Astronomical Catalogue Matching as a Mixture Model problem

David Rohde*, Marcus Gallagher† and Michael Drinkwater**

*Instituto de Matemàtica, UFRJ, Rio de Janeiro, Brazil
†School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia
**School of Physics, University of Queensland, University of Queensland, Brisbane, Australia

Abstract. Astronomical telescopes increasingly operate in survey mode sweeping the sky systematically and producing highly processed data products such as astronomical catalogues which are lists of objects with positional information and other measurements usually including flux in a particular band. An important problem in electronic astronomy is the appropriate way to combine information from different catalogues produced by different telescopes. A key problem in combining this information is to establish different observations of the same object in the two catalogues i.e. the problem of catalogue matching. Positional information is not always sufficient in establishing matches reliably in these cases additional information from the non-positional measurements may also be used. This non-positional information is often scientifically interesting and its inter-catalogue properties may be the main object of study. In previous studies it is argued that while models of non-positional properties may assist in catalogue matching if these properties are scientifically interesting then the conclusions drawn from the analysis may be distorted by using this non-positional information. In this paper it is demonstrated that by employing a predictive Bayesian formalism it is possible to use all available information to assist in obtaining the most reliable matches and still obtain undistorted conclusions. Distortions are avoided because predictive distributions are computed where all the configurations of matches are marginalized over, rather than other approaches which choose a single most likely configuration of matches.

Keywords: MCMC, Statistical Astronomy
PACS: 02.50.Ng,02.50.Tt

INTRODUCTION

Increasingly telescopes operate in survey mode, sweeping the sky systematically and producing highly processed data products, one of the most highly processed of these data products is the astronomical catalogue which consists of a list of objects with two dimensional positions and measurements of flux and other attributes depending on the telescope such as colour, shape or redshift. An emerging problem in electronic astronomy is that of combining information from two or more of these catalogues together. An important statistical problem emerges in identifying different observations in each of the catalogues of the same object.

The most common approach is to use the position alone and match the closest objects together. While this approach is often very effective, there are important problems where this method is not satisfactory. Here we focus on problems where we would like to employ non-positional information in order to match more reliably. An illustration of the problem is given in Figure 1.

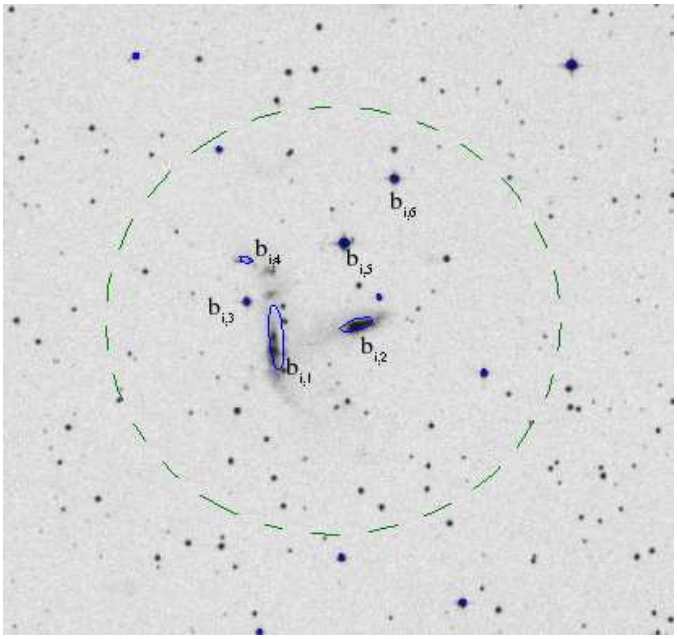

FIGURE 1. An example of a matching problem. An HI detection from the HIPASS catalogue is matched to an optical object in SuperCOSMOS. The object from the sparse HIPASS catalogue a_i is located at the centre of this image, the circle represents the 2σ limit of the positional uncertainty. There are a number of candidate optical counterparts from the denser SuperCOSMOS catalogue, $b_{i,1} \ldots b_{i,7}$, (circled).

Improving in sophistication upon finding the closest match another common approach in the astronomy literature is that of [9] which develops a probabilistic model, and suggests a histogram subtraction procedure for producing estimates. The probabilistic model is a useful framework for developing catalogue matching which has been used and extended here and elsewhere, but the histogram subtraction procedure has been found to give noisy unsatisfactory estimates in [3]. Supervised machine learning methods may also be an option which can exploit large number of catalogue features both positional and non-positional but are only applicable in the unusual case where a training set or subset of pre-matched examples are available [6] [7]. Bayesian approaches have also been suggested in [1] which focuses on identifying individual matches assuming knowledge of physical properties such as the spectral energy distribution which is not a probability distribution, but rather the physical properties of an astronomical object at every flux band. Here the authors suggest rigid a priori scientific knowledge could be used on the non-positional properties, however rigid scientific knowledge on non-positional properties may rarely be available, and in fact obtaining information about this may be the purpose in performing catalogue matching in the first place.

All of these approaches take as there goal establishing a list of the most likely matches as an end goal, however in previous work it has been demonstrated that this can lead to distorted analysis. In particular if only the most likely matches are obtained and an analysis of this follows not considering that this is but one possible configuration of matches that distorted conclusions will follow from the analysis. These problems

are particularly acute when positional information is insufficient to establish matches reliably and scientifically interesting quantities such as fluxes are used in the matching process. This distortion issue is extensively discussed in [7] and the issue is nicely captured by [1] when they state:

> Picking the correct combination of sources from various spatially similar configurations is a degenerate problem that requires extra information to resolve. The use of photometric (flux) information is a natural choice for its wide availability; however, its application requires further assumptions on the spectral energy distributions (SEDs). Often models exist to help out with the solution, but extra caution is needed to avoid any undesirable effect. For example, when the goal is to discover new types of objects with unknown SEDs, one should not apply known SEDs as priors but rather look for combinations that are likely matches based on spatial detections but excluded by SED modeling.

According to these authors if the spectral energy distribution is the object of study then this information must be ignored in the matching process. These authors follow a Bayesian approach where the probability of matches are obtained using a Bayesian framework, however they apparently advocate ignoring useful information even when it could in fact help in establishing correct matches.

The main contribution here is to demonstrate that it is possible to incorporate knowledge of non-positional information in particular the spectral energy distribution into the model without causing distortions. This is achieved by incorporating the matching state for each astronomical object a latent variable, that will be marginalized out. This is achieved by combining two ideas. Firstly we introduce a mixture model where matches and non-matches are mixtures (of different dimensions) and employ standard Gibbs sampling methods for the mixture model. Secondly we suggest that the astronomer adopt a predictive framework using current observations to make predictions of future observations, this means we are able to make the establishing of matches not an end in itself but an intermediate step and as such the matching state can be marginalized out. This allows Bayesian tools to use flexible prior information of scientific relevant quantities such as the spectral energy distribution and update this information by conditioning on unmatched or partially matched catalogues. This marks a key point of departure from [1], which does not update information about the spectral energy distribution, but may uses a rigid model.

The model introduced here is a fully Bayesian refinement of the model presented in [8]. Where they introduce a maximum likelihood approach using the Expectation Maximisation (EM) algorithm based on a histogram model, here a fully Bayesian approach using Gibbs sampling to a mixture based model. The model introduced here is also more general as inter-catalogue properties are also incorporated. The inclusion of inter-catalogue properties is particularly important as learning about these properties are often the main motivation for the whole matching process. In addition the inclusion of the inter-catalogue properties require a slight elaboration beyond standard mixture models. These enhancements allow us to show that matching can include all information without distortion in a fully Bayesian predictive approach. As noted this contrasts with previous approaches, including Bayesian approaches which have suggested in some circumstances it may be good to discard relevant information about the joint flux properties

(spectral energy distribution) as it may distort the matching process. The key contribution we make is to demonstrate that by changing the goal of the analysis from establishing correct matches to computing a predictive distribution, then in agreement with common sense all information can and should be used in the analysis.

MIXTURE MODEL FORMULATION OF MATCHING

Difficult catalogue matching problems usually consist of matching a catalogue with a relatively large number of objects per unit solid angle (the dense catalogue) and a relatively sparse catalogue (the sparse catalogue). This framework and terminology is adopted throughout this paper. The measurements of the ith element of the sparse catalogue is denoted α_i. In our studies this is a three dimensional vector including a two dimensional position right ascension $\alpha_{i_{\mathrm{RA}}}$ and declination $\alpha_{i_{\mathrm{Dec}}}$ and flux in a particular band α'_i. Within a reasonable positional distance of the ith sparse object there are N_i dense candidates that due to position alone are plausible candidates for a match. The jth of these N_i dense candidates near sparse object i is denoted $\beta_{i,j}$ again it is three dimensional with a two dimensional position again consisting of right ascension and declination and a flux in a different band the position these are denoted $\beta_{i,j_{\mathrm{RA}}}$, $\beta_{i,j_{\mathrm{Dec}}}$ and $\beta'_{i,j}$ respectively.

In the model the distribution of positional separation of RA and Dec is assumed to be a symmetrical bivariate normal distribution, the distribution on positional separation is independent of flux properties and also of other entries in the catalogue i.e. the covariance matrix of the distribution is assumed a priori, this can often be obtained from astronomical papers that present the characteristics of the telescope and its catalogue. In contrast the distribution over the inter-catalogue flux of the objects represents scientifically relevant knowledge about the spectral energy distribution that we assume is the main goal of this study as such the distribution of matching objects is given a semi-parametric Gaussian mixture model form. This results in

$$P(\alpha_{i_{\mathrm{RA}}} - \beta_{i,j_{\mathrm{RA}}}, \alpha_{i_{\mathrm{Dec}}} - \beta_{i,j_{\mathrm{Dec}}}, \alpha'_{i,j}, \beta'_{i,j} | Z_{i,j} = 1, \Theta_J) = \\ \mathrm{N}(\alpha_{i_{\mathrm{RA}}} - \beta_{i,j_{\mathrm{RA}}}, 0, \sigma^2)\mathrm{N}(\alpha_{i_{\mathrm{Dec}}} - \beta_{i,j_{\mathrm{Dec}}}, 0, \sigma^2) P_J(\alpha'_i, \beta'_{i,j} | \Theta_J)$$

where $Z_{i,j}$ is latent indicator variable which is one only if sparse object i matches dense object j and is zero otherwise and θ_J is the parameters of the semi-parametric density on the joint or matching catalogue properties, and similarly that θ_F are the semi-parametric density on the dense non-matching catalogue properties. While any semi-parametric, or non-parametric Bayesian model could be employed here we use here a Gaussian Mixture model, bivariate in the case of $P_J(\alpha'_i, \beta'_{i,j} | \Theta_J)$ and univariate in the case of $P_F(\beta'_{i,j} | \Theta_F)$ which will be used shortly. It follows that Θ_J and Θ_F are the parameters of a mixture of normal i.e. coefficients, means and covariances bivariate and univariate respectively.

We note that because strong prior knowledge on σ is often available this may be sufficient to establish likely configurations of matches and thereby informative inference about Θ_J may result be possible even if relatively vague priors are put on Θ_J. A uniform prior over the matches Z is usually appropriate such that all matches are equally likely

although on occasions there may be a subset of known matches available as in the dataset in [7], here we also assume that a subset of matches are available.

In order to formulate a complete likelihood for the model it is also necessary to specify the probability for non-matching objects. The model used here departs slightly from the standard mixture model framework in that while the matching objects have a distribution over both α and β the non-matching objects only have 'dense' properties β' and as such there is only a distribution over β' the dense flux. This can be handled in a fully Bayesian framework again by employing a semi-parametric model over β' i.e. $P(\beta'|\Theta_F)$.

These two expressions can be combined in order to produce a complete model specification for unmatched astronomical catalogues.

$$P(Z,\Theta_J,\Theta_F|D) \propto \prod_{i=1..M} \prod_{j=1..N_i} P_J(\alpha_i,\beta_{i,j}|\Theta_J)^{z_{i,j}} P_F(\beta_{i,j}|\Theta_F)^{1-z_{i,j}} P(\Theta_J)P(\Theta_F). \quad (1)$$

This fixed dimensional model consists of a mixture model of different dimensions for the sparse and dense objects, with slightly altered constraints on the indicator variables such that exactly one object in the sparse catalogue matches an object in the dense catalogue. In practice a fully Bayesian model for the flux distribution of the dense objects may not be necessary as the number of dense dataset is often very large and uncertainty about Θ_f can be neglected and maximum likelihood can be used to estimate Θ_F.

Positional information is always of high importance in matching. In this problem α and β will contain position information. It is reasonable for the $P(\alpha_i,\beta_{i,j}|\Theta_J)$ to factor such that $P_J(\alpha_i,\beta_{i,j}|\Theta_J) = \mathrm{Normal}\left(\begin{pmatrix}\alpha_{i_{\mathrm{RA}}}\\ \alpha_{i_{\mathrm{Dec}}}\end{pmatrix} - \begin{pmatrix}\beta_{i,j_{\mathrm{RA}}}\\ \beta_{i,j_{\mathrm{Dec}}}\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right) \mathrm{P}(\alpha'_{\mathrm{i}},\beta'_{\mathrm{i,j}}|\Theta_{\mathrm{J}})$ where α'_i and $\beta'_{i,j}$ are the parameters with positional information removed and the positional information for the sparse object is $\alpha_{i_{RA}},\alpha_{i_{Dec}}$ and for the dense catalogue object is $\beta_{i,j_{RA}},\beta_{i,j_{Dec}}$.

A Markov chain with a stationary distribution of the posterior can be simulated in the following way, here D is used to abbreviate all the data.

1. Sample from $P(Z|\Theta_J,\Theta_F,D)$.
2. Sample from $P(\Theta_J|D,Z)$.
3. Sample from $P(\Theta_F|D,Z)$.
4. Repeat from step 1.

Step 1, samples a possible configuration of matches and non-matches conditional on the parameters. It is very close to the standard procedure for sampling from latent indicator variables, but differs slightly because it is a mixture of distributions with different dimensions i.e. matching objects have a joint density of sparse and dense properties where non-matching objects only have dense properties. The slightly modified expression for sampling the matching or non-matching state is

$$P(z_{i,j}|\Theta_j,\Theta_f,\alpha_i,\beta_{i,1},...,\beta_{i,N_i}) \propto P(\alpha_i,\beta_{i,j}|\Theta_j)^{z_{i,j}} \prod_{k=1..N_i,k\neq j} P(\beta_{i,k}|\Theta_f)^{1-z_{i,k}}. \quad (2)$$

In our work a constrain on $P(Z)$ is that there is exactly one dense match for every sparse object.

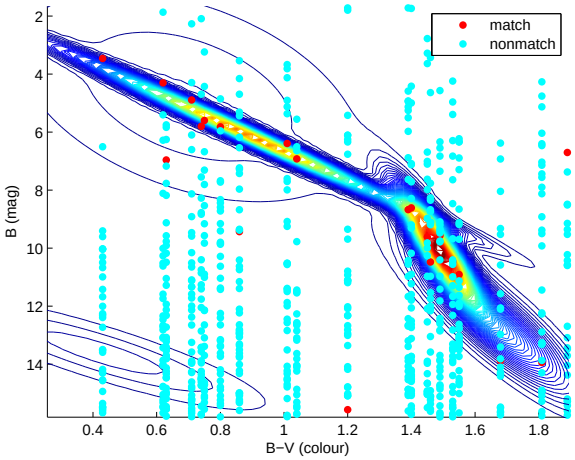

FIGURE 2. The unmatched data are overlaid over the estimated distribution. The vertical streaks are all the dense candidates to a sparse object. The 'true' match is marked in red – the other candidates are marked in blue. Often the true match will not be given the highest probability and a non-match will be detected instead.

Step 2 and Step 3 are standard methods for sampling from the posterior of a finite normal mixture model bivariate in the case of Step 2 and Step 3 which are extensively discussed in the literature e.g. see [4] or [5].

APPLICATION OF MODEL

Although simulations were also applied to a real dataset, the demonstration of a Bayesian predictive approach overcoming distortion when employing all available information is most easily achieved using a contrived matching problem. The true matches in the matching problem is based upon real dataset used for constructing the Hertzsprung-Russell (HR) diagram which has a distinctive shape. Although this is a real dataset [2] in practice it does not require matching to get the relevant measurements of the flux in different bands. We construct a partially matched dataset which consists of 2246 matches, the remaining 749 are considered in the unmatched portion. The flux of the background objects are given a uniform distribution between magnitudes 0 and 20. This is not physically motivated, but is chosen so as to make illustration of distortions from a naive analysis as clear as possible. For the unmatched portion the number of dense candidates is given a Poisson distribution with expectation 50.

Using the maximum likelihood fit of the joint distribution it is possible to determine the most likely matches and to see how these must be distorted to exaggerate the model see Fig. 2. These are shown with the pre-matched data in Fig. 3. A systematic distortion is evident where the most likely matches are less spread out than the original pre-matched dataset. For the purposes of this discussion let us consider the spread of objects perpendicular to the line shown in Fig. 3 . This is calculated by taking $V-(B-V)\times 5.0305-1.6285$ and discarding all objects with $B>9$, this is subsequently

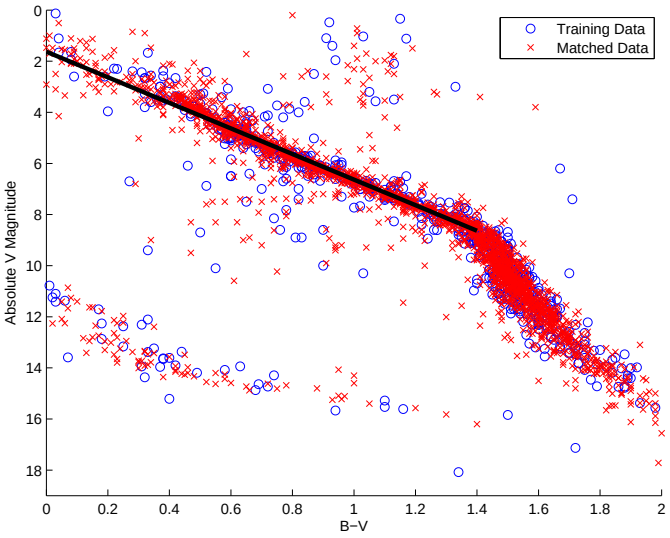

FIGURE 3. The pre-matched data (training set) and most likely matches.

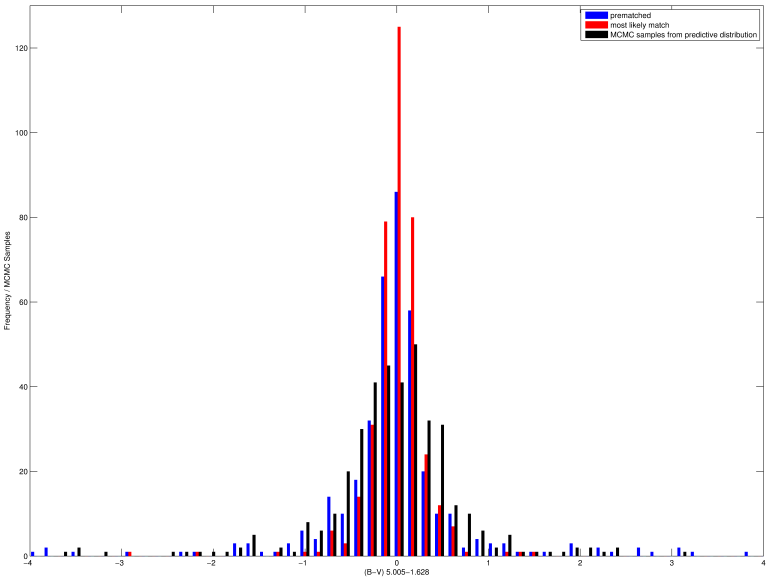

FIGURE 4. Illustration of distortion from using the most likely matches.

referred to as the line width.

A histogram of the spread of the original data, the most likely match and MCMC samples from the predictive distribution is shown in Fig. 4. It is visually obvious that the most likely matches are much more concentrated than the pre-matched data and the

predictive distribution is broader than both the pre-matched and most likely matches. This shows the Bayes-MCMC approach does not have the distortion effect seen in Fig. 3, the predictive distribution is not sharply peaked like the most likely matches.

The sample standard deviation of the line width of the original data is $\hat{\sigma}_{\text{pre}} = 1.2365$, however the standard deviation of the most likely match is much lower at $\hat{\sigma}_{\text{mlm}} = 0.5060$. This is considerably distorted because when errors are made in the most likely match they always happen in systematic ways. The predictive distribution on the other hand gives a standard deviation of $\hat{\sigma}_{\text{pred}} = 1.2552$ which is in reasonable agreement with the true matches. As the histogram of the line width shape was not Gaussian it was necessary to use a number of components to capture the shape.

Comparing the histogram with the predictive distribution we find better agreement with the most likely match although there remains some distance between the two. These difference might be seen as a good reason to consider alternatives to the Gaussian Mixture Model. The successful implementation of this technique requires workable use of Bayesian semi-parametric or non-parametrics of which Gaussian Mixture Models might be considered just one possibility.

The simulation was implemented in Matlab, the sampling of indicator variables required using interpreted Matlab this executed slowly and a single sample took up to 10 minutes to generate, although the speed of this was very dependent on the number of dense candidate objects. A large speed up could be achieved by rewriting this portion of the code in a compiled language, although we opted for long simulation times instead.

It is also interesting to consider how the model specification and prior specification impacts on the results. In general the impact of the priors of the mixture model i.e. the priors over the means and covariances seemed to have little qualitative impact on the outcome of the algorithm. On the other hand two elements that have a strong impact for related reasons are the prior on σ and the general difficulty of the problem, i.e. the number of dense candidates and the amount of the dataset that is pre-matched. It appears that a certain minimal amount of information is required for reasonable inferences to be drawn, if the amount of information is too small the algorithms wanders around a large posterior distribution.

The algorithm was also quite sensitive to the initial conditions of the Markov chain. The following relatively standard procedure was adapted to this context where the EM algorithm was applied to the subset of the data that was pre-matched, being applied to the matched or non-matched components individually in order to get initial estimates of Θ_J and Θ_F. Of course such a procedure is only possible when there is a pre-matched subset of the data available, in other situations other heuristics such as clustering algorithms or closest match algorithms may be employed, but as the difficulty of the problem will increase it is expected that much larger MCMC runs will be needed both for burn in and for averaging.

Judging convergence and discarding a suitable burn-in period is a challenging problem for MCMC methods particularly for mixture models which are known to have multiple separated modes due to symmetries. In general it is not realistic to expect that an MCMC algorithm applied to a mixture model will mix between all of these modes, but rather the algorithm should concentrate on the most important modes and mix between these. The symmetries that are responsible for making the posterior so complex and difficult become advantageous when computing a predictive distribution as adequate

local mixing can be sufficient to approximate the predictive distribution.

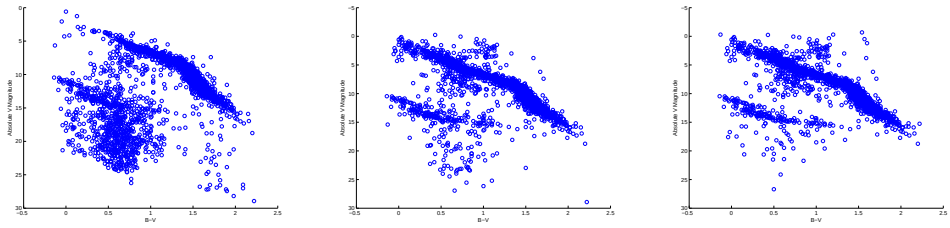

FIGURE 5. Samples of the matches at various points on the Markov chain at 10 iterations (left), 100 iterations (middle) and 150 iterations (right).

While there is a large literature on assessing convergence, including examining trace plots of a particular parameter, autocorrelation or formalized tests we found that a particularly useful and intuitive means of assessing convergence and algorithm behaviour is available in this context in the form of analyzing scatter plots of the matches at different steps of the algorithm. A graphical diagnosis of convergence can be seen in Figure 5 which shows the configuration of matches on the 10th, 100th and 150th iteration, note that Absolute V Magnitude is equal to negative log flux, such that large numbers are fainter than small numbers. The 10th iteration shows that there are many faint objects that are misclassified as matches, on the 100th iteration many more of the matches lie on the standard HR diagram after 150th objects a typical sample closely resembles the HR diagram and shows its distinctive features and there are very few faint objects classified as matches. Depending on the difficulty of the problem under investigation i.e. the number of dense candidates within a reasonable positional uncertainty and the proportion of the catalogue that is pre-matched then the behavior of the algorithm changes. If a lot of information is available then the samples 1-150 will be tail area of the posterior and the algorithm need only escape this configuration once and the algorithm will iterate locally between configurations of matches with very similar shape. On the other hand if less information is available then it may be that the algorithm will revisit these configurations after many iterations as under the model assumptions put forward these are plausible configurations. By varying the difficulty of the problem both of these cases where observed in our work. Although the first case might be the most satisfactory to astronomers, statistically both are interesting cases. The problem represented in Figure 5 is in fact a relatively easy problem where there are a large number of faint non-matching dense objects.

Every iteration of the Gibbs sampling algorithm produces graphical output of combined matches and non-matches, matches and non-matches separated and graphs of the predictive distribution of both the matches and non-matches. These were valuable for diagnosing convergence and understanding both the algorithm behaviour and the inference on the particular dataset. In particular it is interesting to consider how much the configuration of matches varies between iterations.

CONCLUSION

Catalogue matching was formulated as a mixture model problem in a predictive Bayesian framework. It was demonstrated that by operating in a predictive framework rather than establishing the most likely matches it is possible and desirable to include all information in the model. This contrasts with other approaches including other Bayesian approaches which focus on the goal of establishing the most likely matches. This is particularly evident in the example shown in Figure 4 where the line width estimated using an ideal most likely match algorithm is seriously distorted, but the predictive Bayes approach is undistorted. Alternative approaches encourage discarding useful information in order to avoid distorted interpretations of the most likely matches. It is demonstrated by example that these distortions are not evident in the predictive distribution.

ACKNOWLEDGMENTS

David Rohde was partially supported by a University of Queensland Confirmation Scholarship and the Programa Professor Visitante do Exterior from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior(CAPES).

REFERENCES

1. T. Budavari and A. S. Szalay. Probabilistic Cross-Identification of Astronomical Sources. *The Astrophysical Journal*, 679(1), July 2008.
2. W. Gliese and H. Jahreiss. Nearby Stars, Preliminary 3rd Version (Gliese+ 1991). *VizieR Online Data Catalog*, 5070, November 1995.
3. R. G. Mann, S. J. Oliver, S. B. G. Serjeant, M. Rowan-Robinson, A. Baker, N. Eaton, A. Efstathiou, P. Goldschmidt, et al. Observations of the Hubble Deep Field with the Infrared Space Observatory - IV. Association of sources with Hubble Deep Field galaxies. *Monthly Notices of the Royal Astronmical Society*, 289:482–489, August 1997.
4. J.M. Marin, K. Mengersen, and C.P Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics 25*, 2006.
5. G. J McLachlan and D. Peel. *Finite Mixture Models*. John Wiley and Sons Inc, 2000.
6. D. J. Rohde, M. J. Drinkwater, M. R. Gallagher, T. Downs, and M. T. Doyle. Applying machine learning to catalogue matching in astrophysics. *Monthly Notices of the Royal Astronomical Society*, 360(1):69–75, 2005.
7. D. J. Rohde, M. R. Gallagher, M. J. Drinkwater, and K. A. Pimbblet. Matching of catalogues by probabilistic pattern classification. *Monthly Notices of the Royal Astronmical Society*, 369:2–14, June 2006.
8. A. Storkey, C. Williams, E. Taylor, and B. Mann. An expectation maximisation algorithm for one-to-many record linkage, illustrated on the problem of matching far infra-red astronomical sources to optical counterparts. Technical Report EDI-INF-RR-0318, School of Informatics, University of Edinburugh, 2005.
9. W. Sutherland and W. Saunders. On the likelihood ratio for source identification. *Monthly Notices of the Royal Astronmical Society*, 259:413–420, December 1992.