS 02-14
Department of Information and Computer Science
University of California, Irvine

Sergey Kirshner, Igor V. Cadez, Padhraic Smyth,
Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
{skirshne,icadez,smyth}@ics.uci.edu

Chandrika Kamath and Erick Cantú-Paz
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, CA 94551
{kamath2,cantupaz}@llnl.gov

August 27, 2002

**Abstract**

We describe an application of probabilistic modeling to the problem of recognizing radio galaxies with a bent-double morphology. The type of galaxies in question contain distinctive signatures of geometric shape and flux density that can be used to be build a probabilistic model that is then used to score potential galaxy configurations. The experimental results suggest that even relatively simple probabilistic models can be useful in identifying galaxies of interest in an automatic manner.

# 1 Introduction

In this paper we investigate the problem of identifying bent-double radio galaxies in the FIRST (Faint Images of the Radio Sky at Twenty-cm) Survey data set [1]. FIRST produces large numbers of radio images of the deep sky using the Very Large Array at the National Radio Astronomy Observatory. It is scheduled to cover more that 10,000 square degrees of the northern and southern caps (skies). Of particular scientific interest to astronomers is the identification and cataloging of sky objects with a "bent-double" morphology, indicating clusters of galaxies. (For an example, see Figure 1.) Due to the very large number of observed deep-sky radio sources (over 600000 clusters as of 2000), it is infeasible for the astronomers to label all of them manually.

In this paper we propose a probabilistic approach for classification of bent-double configurations. We describe our model for bent-doubles and discuss possible usage of the model with both parametric and non-parametric methods. Since part of the problem is to properly orient the configuration, we describe an iterative algorithm to find proper orientations for the given set of configurations.

# 2 Data

The data from the FIRST Survey is available in two different formats. In the "raw image" format, image cut-outs are available from the FIRST website (http://sundog.stsci.edu/) with a resolution of 1.8 seconds squared per pixel. The second data format is in the form of a catalog of features that have been automatically derived from the raw images by an image analysis program [7]. Each entry corresponds to a single detectable "blob" of bright intensity (a sky object) relative to the sky background: these entries are called **components**. The "blob" of intensities for each component is fitted with an ellipse (details in [7]). The ellipses and intensities for each ellipse are described by a set of estimated features such as sky position of the centers, (RA (right ascension) and Dec (declination)), peak density flux and integrated flux, root mean square (RMS) noise, lengths of the major and minor axes, and the position angle of the major axis of the ellipse counterclockwise from the north. The goal is to find sets of components that are spatially close and that resemble a bent-double. In this paper we focus on the classification of candidate sets of components that have

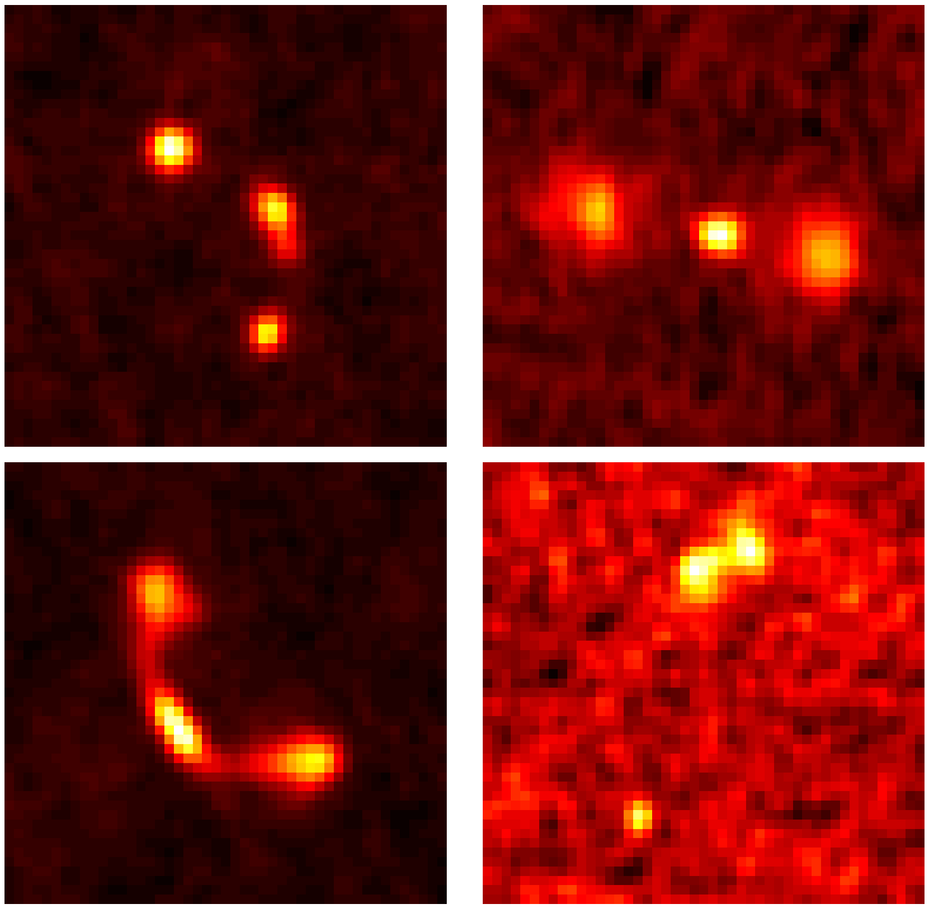

Figure 1: Examples of bent-double (left) and non-bent-double (right) configurations. Notice that the configuration on the top right does not have enough "bend" while the configuration on the bottom right does not exhibit symmetry . Cutouts are $47 \times 47$ pixels each.

| Number of Catalog Entries | Number Radio Sources |
|---|---|
| 1 | 514637 |
| 2 | 66571 |
| 3 | 15059 |
| 4+ | 6333 |

Table 1: Number of components per cluster

been detected by an existing spatial clustering algorithm [3] where each set consists of three components from the catalog (three ellipses). As of 2000, the catalog contained over 15,000 three-component configurations and over 600,000 configurations total (see Table 1 for more details). It is worth mentioning that vast majority of bent-doubles consist of three components. Three-component bent-double configurations typically consist of a center or "core" component and two other side components called "lobes".

The labeled set which we use to build and evaluate our models consists of a total of 128 examples of bent-double galaxies and 22 examples of non-bent-double configurations. A configuration is labeled as a bent-double if at least two astronomers labeled it as such. Note that the visual identification process is the bottleneck in the process since it requires significant time and effort from the scientists, and is subjective and error-prone. This motivates the creation of automated methods for identifying bent-doubles. This data set is also considerably biased towards the bent-double class (i.e., bent-doubles are far more prevalent in this training data set than they are in the catalog in general). This is an artifact of the manner in which scientists generated a labeled data set. However, since we use a likelihood-based approach for ranking candidate objects, where a model is built only on positive examples (bent-doubles), the training methodology presented below is not sensitive to such an imbalance in the training data.

Previous work on automated classification of three-component candidate sets has focused on the use of decision-tree classifiers using a variety of geometric and image intensity features [2] [3] [4]. A limitation of the decision-tree approach is its relative inflexibility in handling uncertainty about the object being classified, e.g., the identification of which of the three components should be treated as the core of a candidate object. A primary motivation for the development of a probabilistic approach is to provide a framework that can handle such uncertainties in a coherent manner. In particular, in this paper, we focus on a probabilistic mixture model that treats the identification of the center component as a hidden variable, providing a natural framework for handling this uncertainty both in the model-building phase (on training data) and in the detection phase (on test data).

## 3 Probabilistic Modeling of Bent-Double Galaxies

We denote a three-component **configuration** by $\mathcal{C} = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3)$, where the $\mathbf{c}_i$'s are the elliptical components as described in the previous section. Each component $\mathbf{c}_x$ is represented as a feature vector, where the specific features will be defined later. Our approach focuses on building a probabilistic model for bent-doubles: $p(\mathcal{C}) = p(\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3)$, the likelihood of the observed $\mathbf{c}_i$ under a bent-double model where we implicitly condition on "bent-double". Our general approach is to define this likelihood, then estimate its parameters from training data, and use it to rank candidate configurations.

### 3.1 Modeling Orientation

By looking at examples of bent-double galaxies and by talking to the scientists studying them, we have been able to establish a number of potentially useful characteristics of the components, the primary one being geometric symmetry. In bent-doubles, two of the components will look close to being mirror images of one another with respect to a line through the third component. We will call mirror-image components *lobe* components, and the other one the *core* component. It also appears that non-bent-doubles either don't exhibit such symmetry, or the angle formed at the core component is too straight — the configuration is not "bent" enough. Once the core component is identified, we can calculate symmetry-based features. However, identifying the most plausible core component requires either an additional algorithm or human expertise. In our approach we use a probabilistic framework that averages over different possible orientations weighted by their likelihood.

To formalize the estimation of the core and the lobes, consider the following. Without loss of generality assign the numbers $1, 2, 3$ to the components. In general we do not know which of 1, 2, or 3 is the core (under a bent-double assumption). By an **orientation** we mean a mapping of vertices to a set

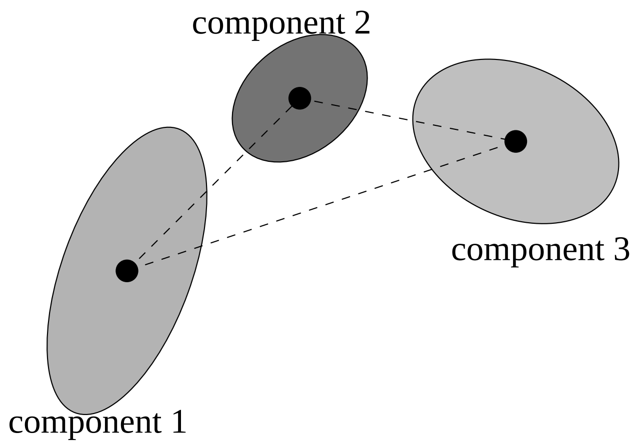


Figure 2: Elliptical components of a hypothetical bent-double. Assuming that label $a$ would correspond to a core component, a good choice of orientations would be $\{1 \to b, 2 \to a, 3 \to c\}$ or $\{1 \to c, 2 \to a, 3 \to b\}$.

of labels $\{a, b, c\}$ which preserves the neighbor relation in a cyclical order. Figure 2 shows an example of elliptical representation with possible orientations. For the set of three vertices, all 6 mappings preserve the neighbor relation. (In general, for configurations of $n$ components, there will be $2n$ such mappings.) The mapping from components $1, 2, 3$ to $a, b, c$ is defined by orientation $\theta_i$. We can then write

$$p(\mathbf{c}_a, \mathbf{c}_b, \mathbf{c}_c) = \sum_{i=1}^{6} p(\mathbf{c}_a, \mathbf{c}_b, \mathbf{c}_c | \theta_i) p(\theta_i), \quad (1)$$

i.e., a mixture over all possible orientations. Each orientation is assumed a priori to be equally likely, i.e., $p(\theta_i) = \frac{1}{6}$. Intuitively for a configuration that clearly looks like a bent-double, the terms in the mixture corresponding to the correct core component would dominate, while the other core interpretations would have much lower likelihood.

We represent each component $\mathbf{c}_x$ by three features. Note that the features can only be calculated conditioned on a particular mapping since they rely on properties of the (assumed) core and lobe components. Thus, conditioned on a particular mapping or orientation $\theta$, assuming label $x \in \{a, b, c\}$ where $a$,$b$,$c$ are defined in a cyclical order, the features are defined as:

- Angle $\alpha_{x,\theta}$—the angle formed at the center of the component (a vertex of the configuration) mapped to label $x$;

- Side ratios $sr_{x,\theta} = \frac{|\text{center of } x \text{ to center of } next(x)|}{|\text{center of } x \text{ to center of } prev(x)|}$;

- Intensity ratios $ir_{x,\theta} = \frac{\text{peak flux of } next(x)}{\text{peak flux of } prev(x)}$,

and so $\mathbf{c}_x|\theta = (\alpha_{x,\theta}, sr_{x,\theta}, ir_{x,\theta})$. Other features are also possible. Nonetheless this particular set appears to capture the more obvious visual properties of bent-doubles.

Rather than modeling the full joint distribution of all features, we make some approximating conditional independence assumptions (motivated by the relatively small amount of training data). In particular, we assume that

$$\begin{aligned} & P\left((\mathbf{c}_a, \mathbf{c}_b, \mathbf{c}_c)\,|\theta\right) \\ & \quad = \quad P\left(\alpha_{a,\theta}, \alpha_{b,\theta}, \alpha_{c,\theta}\right) P\left(sr_{a,\theta}, sr_{b,\theta}, sr_{c,\theta}\right) \\ & \quad \times \quad P\left(ir_{a,\theta}, ir_{b,\theta}, ir_{c,\theta}\right). \end{aligned}$$

For all ratio features $r$ (either of $sr, ir$), $r_{a,\theta} \cdot r_{b,\theta} \cdot r_{c,\theta} = 1$. For the angle features, $\alpha_{a,\theta} + \alpha_{b,\theta} + \alpha_{c,\theta} = \pi$. Assume that label $a$ corresponds to the choice of the core component. If we further assume conditional independence for the features of any two components we can obtain further simplifications:

$$\begin{aligned} & P\left(\alpha_{a,\theta}, \alpha_{b,\theta}, \alpha_{c,\theta}\right) \\ & \quad = \quad P\left(\alpha_{a,\theta}\right) P\left(\alpha_{b,\theta}|\alpha_{a,\theta}\right) P\left(\alpha_{c,\theta}|\alpha_{a,\theta}, \alpha_{b,\theta}\right) \\ & \quad = \quad P\left(\alpha_{a,\theta}\right) P\left(\alpha_{b,\theta}\right); \\ & P\left(r_{a,\theta}, r_{b,\theta}, r_{c,\theta}\right) \\ & \quad = \quad P\left(r_{a,\theta}\right) P\left(r_{b,\theta}|r_{a,\theta}\right) P\left(r_{c,\theta}|r_{a,\theta}, r_{b,\theta}\right) \\ & \quad = \quad P\left(r_{a,\theta}\right) P\left(r_{b,\theta}\right). \end{aligned}$$

### 3.2 Estimation of Probabilities

Given $\theta$, let $P_a(\alpha) = P(\alpha_{a,\theta})$, $P_a(r) = P(r_{a,\theta})$, and let $P_b(\alpha) = P(\alpha_{b,\theta})$, $P_b(r) = P(r_{b,\theta})$. If we know for every training example which component is the core (and is mapped to label $a$) we can then unambiguously estimate each of these distributions by using either parametric or non-parametric methods. We used two methods, fitting Gaussian probability distributions and using kernel density estimators (KDE), to calculate $P_a$ and $P_b$.

For the KDE method, we can estimate $P(x)$ (either $P_a$ or $P_b$) given a set $v_1, \ldots, v_k$ of appropriate values of features over the training set as

$$P(x|v_1, \ldots, v_k) = \frac{1}{k} \sum_{i=1}^{k} K(x, v_i, w)$$

where $K$ is the kernel function and $w$ is the width of the kernel. The choice of kernel for this problem is complicated since all of the used features have bounded domains — each $\alpha$ lies in $(0, \pi)$, and each ratio lies in $(0, +\infty)$. The domain for the ratios suggests using a log-normal kernel while the domain for angles suggests using a transformation-based kernel where the original domain is first mapped to $(-\infty, +\infty)$ [6]. More specifically, for ratio features

$$K_f(x, v_i, w) = \frac{1}{\sqrt{2\pi} w x} e^{-\frac{\left(\left(\log(v_i) + w^2\right) - \log(x)\right)^2}{2w^2}},$$

and for angles

$$K_f(x, v_i, w) = \frac{1}{C} e^{-\frac{(H(v_i) - H(x))^2}{2w^2}}$$

where $H(x) = \frac{1}{\beta} \log \frac{x}{\pi - x}$ with $\beta > 0$ and $C$ being a normalization constant computed by numerical integration. We experimented with other functional forms for the kernels and found that the above kernels produced better accuracy in our experiments.

For the parametric method, we fitted univariate normal (Gaussian) distributions to the transformed features. Since the domain for the Gaussian univariate distribution is $(-\infty, \infty)$, we first transformed the values of the features into $(-\infty, \infty)$. For ratio features, we used the log transformation while for angle features we used the transformation $H$ as described above. For each univariate Gaussian, both mean $\mu$ and variance $\sigma^2$ can be estimated from the transformed features derived from the training set.

### 3.3 Determining Core Components

In the previous subsection we described how to estimate the probabilities of individual features given a proper orientation or the core component (or equivalently, knowing the true identity of the core component). In practice, however, the identity of the core component is unknown.

We use our model to estimate which components are likely to be cores, using the following iterative scheme. Initially, core components for the bent-double examples in the training set are chosen at random. At each step of the iteration, we build the corresponding $P_a$ and $P_b$ distributions from the training set using the currently estimated orientations (and labels $a$). The estimated $P_a$ and $P_b$ distributions are then used on all of the examples in the training set to calculate the probability of each component being a core. This is done by summing $P\left(\mathbf{c}_a, \mathbf{c}_b, \mathbf{c}_c \middle| \theta_i\right)$ in Equation 1 over the 2 (out of 6 possible) orientations $\theta_i$ that map that component to label $a$. The most likely core components for each example are chosen to be the cores for the next iteration (in effect this is an approximation to a full expectation-maximization procedure, where the most likely core component is chosen rather than averaging over core components). The likelihood (probability of the training set under the currently estimated distributions) is recorded at each iteration. The algorithm stops either after a prespecified maximum number of iterations or when there are no changes from one iteration to the next.

This procedure yields estimates of the $P_a$ and $P_b$ distributions for each feature, allowing calculation of $P\left(\mathbf{c}_a, \mathbf{c}_b, \mathbf{c}_c \middle| \theta_i\right)$ for any particular orientation $\theta_i$. Thus, for a new unlabeled example we can now calculate a full likelihood $P\left(\mathbf{c}_a, \mathbf{c}_b, \mathbf{c}_c\right)$ (Equation 1), i.e. we average over all 6 possible orientations. For a set of unlabeled examples this yields a set of likelihood scores under the bent-double model, which can be sorted and thresholded to yield a receiver-operating characteristic. If the likelihood of the data under a non-bent-double model is assumed to be roughly uniform in feature-space, then these likelihoods will be roughly monotonically proportional to the posterior probability of a bent-double given the observed data. Here we choose not to build an explicit model of non-bent-doubles given that they can exhibit considerable variation, and instead rely on a model only of the positive examples for detection.

# 4 Experimental Results

For our experiments we use leave-one-out cross-validation, where for each of the 150 examples we build a model using the positive examples from the set of 149 "other" examples, and then score the original example with this model. The examples are then sorted in decreasing order by their likelihood score and analyzed using receiver operating characteristics (ROC curves). If the two classes can be perfectly separated by these scores, i.e. scores of all negative examples would appear after scores of all positive examples, then the curve would coincide with the left and upper sides of the $[0,1] \times [0,1]$ square. We use $A_{ROC}$, the area above the curve, as a measure of goodness of the model. A random score assignment would yield $A_{ROC} = 0.5$ while perfect assignment would have $A_{ROC} = 0$.

We experimented with both parametric and non-parametric estimators of the distributions $P_a$ and $P_b$. In a non-parametric setup, we used kernel density estimators (KDE) with a number of different choices of bandwidth. The results appear relatively insensitive to the particular bandwidths chosen. One set of bandwidths resulted in the plot shown in Figure 3. Alternatively, we tried estimating $P_a$ and $P_b$ with normal distributions on transformed features with one set of transformations resulting in the plot shown in Figure 3. From the plot we can infer, among other things, that the highest score for a negative example first appears after scores of 95 out of 128 positive examples (74%) for the KDE-based method and 103 out of 128 (80%) for the parametric method. Thus, the model appears to be quite accurate at detecting bent-double galaxies.

# 5 Conclusions

We proposed a probabilistic model for the identification of bent-double galaxies. A general mixture model framework allows for a principled and effective approach to orientation estimation. Experimental results based on cross-validation of likelihood scores under the model are accurate enough to suggest that the technique may be quite useful for automated identification of likely bent-double candidates from very large astronomy catalogs. In the future, we plan to choose an operating point and to compare this method with decision trees. We are also investigating whether bent-double configurations can be identified using unsupervised learning techniques [5].

# 6 Acknowledgements

This work was performed under a sub-contract from the ASCI Scientific Data Management Project of the Lawrence Livermore National Laboratory. The work of S. Kirshner and P. Smyth was supported by research grants from NSF (award IRI-9703120), Lawrence Livermore National Laboratories, the Jet Propulsion Laboratory, IBM Research, and Microsoft Research. I.

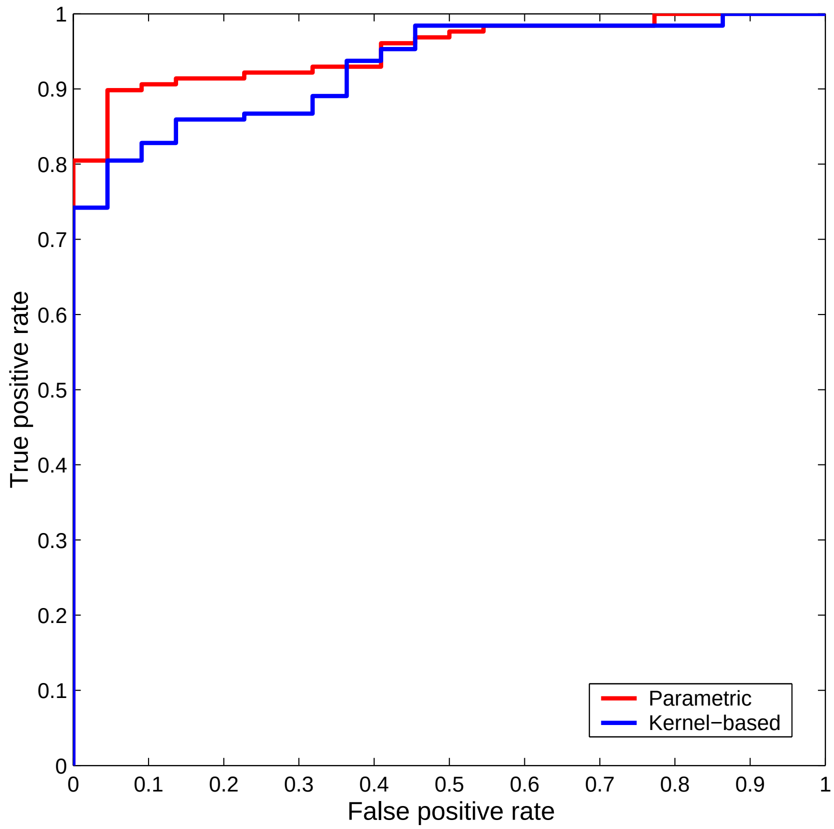


Figure 3: ROC curve plot for a model using angle, ratio of sides, and ratio of intensities, as features. For the parametric method, $A_{ROC} = 0.0469$. For the KDE-based method, $A_{ROC} = 0.0696$.

Cadez was supported by a Microsoft Graduate Fellowship. The work of C. Kamath and E. Cantú-Paz was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48. We gratefully acknowledge our FIRST collaborators, in particular, Robert H. Becker for sharing his expertise on the subject.

## References

[1] Robert H. Becker, Richard L. White, and David J. Helfand. The FIRST Survey: Faint Images of the Radio Sky at Twenty-cm. *Astrophysical Journal*, 450:559, 1995.

[2] Erick Cantú-Paz and Chandrika Kamath. Combining evolutionary algorithms with oblique decision trees to detect bent-double galaxies. In *Proceedings of International Symposium on Optical Science and Technology (SPIE Annual meeting)*, San Diego, July 30-August 4 2000.

[3] I. K. Fodor, E. Cantú-Paz, C. Kamath, and N. A. Tang. Finding bent-double radio galaxies: A case study in data mining. In *Proceedings of the Interface: Computer Science and Statistics Symposium*, volume 33, 2000.

[4] C. Kamath, E. Cantú-Paz, I. K. Fodor, and N. Tang. Searching for bent-double galaxies in the FIRST survey. In R. Grossman, C. Kamath, W. Kegelmeyer, V. Kumar, and R. Namburu, editors, *Data Mining for Scientific and Engineering Applications*, pages 95–114. Kluwer Academic Publishers, September 2001.

[5] Sergey Kirshner, Igov V. Cadez, Padhraic Smyth, and Chandrika Kamath. Learning to classify galaxy shapes using EM algorithm. In *Neural Information Processing Systems*, December 2002. (to appear).

[6] B. W. Silverman. *Density Estimation for Statistical and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, 1986.

[7] Richard L. White, Robert H. Becker, David J. Helfand, and Michael D. Gregg. A catalog of 1.4 GHz radio sources from the FIRST Survey. *Astrophysical Journal*, 475:479, 1997.