Bayesian inference of T Tauri star properties using multi-wavelength survey photometry

Geert Barentsen[1,2⋆], J. S. Vink[1], J. E. Drew[2], S. E. Sale[3]
[1] *Armagh Observatory, College Hill, Armagh BT61 9DG, U.K.*
[2] *Centre for Astrophysics Research, Science and Technology Research Institute, University of Hertfordshire, Hatfield AL10 9AB, U.K.*
[3] *Rudolf Peierls Centre for Theoretical Physics, Keble Road, Oxford OX1 3NP, U.K.*

Received 2012 April 15; Accepted 2012 November 21.

ABSTRACT

There are many pertinent open issues in the area of star and planet formation. Large statistical samples of young stars across star-forming regions are needed to trigger a breakthrough in our understanding, but most optical studies are based on a wide variety of spectrographs and analysis methods, which introduces large biases.

Here we show how graphical Bayesian networks can be employed to construct a hierarchical probabilistic model which allows pre-main sequence ages, masses, accretion rates, and extinctions to be estimated using two widely available photometric survey databases (IPHAS r'/Hα/i' and 2MASS J-band magnitudes). Because our approach does not rely on spectroscopy, it can easily be applied to homogeneously study the large number of clusters for which Gaia will yield membership lists.

We explain how the analysis is carried out using the Markov Chain Monte Carlo (MCMC) method and provide Python source code. We then demonstrate its use on 587 known low-mass members of the star-forming region NGC 2264 (Cone Nebula), arriving at a median age of 3.0 Myr, an accretion fraction of $20 \pm 2\%$ and a median accretion rate of $10^{-8.4}\,\mathrm{M}_\odot/\mathrm{yr}$.

The Bayesian analysis formulated in this work delivers results which are in agreement with spectroscopic studies already in the literature, but achieves this with great efficiency by depending only on photometry. It is a significant step forward from previous photometric studies, because the probabilistic approach ensures that nuisance parameters, such as extinction and distance, are fully included in the analysis with a clear picture on any degeneracies.

Key words: stars: pre-main sequence, methods: data analysis, astronomical data bases: surveys, accretion, open clusters and associations: individual: NGC 2264

1 INTRODUCTION

Large uncertainties remain with respect to the mechanisms and timescales of star and planet formation. While it has been established that young solar-like stars stop accreting and lose their protoplanetary discs on a timescale of ∼1 to 10 Myr (Haisch et al. 2001; Fedele et al. 2010), there is no full understanding of the interplay between the various physical mechanisms which affect disc evolution (Williams & Cieza 2011).

Large statistical samples of young stars are needed to make the breakthrough in our understanding. Whilst such samples have recently become available through infrared photometry which traces circumstellar *dust* (Evans et al. 2009), they are not available for emission-line studies which traces material in the *gas* phase. Understanding the evolution of the gas with respect to the dust is of critical importance for testing competing models of disc evolution and planet formation (e.g. Najita et al. 2007; Owen et al. 2011; Espaillat et al. 2012).

Existing gas emission-line studies have mostly relied on spectroscopy, which could only be obtained for limited numbers of stars in nearby star-forming regions (e.g. Gullbring et al. 1998; Natta et al. 2004; Herczeg & Hillenbrand 2008). Moreover, it is often hard to inter-compare the results from different regions, because they have been obtained using a variety of spectrographs and analysis methods which complicate the analysis.

The increasing availability of data from large photometric surveys allows samples to be obtained across star-forming

⋆ E-mail: geert@barentsen.be

regions which are both larger and more homogeneous. For example, the INT Photometric Hα Survey (IPHAS) covers 1800 deg^2 of the Northern Galactic Plane using r'/i' broad-band and Hα narrow-band filters (Drew et al. 2005; González-Solares et al. 2008). This survey is particularly relevant to star formation studies, because Hα photometry allows a statistical appraisal of gas accretion rates to be made for massive clusters at large distances (De Marchi et al. 2010; Spezzi et al. 2012).

So far, we have used the IPHAS survey to study pre-main sequence stars in just one star-forming region: IC 1396 in Barentsen et al. (2011), hereafter Paper I. We used Hα narrow-band photometry to identify T Tauri stars and estimate their accretion rates, whilst simultaneously estimating ages and masses from the (r'-i')/r' colour-magnitude plane.

Before we can apply our methodology to the entire IPHAS (and future VPHAS) Galactic Plane region, we first need to develop more powerful tools. Whilst the results of Paper I were in good agreement with independent spectroscopic measurements, the method suffered from the drawback that a fixed amount of extinction was assumed for all objects, as there is no straightforward method to obtain this parameter simultaneously with ages, masses and accretion rates from colour-magnitude or colour-colour diagrams. The increasing sophistication of models and the so-called "data deluge" from surveys requires more powerful inference tools to be adopted, as there are limitations to the information content to be extracted from two-dimensional diagrams.

The generic mathematical solution to the problem of understanding which parameter-space regions match a set of observations is called *Bayesian inference*. The theoretical principles of the method have been understood for decades, but the widespread adoption in astrophysics has only taken off in recent years owing to the advances in both algorithms and computing power. So far, the Bayesian framework has become the favoured tool for e.g. the determination of cosmological parameters (Trotta 2008), the analysis of transit light curves (Ford 2005; Kipping et al. 2012) or the determination of meteor rates (Barentsen et al. 2011b). The approach has also been explored for estimating ages and extinctions for main-sequence stars (Pont & Eyer 2004; Jørgensen & Lindegren 2005; Bailer-Jones 2011). In the context of star formation, the method has recently been used to perform a dynamical membership analysis of the Sco OB2 association (Rizzuto et al. 2011) and to assess the accuracy of pre-main sequence models (Gennaro et al. 2012). However, the method has so far not been used to tackle the common problem of estimating the basic parameters of individual pre-main sequence stars.

In this paper we show how Bayesian inference can be used to simultaneously determine extinction, stellar ages & masses and accretion rates for known members of a star-forming region. We will demonstrate the method on NGC 2264, which is one of the best-studied regions within the IPHAS survey area and has a very complete membership list (Dahm 2008; Sung et al. 2008). We note that the future Gaia astrometric survey is expected to yield accurate membership lists for hundreds of clusters (Bailer-Jones 2009), therefore our strategy to re-analyse a large sample of known cluster members in a homogeneous way is likely to become an increasingly important tool.

In §2 we motivate our approach and specify the model. In §3 we explain the application to NGC 2264 and in §4 we present the results. In §5-§7 we discuss the outcome, present future extensions and summarise the conclusions.

2 METHOD: BAYESIAN INFERENCE

Our aim is to determine ages, masses and accretion rates from IPHAS r'/i'/Hα magnitudes, while simultaneously constraining the extinction by adding 2MASS J-band magnitudes to the dataset. We employ Bayesian Inference for this purpose. In §2.1 we describe the motivation for the method, in §2.2-2.3 we explain the formalisms and implementation, while in §2.4-2.5 we explain the practical use.

2.1 Motivation

2.1.1 Characteristics of T Tauri stars

In the current picture of star formation, solar-like stars are thought to assemble the majority of their mass during the first few 10^5 years after the initial collapse of their parent molecular cloud (Evans et al. 2009). Within ∼1 Myr the envelope of gas and dust clears and the newly formed stars become visible at optical wavelengths, from which point they are commonly called *T Tauri* stars. These objects continue to grow by accreting material from a circumstellar accretion disc, which does not exceed 10-20% of the stellar mass and is dispersed within a few million years (Hartmann 2008; Haisch et al. 2001; Fedele et al. 2010).

Mass accretion is thought to take place along magnetic field lines which connect the disc to the star. Infalling gas is essentially on a ballistic trajectory, falling on to the stellar surface at near free-fall velocities, thereby producing hot impact shocks which generate excess UV and optical continuum emission (Calvet & Gullbring 1998; Gullbring et al. 2000). The accretion energy released in these shocks also heats the infalling gas, which in turn produces strong Hα emission.

This is illustrated in Fig. 1, where we show literature spectra of four T Tauri stars of a similar spectral type (late K), shown in order of accretion rates which have been estimated from the UV/optical continuum excess emission. The spectra illustrate that the Hα line strength is correlated with the blue excess, both thought to be a result of the release of accretion energy. In contrast, the r' and i' bands appear least affected by accretion and are thus the most appropriate tracers for stellar age and mass. Hence, in our past study of IC 1396 (Paper I) we used the (r'-i')/r' plane to estimate ages and masses from model isochrones, whilst using the (r'-i')/(r'-Hα) plane to estimate Hα equivalent widths and accretion rates.

We note that at exceptionally high accretion rates ($\gtrsim 10^{-6}$ M$_\odot$/yr) there is evidence for accretion-induced continuum veiling to occur in the r' and i' bands, which may affect the age and mass estimates. The source of the excess emission at these wavelengths is not well understood at present (Fischer et al. 2011). The effect is small and will be ignored in what follows, because our study includes only objects with lower accretion rates. We will return to this topic at the end of the paper however (§5.3).

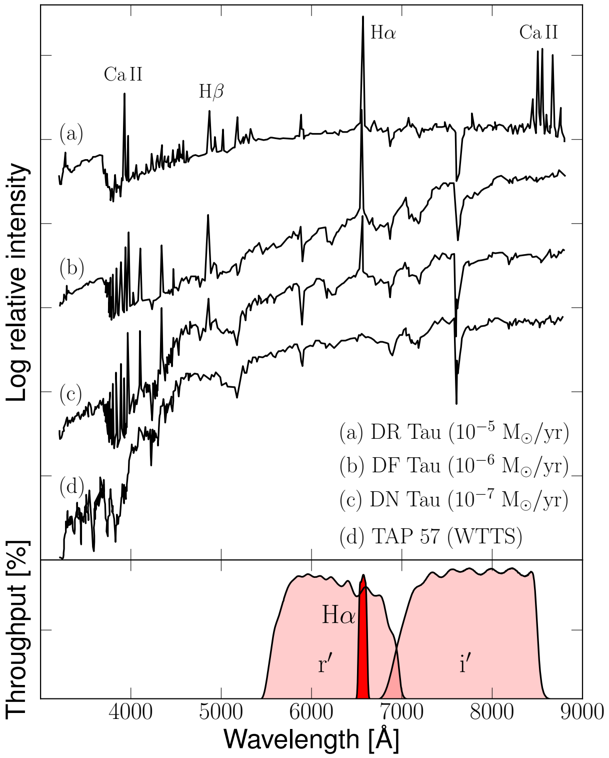

Figure 1. Top panel: spectra of four T Tauri stars of similar spectral type (late K) as published by Bertout (1989) shown in order of decreasing accretion rates as determined by Hartigan et al. (1995). The most notable features are (i) blue continuum excess and line veiling below ∼6000 Å; (ii) hydrogen Balmer lines in emission; (iii) Ca II H/K and Infrared Triplet (IRT) lines in emission. The bottom spectrum is a weak-lined T Tauri star (WTTS) which does not show signs of ongoing mass accretion. Bottom panel: filter transmission curves for the IPHAS r'/Hα/i' filters. The 2MASS J-band filter is located beyond the range of the spectra near 12 000 Å.

2.1.2 Solving the degeneracy between age, mass, extinction and Hα emission

The results obtained in Paper I were in good agreement with spectroscopic measurements from the literature. However, the method suffered from the drawback that a fixed amount of extinction was assumed for all objects, because optical colour-colour diagrams show a well-known degeneracy between reddened high-mass (early-type) stars and unreddened low-mass (late-type) stars. This may surprise some readers, because the (r'-i')/(r'-Hα) plane is known for being able to break this degeneracy due to the strong difference in the Hα line strength between early- and late-type stars (Drew et al. 2008). However, this property cannot be exploited when Hα is in emission.

The degeneracy is illustrated in Fig. 2, where we plot the colour simulations from Paper I to show the intrinsic position of the main sequence as a function of reddening (thick and thin solid lines). These solid curves do not overlap or intersect, i.e. there is a handle on the degeneracy. However, when increasing levels of Hα emission are added (dashed lines), reddened early-type stars with Hα in emission become entangled with unreddened late-type stars.

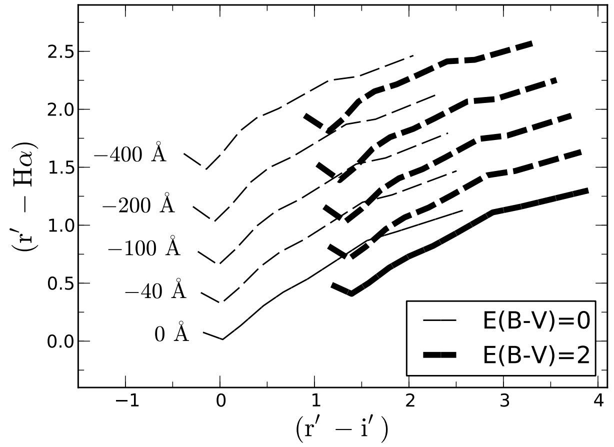

Figure 2. Position of the unreddened and reddened main sequence (thin and thick lines) in the IPHAS (r'-i')/(r'-Hα) plane, shown from early O to late M-type stars. We also mark the position of the main sequence for increasing levels of Hα emission (dashed lines, indicated by their Hα EWs in units Å). The figure illustrates the degeneracy between reddening, spectral type and Hα emission in this plane. These simulations are taken from Paper I.

To resolve this degeneracy, we require additional colour(s) to be added to the IPHAS dataset. Our best-available option is to add the near-infrared J-band magnitude from the 2MASS survey (Skrutskie et al. 2006), because the (r'-i')/(i'-J) plane can be shown to break the degeneracy between reddening and spectral type for K- and M-type objects, which constitute the vast majority of T Tauri stars (Martin 1997). At the same time, the J-band is not commonly affected by excess emission from a circumstellar disc, unlike the 2MASS H- and K-bands at longer wavelengths (Meyer et al. 1997).

However, adding the J-band magnitude to our sample does not allow us to simply read the extinction for every object from the (r'-i')/(i'-J) diagram straight away, as these colours do not depend on mass and extinction alone. For example, the Hα-line falls inside the r'-band, such that objects with Hα in emission show an r'-band excess ranging between -0.05 mag (EW$_{\mathrm{H}\alpha}\cong$ -50 Å) and -1.0 mag (EW$_{\mathrm{H}\alpha}\cong$ -1000 Å), albeit depending on the spectral type (Paper I). In addition, the colours also depend on the stellar age as pre-main sequence stars tend to rise in effective temperature as they approach the main sequence.

Whilst adding the J-band data should offer us sufficient information to constrain the four parameters of interest, the problem remains how these constraints can be inferred in practice?

2.1.3 Inferring parameters from photometry

The traditional method used to estimate stellar parameters from photometry is to place the observed objects in two-dimensional colour/magnitude diagrams, together with the output of evolutionary models. Such an approach is useful when the number of free parameters is small. However, when the number of dimensions in the parameter or observable space increases, there is no obvious way to decide which plane/parameter-combinations should be employed. For example, our r'/Hα/i'/J dataset can be used to construct no less than 15 different colour-colour diagrams and 24 colour-magnitude diagrams, all of which depend to some extent on all of our parameters of interest.

We could devise an ad-hoc algorithm to make use of all the available information, for example: we could iteratively fit model parameters in multiple planes until a certain convergence criterion is met. We can also employ a range of data modelling methods to help us link observations to parameters (e.g. *Neural Networks*, *Principal Component Analysis*). However, it is unclear how the results obtained by such methods depend on ad-hoc choices made in the design of the algorithms (e.g. the structure of the Neural Network). Moreover, there is no clear way for these methods to obtain meaningful error bars which take full account of all known sources of uncertainty (Ford 2005).

A better approach is to obtain *maximum likelihood* estimates using *expectation-maximisation* (EM) algorithms. These methods require the user to define a predictive model which estimates the likelihood of an observation given a set of model parameters. The model is then optimised to find the set of parameters which are *most likely* to explain the observed data (in the special case of a Gaussian likelihood model this is χ^2*-fitting*).

While EM presents a significant improvement over ad-hoc methods, it suffers from the drawback that only a single "best-fit" point estimate is provided, regardless of whether or not a unique solution exists. This is a particular concern in astronomy, where model uncertainties and data sparsity imply that there is often a "family" of likely solutions which occupy degenerate or multi-modal regions in the parameter space. This problem is often solved by keeping one or more of the degenerate "nuisance" parameters fixed, but such assumptions invariably reduce the ability of the model to capture reality.

Rather than focusing on finding a "best-fit" estimate, it is better to employ the predictive model to infer the full probability distribution for all possible solutions. This approach is called *Bayesian inference* (though we note that some authors prefer the term *probabilistic inference* to distinguish the approach from best-fit EM methods which also employ the Bayes' theorem.)

Obtaining the full distribution may seem impractical at first sight, because point estimates are useful for plotting and tabulation. However, a full distribution can be reduced to a point estimate by computing the expectation value or median, which, unlike the maximum likelihood, take full account of the distribution (that is, an expectation value minimises the variance, while the median minimises the mean absolute error.) Moreover, knowledge of the full distribution allows meaningful confidence intervals and covariances to be quantified and visualised, by marginalising over the nuisance parameters.

In what follows we provide a formal description of the method and explain how it is applied to our problem.

2.2 Solution: Bayesian inference

2.2.1 Formalism

The basic idea is to create a parameterised model which is able to reproduce the data and its uncertainty, and then compare that model for different sets of parameters against the observations in a probabilistic way.

Let $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ represent a set of unknown model parameters and let $\boldsymbol{d} = \{d_1, \ldots, d_i\}$ represent a set of observed data. We can construct a *likelihood model* $P(\boldsymbol{d}|\boldsymbol{\theta})$ which computes the probability for an observation to occur under a given a set of parameters. Such a model can be computed using the best-available knowledge, and is limited only by scientific complexity and computing power.

Of course our aim is not to understand which observations are expected given the parameters, but inversely, *which of the various possible sets of parameters best explain a given observation.* This is expressed by the function $P(\boldsymbol{\theta}|\boldsymbol{d})$, called the *posterior* distribution, which can be linked to the likelihood model using the *chain rule* from probability theory:

$$P(\boldsymbol{\theta}|\boldsymbol{d}) \cdot P(\boldsymbol{d}) = P(\boldsymbol{\theta}, \boldsymbol{d}) = P(\boldsymbol{d}|\boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}), \quad (1)$$

which leads to the well-known theorem by Bayes:

$$P(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{\theta}) \cdot P(\boldsymbol{\theta})}{P(\boldsymbol{d})}, \quad (2)$$

where $P(\boldsymbol{\theta})$ is called the *prior*, which encodes any a priori knowledge about the parameters which we wish to include in our model (including the allowed physical bounds). The denominator, $P(\boldsymbol{d})$, may be thought of as a normalising constant which does not effect the shape of the posterior distribution and can be ignored, i.e.:

$$P(\boldsymbol{\theta}|\boldsymbol{d}) \propto P(\boldsymbol{d}|\boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}), \quad (3)$$

or simply:

$$P(\boldsymbol{\theta}|\boldsymbol{d}) \propto P(\boldsymbol{\theta}, \boldsymbol{d}). \quad (4)$$

Thus, the key to identify the regions in the parameter space which explain a set of observations is the ability to compute the joint probability distribution $P(\boldsymbol{\theta}, \boldsymbol{d})$.

2.2.2 Constructing the joint distribution

$P(\boldsymbol{\theta}, \boldsymbol{d})$ can be thought of as the model which defines how the theoretical parameters and observations relate to each other. We now explain how this model is formulated for our application.

First, let us define the set of unknown variables $\boldsymbol{\theta}$ and the set of observables $\boldsymbol{d}$. The free parameters of principal interest in our work are mass (M_*), age (τ), mass accretion rate ($\dot{\mathrm{M}}_{\mathrm{acc}}$) and extinction ($\mathrm{A}_0$). For syntactic convenience, we add a set of additional variables which help us formulate the model, such as a star's intrinsic spectral energy distribution ($\mathrm{SED}_{\mathrm{int}}$). Their meaning is explained in Table 1:

$$\boldsymbol{\theta} = \{\mathrm{M}_*,\ \tau,\ \dot{\mathrm{M}}_{\mathrm{acc}},\ \mathrm{A}_0,\ \mathrm{SED}_{\mathrm{int}},\ \mathrm{R}_{\mathrm{in}},\ \mathrm{L}_{\mathrm{H}\alpha}, \quad (5)$$

$$\mathrm{EW}_{\mathrm{H}\alpha},\ d,\ \mathrm{SED}_{\mathrm{app}}\}. \quad (6)$$

We note that the additional variables are either determined by the four free parameters of interest, or they are so-called *nuisance parameters* which will be constrained by a strong prior assumption (e.g. in what follows the distance d will be constrained by a strong prior based on literature estimates).

M_*	Stellar mass ($\mathrm{M}_\odot$)
τ	Stellar age (Myr)
$\dot{\mathrm{M}}_{\mathrm{acc}}$	Accretion rate ($\mathrm{M}_\odot$ yr^{-1})
R_{in}	Inner disc truncation radius (R_*)
$\mathrm{SED}_{\mathrm{int}}$	Modelled intrinsic SED $\{\mathrm{M}_{\mathrm{r'}},\ \mathrm{M}_{\mathrm{H}\alpha},\ \mathrm{M}_{\mathrm{i'}},\ \mathrm{M}_{\mathrm{J}}\}$
$\mathrm{L}_{\mathrm{H}\alpha}$	Excess Hα luminosity ($\mathrm{L}_\odot$)
$\mathrm{EW}_{\mathrm{H}\alpha}$	Hα emission Equivalent Width (Å)
d	Distance (pc)
A_0	Extinction parameter (mag)
$\mathrm{SED}_{\mathrm{app}}$	Modelled apparent SED $\{\mathrm{m}_{\mathrm{r'}},\ \mathrm{m}_{\mathrm{H}\alpha},\ \mathrm{m}_{\mathrm{i'}},\ \mathrm{m}_{\mathrm{J}}\}$
$\mathrm{SED}_{\mathrm{obs}}$	Observed SED $\{$r', Hα, i', J, $\sigma_{\mathrm{r'}}$, $\sigma_{\mathrm{H}\alpha}$, $\sigma_{\mathrm{i'}}$, $\sigma_{\mathrm{J}}\}$

Table 1. Notation.

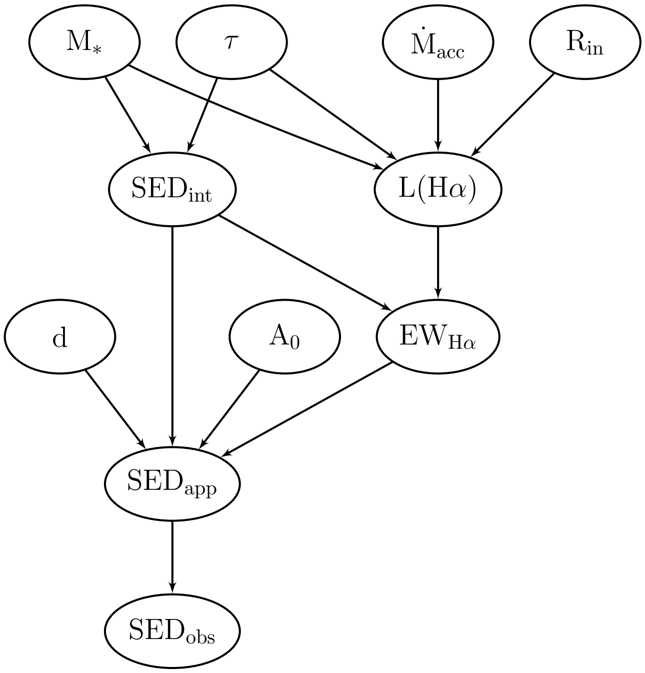
M_*
τ
$\dot{\mathrm{M}}_{\mathrm{acc}}$
R_{in}
$\mathrm{SED}_{\mathrm{int}}$
$\mathrm{L(H}\alpha)$
d
A_0
$\mathrm{EW}_{\mathrm{H}\alpha}$
$\mathrm{SED}_{\mathrm{app}}$
$\mathrm{SED}_{\mathrm{obs}}$

Figure 3. Bayesian network representing the dependencies between the variables in our inference model. Nodes which are not directly connected represent variables which are conditionally independent of each other given their parents.

For the observed data of a star, we adopt the SED as characterised by the four apparent magnitudes and their uncertainties:

$$\boldsymbol{d} = \mathrm{SED}_{\mathrm{obs}} = \{\mathrm{r'},\ \mathrm{H}\alpha,\ \mathrm{i'},\ \mathrm{J},\ \sigma_{\mathrm{r'}},\ \sigma_{\mathrm{H}\alpha},\ \sigma_{\mathrm{i'}},\ \sigma_{\mathrm{J}}\}. \tag{7}$$

Having defined $\boldsymbol{\theta}$ and $\boldsymbol{d}$, we now formulate $P(\boldsymbol{\theta}, \boldsymbol{d})$. This is a complex distribution with a high number of dimensions. It would be tedious to construct a single function which computes its value for all possible combinations of $\boldsymbol{\theta}$ and $\boldsymbol{d}$. We can greatly reduce the complexity however by exploiting the fact that many of the variables can be assumed to have (conditional) independence relationships.

A convenient and concise way to represent the dependencies amongst variables is to use a *directed acyclic graph* (i.e. a graph with no loops), often referred to as a *probabilistic graphical model* or a *Bayesian network*. These terms are synonyms for a syntactic method whereby variables are represented by nodes in a graph, and dependencies between those variables are represented by arrows. Intuitively, an arrow from variable A to B indicates that A has a *direct influence* on B. Formally, the Bayesian Network is constructed as follows: (Russell & Norvig 2009)

(i) each node corresponds to a variable from $\boldsymbol{\theta}$ or $\boldsymbol{d}$;

(ii) a set of arrows connect pairs of nodes. If there is an arrow from node A to node B, then A is said to be a *parent* of B. Pairs of nodes which are *not* directly connected by an arrow represent variables which are assumed to be conditionally independent of each other given their parents;

(iii) the graph should not have directed cycles, this would be a sign of recursive reasoning.

The Bayesian network for our model is shown in Fig. 3. For example, the graph shows arrows pointing from M_* and τ to the intrinsic SED, which represents our assumption that we only need to know M_* and τ to infer $\mathrm{SED}_{\mathrm{int}}$, i.e.:

$$P(\mathrm{SED}_{\mathrm{int}} \mid \boldsymbol{\theta}, \boldsymbol{d}) = P(\mathrm{SED}_{\mathrm{int}} \mid \mathrm{M}_*, \tau) \tag{8}$$

We may use this property to simplify the formulation of $P(\boldsymbol{\theta}, \boldsymbol{d})$ as follows. Let $(n_1, \ldots, n_k)$ be the set of all variables. The chain rule allows us to write its joint distribution as a product of conditional probabilities:

$$P(n_1, \ldots, n_k) = \prod_{i=1}^{k} P(n_i \mid n_{i+1}, \ldots, n_k). \tag{9}$$

Given the specification of a Bayesian network over all nodes n_i, Eqn. 9 may be simplified using the property of conditional independence:

$$P(n_1, \ldots, n_k) = \prod_{i=1}^{k} P(n_i \mid \mathit{Parents}(n_i)), \tag{10}$$

where $\mathit{Parents}(n_i)$ is the set of nodes which have an arrow pointing to n_i in the network, and $\mathit{Parents}(n_i) \subseteq \{n_{i+1}, \ldots, n_k\}$. The last condition can always be satisfied by pre-ordering the nodes in a way that is consistent with the partial order given by the network.

$P(\boldsymbol{\theta}, \boldsymbol{d})$ can now be written as the product of a series of lower-dimensional distributions:

$$\begin{aligned}
P(\boldsymbol{\theta}, \boldsymbol{d}) &= \prod_{i=1}^{k} P(n_i \mid \mathit{Parents}(n_i)) && (11)\\
&= P(\mathrm{M}_*) \cdot P(\tau) \cdot P(\dot{\mathrm{M}}_{\mathrm{acc}}) && (12)\\
&\quad \cdot P(\mathrm{R}_{\mathrm{in}}) \cdot P(d) \cdot P(\mathrm{A}_0) && (13)\\
&\quad \cdot P(\mathrm{SED}_{\mathrm{int}} \mid \mathrm{M}_*, \tau) && (14)\\
&\quad \cdot P(\mathrm{L}_{\mathrm{H}\alpha} \mid \mathrm{M}_*, \tau, \dot{\mathrm{M}}_{\mathrm{acc}}, \mathrm{R}_{\mathrm{in}}) && (15)\\
&\quad \cdot P(\mathrm{EW}_{\mathrm{H}\alpha} \mid \mathrm{L}_{\mathrm{H}\alpha}, \mathrm{SED}_{\mathrm{int}}) && (16)\\
&\quad \cdot P(\mathrm{SED}_{\mathrm{app}} \mid \mathrm{SED}_{\mathrm{int}}, \mathrm{EW}_{\mathrm{H}\alpha}, d, \mathrm{A}_0) && (17)\\
&\quad \cdot P(\mathrm{SED}_{\mathrm{obs}} \mid \mathrm{SED}_{\mathrm{app}}) && (18)
\end{aligned}$$

It is trivial to see that rewriting the equation in this way makes it far easier to define our probabilistic model, which is now a combination of small hierarchical “sub-models”. The models for parameters which do not have parents are called the *priors* (Eqns. 12-13), whilst the others are called the *likelihoods* (Eqns. 14-18). In what follows we explain how each of the above factors are computed, which acts as a formal specification of our parameter inference model.

Prior	Distribution	
$P(M_*)$	$\sim \begin{cases} {M_*}^{-1.3} & \text{if } 0.1 < M_* < 0.5 \\ 0.5\,{M_*}^{-2.3} & \text{if } 0.5 \leqslant M_* < 7 \end{cases}$	
$P(\log \tau)$	$\sim \mathcal{U}(5, 8)$	
$P(\log \dot{M}_{acc})$	$\sim \mathcal{U}(-15, -2)$	
$P(R_{in})$	$\sim \mathcal{N}(5, \sigma = 2)$	$(R_{in} > 1)$
$P(d)$	$\sim \mathcal{N}(760, \sigma = 5)$	
$P(\log A_0)$	$\sim \mathcal{N}(-0.27, \sigma = 0.46)$	

Table 2. Summary of priors. $\mathcal{U}(min, max)$ denotes a Uniform distribution, while $\mathcal{N}(\mu, \sigma)$ denotes a Gaussian.

2.3 Priors and Likelihoods

2.3.1 Priors

A challenge that comes with the use of Bayes' theorem is the choice of the priors. We should not shy away from this task however, as there is no inference without assumptions. Being forced to define them in a clear way can be seen as an advantage over other methods (e.g. χ^2-fitting bears the implicit assumption of uniform priors on the free parameters, yet the consequence of this assumption is rarely considered.) In the case of our most influential prior, $P(A_0)$, we will quantify the extent of its influence in §5.4 at the end of the paper.

The prior distributions are summarised in Table 2 and explained as follows:

- $P(M_*)$: the mass prior follows the Initial Mass Function (IMF) due to Kroupa (2001), truncated between 0.1 and $7\,M_\odot$ which are the limits of the stellar evolutionary model that we adopt in what follows. Objects outside this range are either saturated or fall below the detection limit of the IPHAS survey and so these truncation limits do not affect our results, other than constraining the parameter space to a sensible domain.
- $P(\tau)$: the age prior is assumed uniform in the logarithm and truncated between 0.1 and 100 Myr, which again corresponds to the limits of the evolutionary model.
- $P(\dot{M}_{acc})$: the accretion rate prior is assumed uniform in the logarithm and is truncated between 10^{-15} and $10^{-2}\,M_\odot\,yr^{-1}$. This range entails all the accretion rates commonly reported in the literature and goes well below the typical detection limit of $\sim 10^{-10}\,M_\odot\,yr^{-1}$ we found in Paper I.
- $P(R_{in})$: the disc truncation radius follows a Gaussian distribution with a mean of $5\,R_*$ and standard deviation $2\,R_*$. This is a commonly used assumption based on the typical co-rotation radius of T Tauri stars (Gullbring et al. 1998).
- $P(d)$: for the distance prior we adopt a Gaussian distribution centred on the widely cited distance towards NGC 2264 of 760 pc (Sung et al. 1997). The standard deviation of 5 pc reflects the approximate diameter of the cluster and therefore the uncertainty in our results will reflect only the relative distance errors. We note that a recent distance estimate by Baxter et al. (2009) puts the region at 913 pc and so the systematic error in the distance may be significantly larger than 5 pc. However, because we aim to investigate the relative properties of objects in the cluster, rather than systematic errors for the cluster as a whole, we opt not to model the absolute distance uncertainty here. (Note that the Gaia survey will remove most of this uncertainty in the future.)
- $P(A_0)$: for the extinction parameter we adopt the empirical distribution of extinction for 202 candidate members as determined by Rebull et al. (2002) on the basis of moderate-resolution spectroscopy of low-mass objects in NGC 2264. We found that their distribution can be well-approximated as a log-normal with mean $\log A_0 = -0.27$ ($A_0 = 0.54$) and a broad standard deviation $\sigma \log A_0 = 0.46$.

2.3.2 $SED_{\rm int}$ likelihood (Eqn. 14)

The intrinsic SED of a young star is predicted as a function of mass and age as follows.

First, intrinsic broad-band magnitudes $M_{r'}$, $M_{i'}$ and M_J are interpolated from the evolutionary model due to Siess et al. (2000) for solar metallicity ($Z = 0.02$). Their model consists of 29 separate mass tracks from 0.1 to 7 $M_\odot$ from which isochrones can be computed using an online tool[1]. We used this tool to download the model at 50 different ages ranging between 0.1 Myr and 100 Myr, i.e. we obtained a dense sampling of the model in 1450 discrete points as a function of mass and age ($= 29 \cdot 50$ points).

The Siess et al. model provides intrinsic magnitudes on the basis of the empirical conversion tables presented by Kenyon & Hartmann (1995), which provide intrinsic colours and bolometric corrections as a function of stellar effective temperature. These tables only provide a calibration for the Cousins photometric system however, so we had to convert the model R_C/I_C magnitudes to IPHAS r'/i' using the transformations given in Paper I.

We then approximated the evolutionary model as a continuous function by fitting 1450 Radial Basis Functions (RBF) to the collection of discrete model points using the Python module SCIPY.INTERPOLATE.RBF (cf. Appendix A). The resulting set of basis functions provides us with an accurate and fast tool to interpolate values from the grid.

Having obtained intrinsic magnitudes as a function of age and mass, we then predict the narrow-band magnitude $M_{H\alpha}$ using the grid of IPHAS colour simulations presented in Paper I, where we determined the intrinsic colour (r'-Hα) as a function of (r'-i') on the basis of a library of observed spectra.

These steps provide us with a forward model of the intrinsic magnitudes as a function of age and mass, i.e. $f_{\rm SED}(M_*, \tau) = \{M_{\rm r'},\ M_{\rm H\alpha},\ M_{\rm i'},\ M_{\rm J}\}$. For simplicity, we assume here that this model predicts absolute magnitudes with zero uncertainty. In reality, there are known to be significant systematic differences between pre-main sequence models, ranging up to 2-4 Myr in age and 0.2 $M_\odot$ in mass (Paper I). However, the modelling and study of these systematic effects are beyond the scope of this work, and so we adopt a deterministic likelihood:

$$P(\mathrm{SED_{int}} \mid M_*, \tau) = \begin{cases} 1 & \text{if } \mathrm{SED_{int}} = f_{\rm SED}(M_*, \tau) \\ 0 & \text{if } \mathrm{SED_{int}} \neq f_{\rm SED}(M_*, \tau) \end{cases} \tag{19}$$

[1] http://www.astro.ulb.ac.be/∼siess/WWWTools/Isochrones

2.3.3 $\mathrm{L_{H\alpha}}$ *likelihood (Eqn. 15)*

Hα is the strongest emission line due to accretion, and its intensity can be used to trace the accretion luminosity.

As explained in §2, accretion is thought to take place along magnetic field lines which act as channels connecting the disc to the star from an inner disc truncation radius $\mathrm{R_{in}}$. The infalling gas is essentially on a ballistic trajectory, falling on to the star at near free-fall velocities, producing a hot impact shock (Calvet & Gullbring 1998; Gullbring et al. 2000). The energy released in these shocks heats the infalling circumstellar gas, which explains the Hα emission.

Under the assumption that the free-fall gravitational energy released in the impact accretion shock is reprocessed entirely in the accretion energy $\mathrm{L_{acc}}$, we may write:

$$\mathrm{L_{acc}} = \frac{\mathrm{GM_*\dot{M}_{acc}}}{\mathrm{R_*}}(1 - \frac{\mathrm{R_*}}{\mathrm{R_{in}}}), \tag{20}$$

and therefore:

$$\log \frac{\mathrm{L_{acc}}}{\mathrm{L_\odot}} = 7.496 + \log \frac{\mathrm{M_*}}{\mathrm{M_\odot}} + \log \frac{\mathrm{\dot{M}_{acc}}}{\mathrm{M_\odot} yr^{-1}} - \log \frac{\mathrm{R_*}}{\mathrm{R_\odot}} + \log(1 - \frac{\mathrm{R_*}}{\mathrm{R_{in}}}), \tag{21}$$

where $\mathrm{R_*}$ is derived from a forward model $f_{\mathrm{R_*}}(\mathrm{M_*}, \tau)$ obtained by interpolating the Siess et al. model in the same way as described previously.

The accretion luminosity $\mathrm{L_{acc}}$ has previously been found to relate to $\mathrm{L_{H\alpha}}$ as a power-law relationship (e.g. Herczeg & Hillenbrand 2008; De Marchi et al. 2010). In Paper I (fig. 8) we presented a compilation of 107 objects from the literature for which both $\mathrm{L_{acc}}$ and $\mathrm{L_{H\alpha}}$ estimates are available, whereby $\mathrm{L_{acc}}$ has been derived from blue continuum excess measurements. We used the STATS.LM function in the R statistical environment to determine the linear least squares regression:

$$\log \mathrm{L_{H\alpha}} = (0.64 \pm 0.04) \log \mathrm{L_{acc}} - (2.12 \pm 0.08). \tag{22}$$

This empirical relationship shows a significant scatter however (rms = 0.43), which is commonly assumed to be caused by a combination of physical effects (e.g. absorption by stellar winds, uncertain extinction corrections, emission not due to accretion). The scatter in this relationship dominates the uncertainty in the inferred accretion rates (Paper I). We take this into account by modelling the likelihood function as a Log-normal distribution:

$$P(\mathrm{L_{H\alpha}} \mid \mathrm{M_*}, \tau, \mathrm{\dot{M}_{acc}}, \mathrm{R_{in}}) \sim \log \mathcal{N}(\log \mathrm{L_{H\alpha}}, \sigma = 0.43), \tag{23}$$

where $\log \mathrm{L_{H\alpha}}$ is obtained by combining Eqns. 21 & 22.

2.3.4 $EW_{\mathrm{H\alpha}}$ *likelihood (Eqn. 16)*

To infer $\mathrm{EW_{H\alpha}}$ we need to predict the stellar continuum luminosity in the Hα band:

$$\mathrm{L(H\alpha)_{cont}} = \mathrm{L}_V(\mathrm{H\alpha}) \cdot 10^{-0.4 \cdot [\mathrm{M_{H\alpha}} + 0.03]} \tag{24}$$

where $\mathrm{M_{H\alpha}} \in \mathrm{SED_{int}}$ is the previously estimated intrinsic magnitude of the star and $\mathrm{L}_V(\mathrm{H\alpha}) = 0.316\ \mathrm{L_\odot}$ is the luminosity of Vega in the IPHAS Hα passband (Paper I). We may then obtain the equivalent width from its definition:

$$\mathrm{EW_{H\alpha}} = -RW \cdot \frac{\mathrm{L_{H\alpha}}}{\mathrm{L(H\alpha)_{cont}}}, \tag{25}$$

where $RW = 95\,\mathrm{\AA}$ is the rectangular width of the IPHAS Hα filter.

The uncertainty associated with the conversion of $\mathrm{L_{H\alpha}}$ into $\mathrm{EW_{H\alpha}}$ can be assumed negligible and for simplicity we take the likelihood to be deterministic.

2.3.5 SED_{app} *likelihood (Eqn. 17)*

The apparent SED is obtained by correcting the intrinsic SED for the effects of distance, extinction and Hα emission in three steps.

First, the distance modulus $5 \log(d) - 5$ is added to each of the intrinsic magnitudes. Second, offsets for the r'/Hα/i' magnitudes are obtained as a function of $\mathrm{A_0}$ and $\mathrm{EW_{H\alpha}}$ by means of RBF-interpolation from the pre-computed grid of simulated photometry from Paper I. Finally, the offset for the J magnitude for extinction is computed using the reddening law due to Schlegel et al. (1998). The J-band offset cannot be computed in the same way as the other bands, because the optical spectral library on which our grid of simulated colours is based does not extend far enough into the near-infrared.

Again, this likelihood is assumed deterministic.

2.3.6 SED_{obs} *likelihood (Eqn. 18)*

Finally, we compute the likelihood of observing $\mathrm{SED_{obs}}$ when expecting $\mathrm{SED_{app}}$. Assuming normally distributed uncertainties on the observed magnitudes, the likelihood is given by a multivariate Gaussian:

$$P(\mathrm{SED_{obs}} \mid \mathrm{SED_{app}}) = \frac{1}{(2\pi)^{k/2} |\mathbf{\Sigma}|^{1/2}} e^{-D^2/2}, \tag{26}$$

where $\mathbf{\Sigma}$ is the covariance matrix of the magnitudes in $\mathrm{SED_{obs}}$ and D^2 is given by

$$D^2 = (\mathrm{SED_{obs}} - \mathrm{SED_{app}})^T \mathbf{\Sigma}^{-1} (\mathrm{SED_{obs}} - \mathrm{SED_{app}}). \tag{27}$$

We assume that the uncertainties between magnitudes in $\mathrm{SED_{obs}}$ are uncorrelated, in which case $\mathbf{\Sigma}$ is a diagonal matrix and D^2 can be simplified to

$$D^2 = \sum_{i=1}^{n} \frac{[\mathrm{SED_{obs}}(m_i) - \mathrm{SED_{app}}(m_i)]^2}{\sigma_{m_i}^2 + \sigma_{\mathrm{cal}}^2} \tag{28}$$

where $\sigma_{m_i}^2$ is the squared photometric uncertainty for each magnitude m_i and σ_{cal}^2 is an extra uncertainty term which we add to account for absolute calibration errors. This is required because the photometric uncertainty given by the IPHAS database only represents the relative uncertainty due to background noise, which is often smaller than 0.01 mag for bright stars. In reality, ground-based survey photometry rarely reaches an absolute accuracy better than a few percent due to additional sources of noise (e.g. variable atmospheric conditions).

We found empirically that a value of $\sigma_{\mathrm{cal}} = 0.1$ is required to prevent the expected match between the observed data and the modelled SED to be too exact. Leaving this term out has the effect of producing a complex multimodal probability landscape which prevents the sampling algorithm –discussed below– from converging efficiently. The term can be considered as a way to account for all those sources of noise which we could not explicitly model in the other parts of the model.

2.4 Sampling the joint distribution using MCMC

In the previous sections we defined the joint distribution $P(\boldsymbol{\theta}, \boldsymbol{d})$ and thus the posterior $P(\boldsymbol{\theta}|\boldsymbol{d})$ as a product of priors and likelihoods. At this point the question remains how this distribution is computed in practice. A simple "brute force" method which computes the model for the entire parameter space is computationally intractable. Even if we were to sample the parameter space at a coarse resolution, say, 100 values for each of the 10 model parameters, we would require the distribution to be computed in 10^{20} points. This would occupy the author's computer for many billion years.

Fortunately, probability distributions can be sampled efficiently using a specialised class of algorithms called *Markov Chain Monte Carlo* (MCMC). In brief, MCMC algorithms generate a pseudo-random walk in the parameter space in such a way that, over time, points in the space are visited with a frequency that is proportional to a specified probability distribution. Only useful points in the parameter space tend to be evaluated and we avoid wasting time calculating an infinite number of points in the improbable regions. The details of MCMC algorithms are beyond the scope of this paper but we recommend Chib & Greenberg (1995) for an introduction and MacKay (2003) and Gregory (2005) for background reading.

Several libraries are available which allow Bayesian models to be defined and sampled using MCMC. In this work we used the PyMC framework for Python (Patil et al. 2010) which has the advantage that it allows models to be defined in a very concise way. Our annotated model takes less than two pages and is therefore included at the end of this paper for easy reference (Appendix A). As such, this paper contains a precise and repeatable specification of the parameter estimation procedure. The source code and accompanying files are also available from the GitHub repository of the author[2].

For each object under study in this paper (to be explained in §3), we sampled the joint distribution using the default settings of PyMC, which is to use the traditional *Metropolis-Hastings* walking algorithm with a Gaussian *step proposal function* (we refer to the manual of PyMC for definitions of these technical terms). PyMC automatically tunes the size of the step proposal function to ensure the walk is made efficiently with an *acceptance rate* between 20 and 50%.

It is usually sufficient to sample a probability distribution in only a few hundred independent points to obtain a sufficiently accurate approximation. MCMC algorithms typically require far more samples to be obtained however, because the walking algorithm naturally produces chains which are auto-correlated and hence may be stuck at local maxima. The true number of iterations which are required depends on the shape of the probability landscape; a complex or multimodal distribution with sharp hills and valleys tends to require far more samples. In our application we find the chains to be auto-correlated over a typical length of 50 to 250 steps (depending on the properties of the star). For this reason, we decided to sample each object in 250 000 points such that the total number of points is at least 3 orders of magnitude larger than the auto-correlation effect (i.e. at least 1000 truly independent samples are obtained for each object).

It is likely that our application would profit from recent advances in MCMC algorithms which claim to reduce the auto-correlation effect considerably. We draw the reader's attention to an implementation of such algorithm made available by Foreman-Mackey et al. (2012) which we intend to evaluate in future work.

2.5 Results of the sampling procedure

To verify the reliability of our procedure, the sampling was carried out using 5 independent walks of 50 000 steps, each starting at randomised positions (with a *burn-in length* of 1 000 steps). We consistently found these independent chains to converge to the same parameter-space regions of high likelihood within a few hundred iterations, i.e. fast convergence towards a global maximum was reached in all cases.

We visualised the samplings by means of 2D-histograms (Figs. 4-7), which trace the posterior marginalised over all other parameters. The first two examples (Figs. 4-5) are representative for the vast majority of objects in our study which show low levels of extinction ($A_0 < 1$). The main difference is that the first example shows evidence for Hα emission, whereas the second example does not.

Non-typical examples are shown in Figs. 6-7. These objects have (i'-J) colours which are significantly redder compared to the first two examples, in a way that is consistent with a higher level of extinction. We note that the uncertainties are significantly larger for these more highly reddened objects. This is a result of our decision to adopt a log-normal extinction prior with a peak near A_0=0.5 and a long tail towards higher values (corresponding to the empirical distribution for the region which we adopted in §2.3).

If we had not included this prior information then the uncertainty in Figs. 4-5 would have been similar to that in Figs. 6-7. In other words, whilst our dataset offers a rough constraint on the individual extinction, the degeneracy between extinction and mass is not resolved entirely and the prior makes a contribution towards constraining the result. At the same time, Figs. 6-7 demonstrate that the prior does not prevent higher levels of extinction and uncertainty to be revealed when the data are inconsistent with low extinction. We will quantify the contribution of the prior in the discussion at the end of the paper (§5.4).

We draw the reader's attention to the 'banana'-shaped posterior which appears when the uncertainty is large. This is a natural result of curves in the model evolutionary tracks relative to the reddening vector.

We also note that the range of possible masses in Fig. 7 appears to go somewhat beyond the mass range that is provided by the Siess et al. model. Only a few faint objects are affected in this way however, and we decided not to remove these from our study.

Finally, we note that a visual inspection of all objects under study revealed that these marginalised posteriors distributions are single-moded and quasi-symmetric when the logarithm of each parameter is considered. This implies that the marginalised posterior can be well-characterised by means of computing expectation values and standard deviations.

The posterior was sampled and characterised in this

[2] https://github.com/barentsen

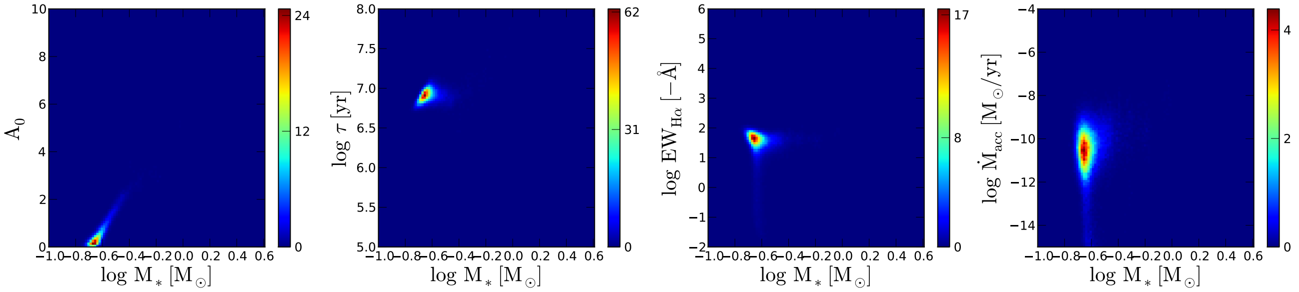

Figure 4. Normalised 2D-histograms of the distribution of points in the MCMC chains for object C11059 (r'=19.4 ± 0.03; r'-Hα=1.3 ± 0.04; r'-i'=1.9 ± 0.04; i'-J=2.1 ± 0.06). These histograms trace the posterior distributions marginalised over all other parameters. Blue regions show areas in the parameter space with low probability, red areas show areas with high probability.

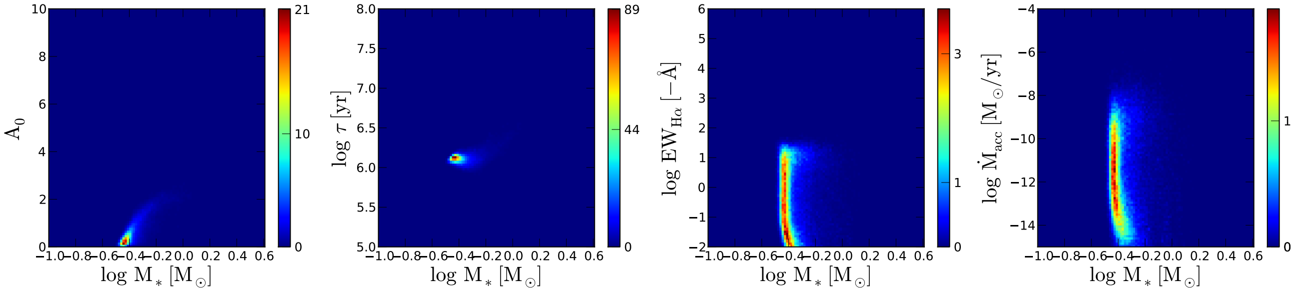

Figure 5. Histograms for object C15152 (r'=16.3 ± 0.002; r'-Hα=0.8 ± 0.004; r'-i'=1.4 ± 0.003; i'-J=1.7 ± 0.02). Compared to the first example, this is a brighter object with near-zero Hα emission and a slightly higher mass. The probability region is marginally more condensed owing to the smaller photometric uncertainties.

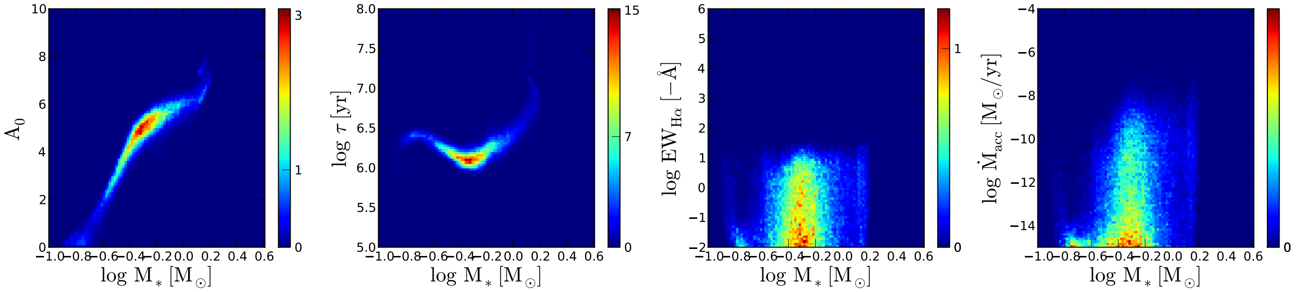

Figure 6. Histograms for object C29493 (r'=19.3 ± 0.04; r'-Hα=0.7 ± 0.08; r'-i'=2.0 ± 0.05; i'-J=3.1 ± 0.04). Compared to the first example shown in Fig. 4, the i'-J colour is significantly redder which results in a higher extinction. We also draw attention to the banana-like shape of the high-probability region, which is a result of changes in the direction of the model isochrones relative to the reddening vector.

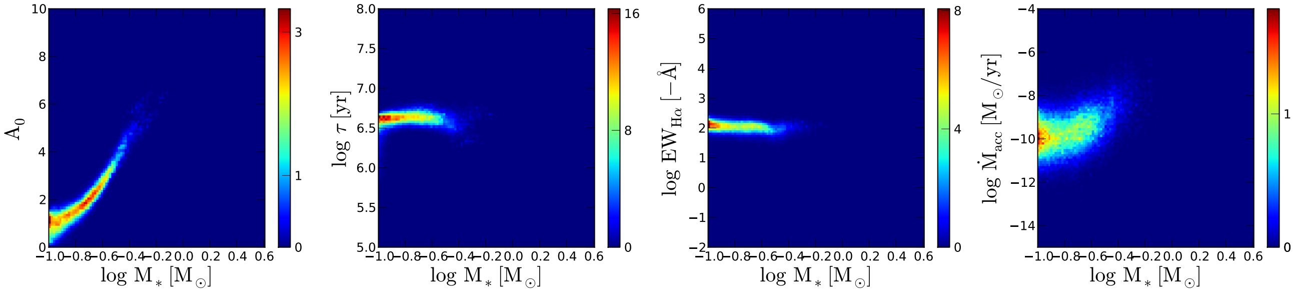

Figure 7. Histograms for object C36902 (r'=20.6 ± 0.09; r'-Hα=1.6 ± 0.11; r'-i'=2.1 ± 0.10; i'-J=3.1 ± 0.07). This is one of the lowest-mass objects in our study. The high-probability region appears to go slightly beyond the range of the Siess et al. model.

way for a total of 587 known members of the NGC 2264 star-forming region. In the following sections, we discuss how the input dataset was obtained (§3) and how the derived parameters make sense of the objects' positions in colour/magnitude diagrams (§4-§5).

3 APPLICATION TO NGC 2264

NGC 2264, also known as the *Cone Nebula* or *Christmas Tree Cluster*, is located at a distance of ∼760 pc in the constellation of Monoceros (Sung et al. 1997). The cluster is estimated to contain ∼1000 members, most of which have been identified using a range of methods including Hα and variability surveys, X-ray observations and mid-infrared imaging (a review is given by Dahm 2008).

Stellar parameters have previously been estimated for members of the region using a wide variety of colour/magnitude diagrams and evolutionary models (e.g. Park et al. 2000; Rebull et al. 2002; Sung et al. 2004; Flaccomio et al. 2006; Dahm et al. 2007). There are significant systematic differences between these studies however, with median age estimates for the cluster ranging between ∼1 and 5 Myr (Dahm 2008).

Moreover, there is only a partial overlap in the membership samples considered in these previous studies, owing to differences in the datasets and member selection criteria. In what follows we attempt to select the largest and most homogeneous membership sample of NGC 2264 to date, and then use it to demonstrate our method.

3.1 Membership list

The most comprehensive catalogue of objects towards NGC 2264 has been presented by Sung et al. (2008, 2009) hereafter referred to as S08 and S09. Their work is based on a compilation of

(i) deep *VRI* and Hα photometry using the 3.6 m Canada France Hawaii Telescope (CFHT);
(ii) bright *BVRI* and Hα photometry using the 1 m telescope at Siding Spring Observatory (SSO);
(iii) low- and moderate-resolution literature spectroscopy from Reipurth et al. (2004) and Dahm & Simon (2005);
(iv) archival X-ray observations from the Chandra and ROSAT Space Telescopes; and
(v) archival infrared observations from the Spitzer Space Telescope.

The authors constructed a catalogue of 69 674 optical objects detected towards the cluster (tables 3, 8 & 9 in S08) and then assigned various "membership codes" by crossmatching the Hα, X-ray and infrared data. We used these codes to select a total of 1191 likely members which satisfy one or more of the following criteria:

(i) Hα emission stronger than chromospherically active main-sequence stars (membership code: 'H', 'E', '+', 'P' or 'p');
(ii) strong X-ray emission (code: 'X', '+', '-', 'M' or 'P');
(iii) Spitzer colours consistent with a protoplanetary disc (code: 'I', 'II', 'II/III', 'pre-TD' or 'TD' in S09).

In this sample of 1191 objects, 42% satisfy the Hα criterion, 62% satisfy the X-ray criterion, and 38% satisfy the Spitzer criterion. There is considerable overlap: 32% satisfy more than one criterion while 10% satisfy all three.

Additional signatures of membership such as radial velocities, or chemical indicators of youth such as Lithium, are currently not available for most of these objects. This is, in part, because half of the sample is fainter than V>18 for which high-resolution spectroscopy becomes increasingly expensive. Based on the clustering properties and positions in the colour-magnitude diagram however, S08 convincingly argued that the vast majority of objects in this sample are genuine members of NGC 2264. Nevertheless, it is likely that our sample contains a small number of foreground/background objects.

The spatial distribution of the sample is shown in Fig. 8 together with the footprints of the observations from which the sample was compiled.

3.2 IPHAS counterparts

IPHAS is a 1800 deg^2 photometric survey of the Northern Galactic Plane ($30^{\circ} \lesssim \ell \lesssim 220^{\circ}$, $-5^{\circ} \lesssim b \lesssim +5^{\circ}$) carried out using a narrow-band Hα filter and the broad-band Sloan r' and i' filters, using the 0.3 deg^2 Wide-Field Camera (WFC) at the 2.5-meter Isaac Newton Telescope (INT) in La Palma. Data towards NGC 2264 were obtained as part of the survey during several observing runs between 2003 and 2009. The central part of the cluster, containing the vast majority of members, was observed on 2008 January 17 with an average seeing of $1.2 \pm 0.1''$(IPHAS field numbers 3773 and 3773o).

We used the Montage toolkit to create an Hα mosaic of 20 fields towards the region which is shown in Fig. 9. The contrast of the mosaic has been stretched using an arcsinh curve to bring out the Hα background emission.

All data were pipeline processed at the Cambridge Astronomical Survey Unit (CASU) as detailed in Irwin & Lewis (2001), Drew et al. (2005) and González-Solares et al. (2008). This routinely includes photometric calibration based on nightly standard star fields, with all magnitudes based on the Vega system. In addition, we have been able to draw upon the results of a global calibration of the survey data which significantly reduces field-to-field magnitude shifts (Drew et al, in preparation).

We crossmatched the sample of 1191 members from S08 against the IPHAS catalogue. The astrometry of both S08 and IPHAS is based on 2MASS reference coordinates which offers a typical accuracy of 0.1 to 0.2″ (González-Solares et al. 2008). For this reason, we decided on a strict matching distance upper bound of 0.5″. A total of 819 members were found to have a counterpart within the matching distance in *all* three IPHAS bands. For 72 of these objects Hα photometry was not previously available in the S08 catalogue.

From the 372 objects which could not be matched in all three bands, 249 fall below the typical detection limit of IPHAS ($R > 20$) and 44 are saturated ($R < 13$). Most of the remaining objects were found to be blended with a nearby neighbour in IPHAS while being resolved in the higher resolution CFHT-based data from S08, which produces an astrometric offset. Increasing our matching distance to 1.0″ would include 35 of these objects, but we decide against this in favour of data reliability.

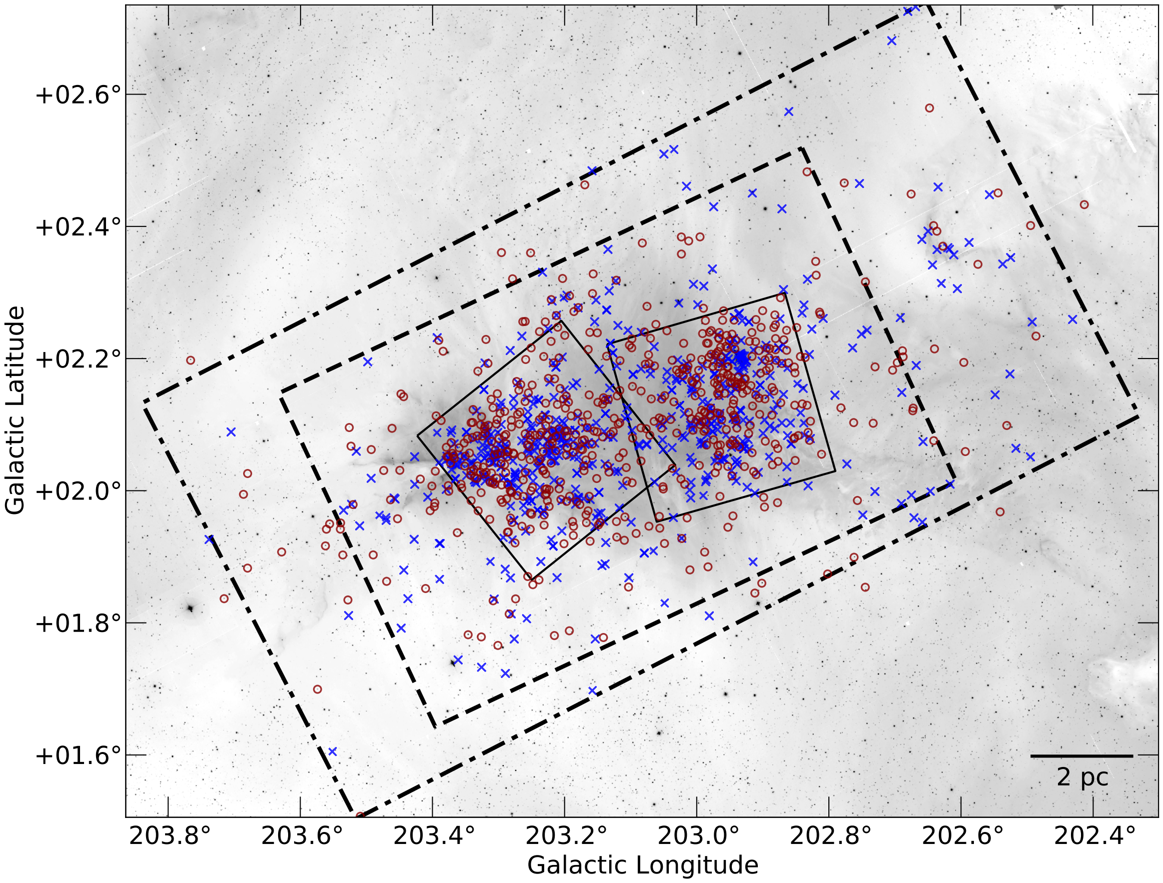

Figure 8. Spatial distribution of known candidate members in NGC 2264 taken from the works by Sung et al. (see text.) Red circles show objects for which high-reliability IPHAS and 2MASS photometry is available which satisfy the strict quality requirements defined in §2.4; these are the objects which are studied in our work. Blue crosses show candidate members for which we could not obtain high-quality IPHAS/2MASS photometry. Large rectangles show the footprints of the deep optical CFHT observations (dash-dotted line), Spitzer IRAC observations (dashed line) and Chandra observations (solid line). A few objects fall outside these footprints, because they were included on the basis of bright optical photometry or literature spectroscopy (footprints not shown). The background image is an inverted version of Fig. 9.

In the majority of cases the matched objects were detected 2 or 3 multiple times by the IPHAS survey, usually in the same night. This is because IPHAS fields are observed twice with a small offset to account for gaps between CCD chips in the camera. When two or more detections of the same object are available, we derived the mean magnitude and updated the uncertainty.

3.3 New candidate members from IPHAS

Having obtained a sample of 819 very likely members from literature, we carried out a search in the IPHAS database for any new Hα emitters which may have been missed during previous searches. Using the method detailed in Paper I we identified a total of 164 objects which are located confidently above the main locus of stars in the IPHAS (r'-i')/(r'-Hα) diagram. The selection threshold explained in Paper I is designed to avoid chromospherically active field stars, and so the vast majority of these objects are likely to be accreting T Tauri stars.

We find that 150 (91%) of these objects were already selected because they are classified as strong Hα emitters in S08, while a further 9 objects were selected on the basis of their X-ray or infrared emission but not on the basis of Hα (object IDs: C27213, C34019, C36055, C37804, C39811, C42487, C8431, W3407, W5604). Only 5 objects were not already selected and are added to our sample (C18650, C19751, C25302, W2320, W5620). This brings the sample size to 824 objects. The fact that we are unable to identify a significant number of new Hα-emitters suggests that the sample is very complete in this respect, at least down to the detection limits of IPHAS.

3.4 Quality criteria

At this point we could continue our investigation with all of the 824 objects for which IPHAS counterparts were found. However, we choose to narrow down the sample to 587 objects using the following strict quality requirements:

(i) the photometric uncertainty on each of the three IPHAS magnitudes must be smaller than 0.1;

(ii) each object must be classified as "strictly stellar" or "probable stellar" in the IPHAS catalogue in all three bands

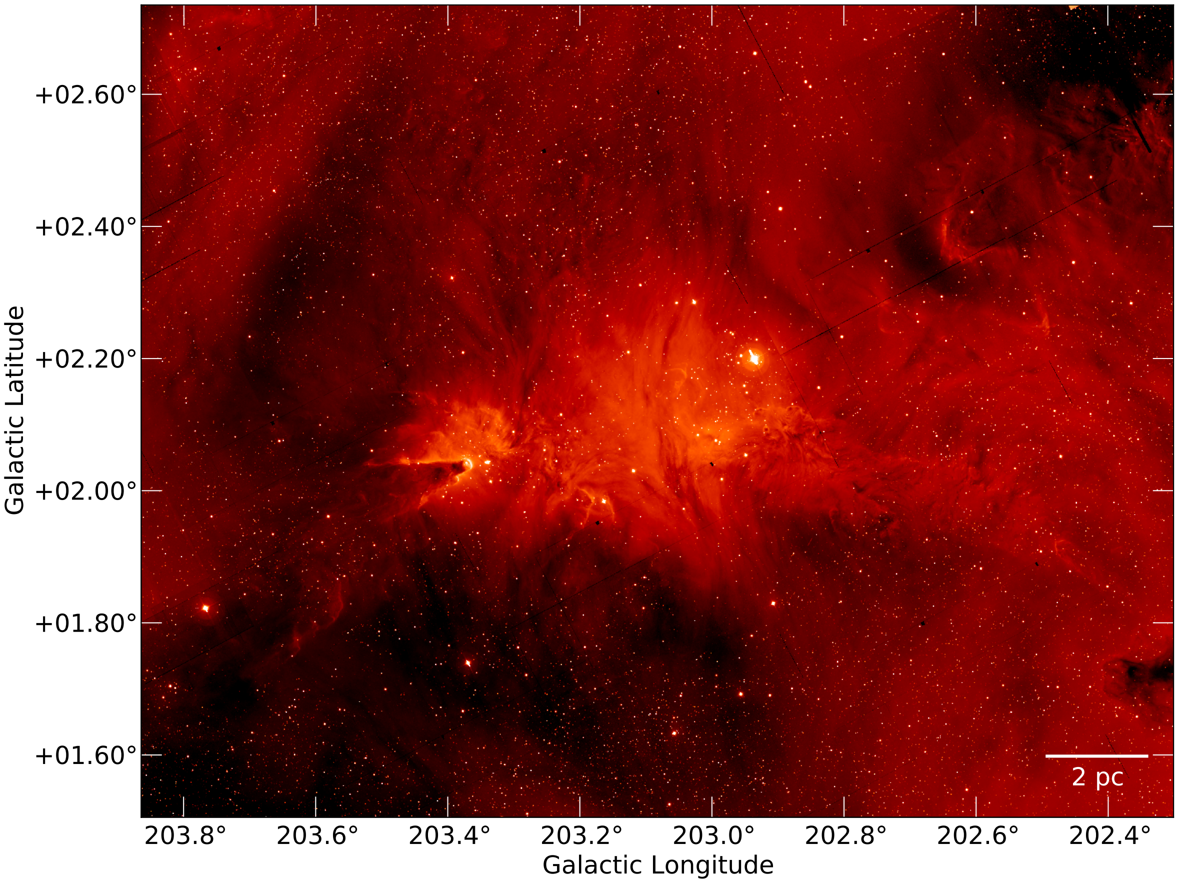

Figure 9. Mosaic of IPHAS observations towards NGC 2264 in the Hα band. A few narrow black bands can be seen, which indicate small gaps of missing data which will be completed in a future IPHAS data release.

(quality flag '-1' or '-2' defined by González-Solares et al. 2008);

(iii) the object must have a J-band counterpart in the 2MASS database with Signal-to-Noise Ratio (SNR) > 5 (quality flag 'A', 'B' or 'C');

(iv) the object must not be marked as an unresolved binary in the catalogue due to S08 or 2MASS (flag 'D') and its nearest neighbour must be further away than 1 arcsecond.

Criterion (i) avoids faint sources with high uncertainties. The criterion corresponds to a SNR of 10 or typical magnitude limits of 20.5 in r' and 19.5 in i'/Hα (González-Solares et al. 2008), although we note that there are small spatial variations in the true magnitude limits depending on the observing conditions and the number of repeat observations. Only 34 objects did not meet this criterion.

Criterion (ii) deals with the issue of flux-contamination. The magnitudes in the IPHAS database are based on aperture photometry which is prone to contamination by nearby neighbours or spatially varying background emission. The IPHAS pipeline solves this problem by tracking variations in the background emission on scales of 20-30 arcsec. In some cases the background changes on scales smaller than 5-10 arcsec however, in which case photometric measurements of faint sources become unreliable.

Fortunately, photometry which is unreliable for this reason is flagged in the pipeline as part of the morphological classification step (see §2.1 in Paper I). In brief, the pipeline derives a curve-of-growth for each object from a series of aperture flux measures with different aperture radii. When this curve deviates from the characteristic point spread function (PSF) of other stars in the field, the object is flagged as "extended" (class 1) or "probable extended" (class -3).

By requiring objects to be classified as "strictly stellar" (code -1) or "probable stellar" (code -2) in all three bands, we ensure that only reliable measurements for objects with a normal-shaped PSF are included in our analysis. A total of 143 objects did not meet this criterion.

Criterion (iii) is introduced because IPHAS magnitudes alone are not sufficient to constrain the extinction of individual objects as we explained in §2. A total of 62 objects did not meet this criterion. Future work could benefit from the deeper J-band data offered by the UK Infrared Deep Sky Survey (UKIDSS, Lucas et al. 2008), but data for NGC 2264 is not yet available from that survey at this time.

Finally, criterion (iv) avoids likely problems due to source confusion. A total of 30 objects did not meet this criterion.

After applying each of the criteria, we are left with 587 objects (because a number of objects failed more than one criterion). The resulting table of IPHAS and 2MASS photometry for these objects is given in Table 3 and forms the basis for our analysis.

4 RESULTS

We obtained Bayesian posterior distributions for each of the 587 objects selected above. We then summarised the posteriors by computing marginalised means and standard deviations, i.e. we derived a point estimate per parameter for each object. These values are listed in Table 4 and visualised by histograms in Fig. 10. In this section we provide a brief overview of the results by inspecting the distributions of these point estimates. This will lead us to a set of preliminary results on the properties of NGC 2264, such as its mass distribution, median age and fraction of accretors.

Whilst point estimates of stellar parameters are widely used as a tool to investigate the properties of a cluster, we warn that the presence of large uncertainties can make a direct analysis of point estimates unreliable. For example, while it is tempting to estimate the age of NGC 2264 from the histogram of mean stellar ages shown in Fig. 10, the result may be biased due to the presence of assymetric and correlated uncertainties (see Pont & Eyer 2004). This does not imply that our results cannot be used to study the properties of the cluster, in fact our posteriors contain all the required information on the uncertainties and degeneracies. In order to exploit this information however, we would need to build a probabilistic model which links the global parameters of the cluster to the posteriors of the stars.

Whilst it is easy to see that our hierarchical model may be extended to include the global properties of the cluster as parameters, the priors and likelihoods of such parameters would have to be chosen with care. They would need to reflect the current knowledge in the field and allow the right questions to be answered. For example, if we were to estimate the age and age spread of the cluster, we would need to make a detailed appraisal of the accuracy of pre-main sequence isochrones and include the information in the model.

Defining such a cluster model is beyond the scope of the present work. In what follows we merely offer the reader a concise summary of the distribution of the point estimates, which may be considered as a first-order approximation of the global properties of NGC 2264. In §6 we will discuss the future prospect of extending our work to obtain a complete model of the cluster.

4.1 Masses, ages & extinction

The mass distribution (Fig. 10, panel a) shows the expected power-law increase towards lower masses. Compared with the Kroupa IMF (blue solid line) we find a deficit of objects with masses below $0.25\,\mathrm{M}_{\odot}$. Our sample is significantly less complete below this mass due to the SNR quality criteria imposed in §2.4. Similarly, many stars with masses heavier than $\sim$1.2 $\mathrm{M}_{\odot}$ are missing due to the saturation limits of the IPHAS survey. The highest inferred mass in our sample equals $1.8^{+0.3}_{-0.2}\,\mathrm{M}_{\odot}$.

The age distribution (panel b) shows a mean age of $\log\tau = 6.48 \pm 0.38$ (which corresponds to the median $\tau = 3.0\,\mathrm{Myr}$), albeit with a large apparent dispersion between 1.8 and 4.5 Myr (25 and 75% quartiles). Our median estimate of 3 Myr is identical to the main-sequence turn-off age obtained in the seminal paper by Walker (1956), and is also consistent with the age of 3.1 Myr reported by Sung et al. (2004) using the same set of isochrones as in our work.

The extinction (panel c) shows a mean of $\log A_0 = -0.37 \pm 0.20$ ($A_0 = 0.43$) which is consistent with the widely reported low levels of foreground extinction. 30 objects show higher levels of extinction ranging between A_0=1 and 3, while only 4 objects show extinctions beyond $\mathrm{A}_0 > 3$ (beware that the vertical axis in panel c is logarithmic for clarity).

Compared to the log-normal extinction prior (solid blue line) there is a deficit of objects with large extinctions, which will be discussed in §5.4.

4.2 Hα emission & accretion rates

The Hα luminosities (panel d) are shown as a fraction of the object's bolometric luminosity L_* which we derived from the Siess et al. model as a function of an object's age and mass. This allows us to show the distribution relative to the *chromospheric saturation limit* at $\log \mathrm{L}_{\mathrm{H}\alpha}/\mathrm{L}_* = -3.3$ (dashed line). This is the maximum level of Hα emission observed in clusters at the age of 65 to 125 Myr, where accretion is thought to have ceased and Hα-emission is produced entirely by chromospheric activity (Barrado y Navascués & Martín 2003).

This saturation limit is a widely used criterion to separate accreting from non-accreting objects. In the remainder of this paper we refer to objects which fall above the limit as "accretors" or *Classical T Tauri Stars* (CTTS) while those which fall below the limit are "non-accretors" or *Weak-lined T Tauri Stars* (WTTS).

According to this definition, the accretion fraction is $20 \pm 2\%$ (115 objects). This fraction may be slightly underestimated because 36 additional objects fall only just below the criterion ($\log \mathrm{L}_{\mathrm{H}\alpha}/\mathrm{L}_*$ between -4.0 and -3.3). It is likely that some of these objects are undergoing very low levels of mass accretion which we cannot distinguish from chromospheric activity using our dataset. If we were to assume that all these objects are undergoing accretion then the fraction of accreting stars would rise to 25%. We note that a frequency of 20 to 25% for a cluster of 3 Myr is in excellent agreement with other clusters of a similar age (Fedele et al. 2010).

The Hα EW distribution for the accreting stars (panel e) shows a range from -12 to -546 Å. The corresponding mass accretion rates (panel f) range from $10^{-10.7}$ to $10^{-6.4}\,\mathrm{M}_{\odot}/\mathrm{yr}$ with a median of $10^{-8.4}\,\mathrm{M}_{\odot}/\mathrm{yr}$. They are in broad agreement with accretion rates found in clusters of a similar age (e.g. Sicilia-Aguilar et al. 2010).

5 DISCUSSION

The results presented in this paper can be used to investigate a wide range of questions regarding star formation and the history of NGC 2264. However, in the remainder of this paper we choose to restrict ourselves to an evaluation of the method with an eye on future improvements. In what follows we show that the results obtained match (i) those expected from traditional colour-colour and colour-magnitude diagrams, and (ii) those previously reported in the literature using spectroscopy. We also discuss a small number of objects with anomalous colours and discuss the influence of the extinction prior.

Name	IPHAS ID	IPHAS magnitudes			2MASS magnitudes		
(S08)	J[RA(2000)+Dec(2000)]	r'	Hα	i'	J	H	K
C11059	J063955.90+094239.8	19.41±0.03	18.14±0.03	17.56±0.02	15.44±0.06	14.69±0.05	14.28±0.08
C11997	J063957.95+094104.7	17.91±0.01	16.50±0.01	16.16±0.01	14.36±0.04	13.65±0.03	13.35±0.04
C12135	J063958.30+092848.6	18.51±0.01	17.66±0.02	16.74±0.01	14.99±0.04	14.34±0.05	14.11±0.07
C12598	J063959.23+100607.7	17.26±0.00	16.18±0.01	16.07±0.00	13.91±0.02	12.80±0.03	12.10±0.02
C13507	J064001.30+094300.5	19.10±0.04	16.91±0.01	18.14±0.04	15.00±0.04	13.17±0.02	11.94±0.03
C14005	J064002.67+093524.3	16.86±0.00	15.87±0.00	15.35±0.00	13.70±0.03	12.99±0.03	12.71±0.03
C15152	J064005.22+095056.6	16.27±0.00	15.50±0.00	14.87±0.00	13.19±0.02	12.48±0.02	12.26±0.02
C15247	J064005.53+092226.1	16.54±0.00	15.65±0.00	15.44±0.00	13.95±0.03	13.12±0.02	12.73±0.03
C15285	J064005.53+094554.8	18.54±0.02	18.14±0.04	17.16±0.02	15.14±0.05	14.28±0.04	13.93±0.05
C15519	J064006.00+094942.7	16.44±0.00	15.62±0.00	15.18±0.00	13.46±0.03	12.76±0.03	12.45±0.03
. . .							

Table 3. IPHAS and 2MASS photometry for known members of NGC 2264 which satisfy our selection and quality criteria (see text). The first column shows the existing object identifier as defined by Sung et al. (2008), while the second column shows the IAU-registered naming convention for objects detected by the IPHAS survey, which is formed by prefixing "IPHAS" to the position string. Calibrated IPHAS photometry is given in columns 3-5 and matched 2MASS data is given in columns 6-8. This table is available in its entirety in the online journal.

Name	$\log A_0$	$\log M_*$	$\log \tau$	$\log EW_{H\alpha}$	$\log L_{H\alpha}$	$\log \dot{M}_{acc}$	Comments
	[mag]	[$M_\odot$]	[yr]	[-Å]	[$L_\odot$]	[$M_\odot\,yr^{-1}$]	
C11059	-0.41 ± 0.44	-0.61 ± 0.08	6.91 ± 0.09	1.45 ± 0.54	-4.5 ± 0.6		
C11997	-0.39 ± 0.40	-0.49 ± 0.07	6.50 ± 0.10	1.79 ± 0.16	-3.6 ± 0.2	-9.1 ± 0.8	CTTS
C12135	-0.33 ± 0.38	-0.51 ± 0.07	6.77 ± 0.09	-0.20 ± 0.85	-5.8 ± 0.9		
C12598	0.27 ± 0.39	-0.18 ± 0.17	6.56 ± 0.31	1.59 ± 0.86	-3.1 ± 1.0	-8.5 ± 1.4	CTTS
C13507	0.19 ± 0.73	-0.34 ± 0.28	6.83 ± 0.44	1.41 ± 2.86	-3.8 ± 3.0	-8.8 ± 3.3	CTTS
C14005	-0.42 ± 0.42	-0.41 ± 0.07	6.28 ± 0.07	0.49 ± 1.13	-4.5 ± 1.2		
C15152	-0.33 ± 0.40	-0.36 ± 0.09	6.12 ± 0.08	-0.61 ± 1.18	-5.3 ± 1.2		
C15247	-0.48 ± 0.41	-0.29 ± 0.10	6.59 ± 0.19	1.00 ± 0.98	-3.9 ± 1.0		
C15285	0.21 ± 0.27	-0.32 ± 0.14	7.04 ± 0.24	-0.86 ± 0.92	-6.2 ± 0.9		
C15519	-0.36 ± 0.45	-0.32 ± 0.12	6.27 ± 0.13	-0.44 ± 1.44	-5.2 ± 1.5		
. . .							

Table 4. Posterior expectation values and standard deviations for parameters of NGC 2264 members, obtained from IPHAS and 2MASS photometry using Bayesian inference. Shown here are the first 10 entries, the table is available in its entirety in the online journal.

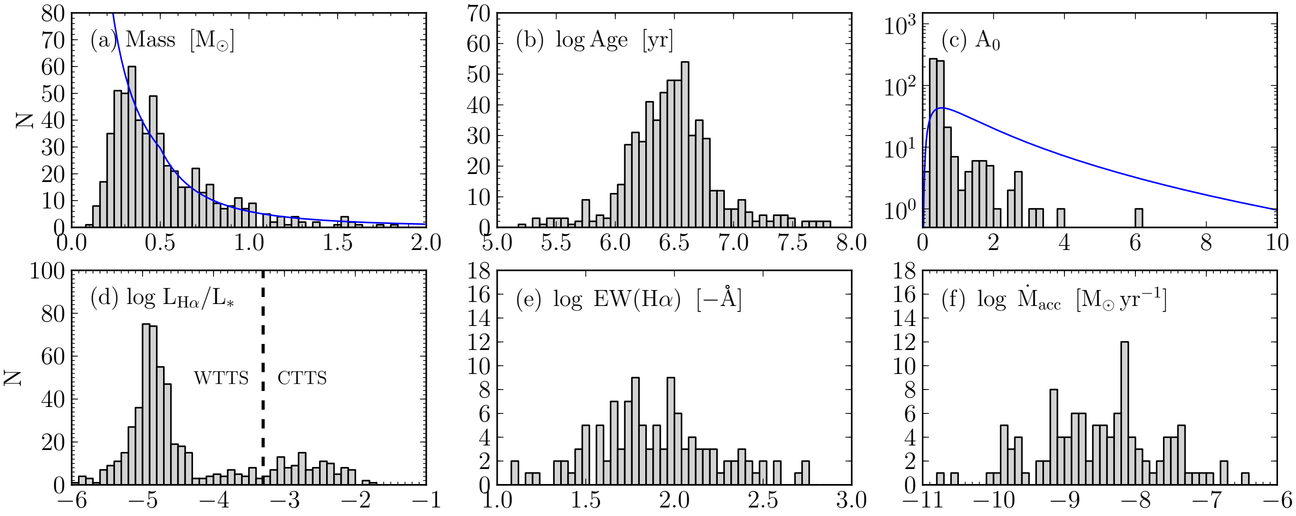

Figure 10. Distribution of inferred parameters listed in Table 4. Blue solid lines in panels (a) and (c) show the priors. We note that panels (e) and (f) only include the CTTS objects. We warn that these histograms do not reflect the underlying uncertainties and degeneracies.

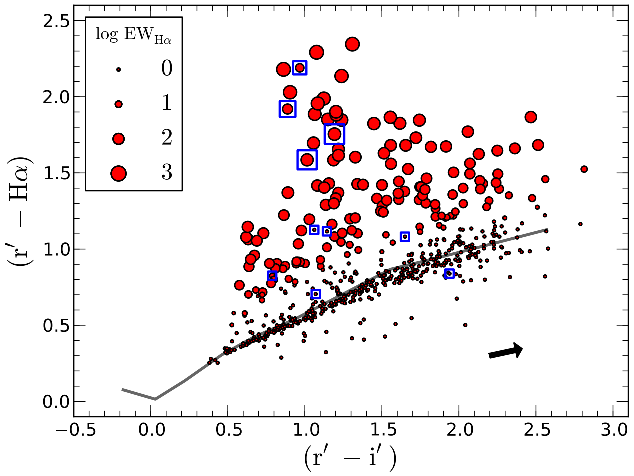

Figure 11. Position of objects in the IPHAS (r'-i')/(r'-Hα) plane. The size of the circles represent the expectation value of the Hα EW posterior. The solid line shows the unreddened main sequence while the arrow shows the typical reddening vector for a unit A$_V$, both taken from Paper I. Blue squares indicate objects with low likelihoods (cf. §5.3)

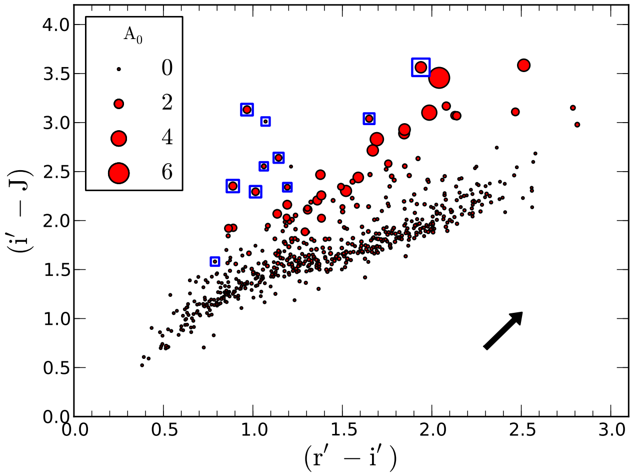

Figure 12. Position of objects in the IPHAS-2MASS (r'-i')/(i'-J) plane. The size of the circles represent the mean extinction posterior. The arrow shows the reddening vector for A$_V$ = 1 due to Schlegel et al. (1998).

5.1 Comparison with colour/magnitude diagrams

To verify that our results are consistent with those which would have been obtained using traditional plane-fitting methods, we show the position of objects as a function of their properties in three relevant colour/magnitude diagrams

First, Fig. 11 shows the (r'-i')/(r'-Hα) plane. This diagram acts mainly as an indicator for the Hα-line strength: objects with Hα in emission show greater r'-Hα values and are therefore located above the main locus. The size of the symbols represent our posterior mean estimate for EW$_{H\alpha}$. We note the good correspondence between these estimates and the distance of an object from the main locus.

Second, Fig. 12 shows the (r'-i')/(i'-J) plane. In this diagram we let the size of the symbols represent the extinction estimate, because the unit reddening vector (black arrow) follows a direction which is somewhat offset from the main locus of stars, such that objects with high extinction tend to be located above the main locus.

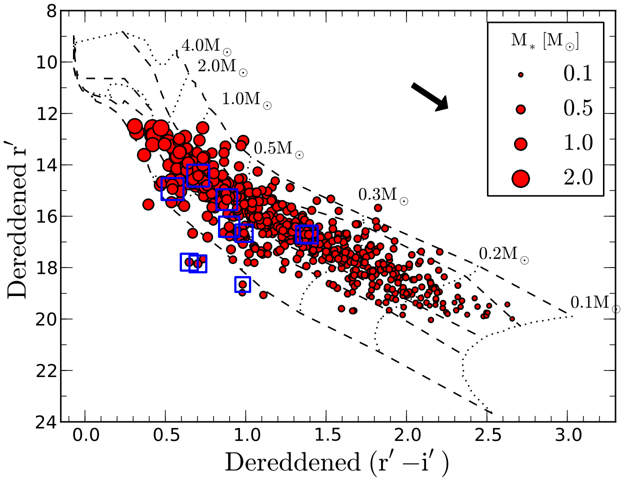

Figure 13. Position of objects in the IPHAS (r'-i')/r' plane, dereddened using the mean extinction posterior values of each object. The size of the circles represent the mean stellar mass posterior. The arrow shows the reddening vector for a unit A$_V$ due to Schlegel et al. (1998). We also show the evolutionary mass tracks (dotted lines) and isochrones (dashed lines) from the Siess et al. model, placed at the adopted distance of 760 pc. The isochrones are for 0.1, 1, 5, 10 and 100 Myr (top to bottom).

At first glance, this diagram shows a good agreement between the position of objects and their estimated extinction. However, we draw the reader's attention to the lack of a one-on-one relationship between the extinction and the apparent distance from the main locus. As discussed in §2.1.2, this is a result of the fact that the Hα-line falls inside the r'-band, which has the effect of moving emission-line objects towards the left of the diagram and above the main locus, such that strong Hα emitters with low extinction occupy the same region in the diagram as objects with weak emission and high extinction. This illustrates the fact that it is difficult to simultaneously estimate extinction and Hα emission from these diagrams, which was a major motivation for adopting the Bayesian approach (cf. §2).

Finally, Fig. 13 shows the (r'-i')/r' plane, which can be used to trace the ages and masses in a way similar to a Hertzsprung-Russell diagram. The objects shown in this plane have been dereddened according to their individual extinction estimates, such that we can compare their position against the theoretical isochrones (dashed lines) and evolutionary mass tracks (dotted lines) from the Siess et al. model. We find a good agreement between the position of objects in the diagram and their estimated masses. An equally good agreement is found for the age estimates (not shown).

5.2 Comparison with existing spectroscopy

The most comprehensive set of spectroscopy towards NGC 2264 that is currently available in the literature is the survey presented by Dahm & Simon (2005), which is based on the Wide-Field Grism Spectrograph (WFGS) on the Uni-

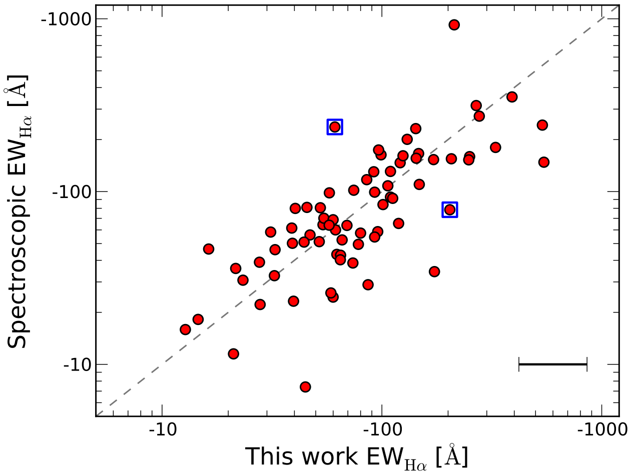

Figure 14. Comparison of our inferred Hα EWs with values obtained from grism spectroscopy by Dahm & Simon (2005). The grey dashed line shows the unity relation. The median error bar is shown in the bottom right. The scatter is thought to be due to a combination of uncertainty and natural Hα emission variability.

versity of Hawaii 2.2 m telescope on Mauna Kea, augmented with spectra from the Gemini North 8.2 m telescope.

Their dataset provides spectroscopic Hα EWs for 74 out of the 115 accreting objects in our sample. We find a good correlation between their values and our estimates (r=0.8) which is shown in Fig. 14, albeit with a significant scatter. The spread in the relationship is very similar to the one we previously found in a different cluster (Paper I), and is thought to be due to a combination of natural variability and uncertainty (the median 1-sigma uncertainty for our estimates is shown in the bottom right of the plot).

We note that the work by Dahm & Simon highlighted the strong natural variability of the Hα line. Using spectra from multiple epochs between 1990 and 2003, the authors reported that nearly all of the CTTS (90%) exhibited changes in the EW at or above the 10% level, while 57% varied at 50% or greater. This confirms that the scatter is at least in part due to natural variability.

5.3 Objects with low likelihoods

An advantage of the Bayesian method is that we may evaluate how well the model matches different objects by comparing their mean likelihood values (obtained from Eqn. 28). Using this information, we found a small number of $\sim$10 objects with significantly lower likelihoods than the main locus of stars. These 10 "worst-fit" objects have been marked by blue rectangles in Figs. 11-14 (object ID's: C13507, C22501, C27963, C28541, C31190, C31352, C33877, C36198, C36493, C37366).

The markers reveal that several of these objects show strong Hα emission in Fig. 11, while the colours appear anomalous in Fig. 12, where they fall beyond the extreme blue edge of the locus. Likewise, a few fall below the main locus in Fig. 13.

We can think of four reasons to explain their position:

(i) The objects might be blue Hα-emitters in the background, e.g. interacting binaries, Be stars or unresolved planetary nebulae (Corradi et al. 2008) not related to NGC 2264.

(ii) If the objects are genuine members, the presence of strong Hα emission suggests that they are undergoing high levels of mass accretion, which is known to be a cause for continuum veiling in the red part of the spectrum. The origin of such emission is unclear however (Fischer et al. 2011).

(iii) The objects might be affected by anomalous extinction. Three of the stars have previously been discussed by Sung et al. (2008, 2009) who classified them as "BMS" (for Below the pre-Main Sequence) based on their outlier position in the colour-magnitude diagram. Sung et al. supported the assumption that these are bona-fide young stars with a nearly edge-on disk. The dust grains in a disk tend to be larger than those in the interstellar medium, and hence the extinction law may differ significantly.

(iv) It is possible that the edge of the disk contaminates the colours of the star, depending on the inner radius, inclination and shape of the disk.

If these outlier objects are genuine members, they provide evidence that our results would profit from a more advanced pre-main sequence model which includes the effects of continuum veiling due to accretion and dust in the circumstellar environment. We will discuss this prospect in §6.

5.4 The extinction prior

We noted previously that the extinction distribution (Fig. 10, panel c) shows a deficit of highly reddened objects when compared to the prior (blue solid line). The prior was chosen based on the distribution of spectroscopy-based extinctions determined by Rebull et al. (2002) for 202 candidate members (§2).

The mismatch between our results and the prior is explained by the use of different selection criteria in both studies. On one hand, the membership selection by Rebull et al. made use of a combination of colour-magnitude diagrams and photometric variability, which is prone to the inclusion of background objects (i.e. out of the 17 objects for which the authors reported extinctions larger than $A_V > 3$, only 2 passed our membership criteria). On the other hand, the use of optical photometry and X-ray observations in our criteria is likely to introduce a bias against highly reddened objects.

The mismatch illustrates that the extinction prior does not determine the results alone, but merely helps the data to constrain the parameters using the additional knowledge which we chose to include. To understand the influence of the prior, we repeated the inference procedure using a log-uniform prior $P(\log A_0) \sim \mathcal{U}(-1, 1)$, which is less informed. We found this prior to produce near-identical results with mean $\log A_0 = -0.32$ ($A_0 = 0.48$), which differs from the original mean extinction by only +0.05 mag. The influence of this prior on the other parameters was found to be negligible, with mass estimates showing a median shift of $+0.01\,M_\odot$ and the median age remaining constant.

We also repeated the experiment using the absolute uniform prior $P(A_0) \sim \mathcal{U}(0, 10)$, which favours high extinctions far more strongly than the log-uniform prior. This was found to change the extinction of individual objects by a factor 2.1 on average, hence raising the mean extinction of the sample to $\log A_0 = -0.03$ ($A_0 = 0.93$). As a result, masses experienced a median shift of $+0.11\,\mathrm{M}_\odot$ and the median age increased from 3.0 Myr to 4.3 Myr.

The factor ∼2 corresponds to the typical 1-σ uncertainty in the extinction of highly reddened objects in our results (e.g. Fig. 6). We therefore estimate that this is the level at which our method is able to constrain the extinction in regions where no prior information is available. The ability to constrain the extinction within a factor ∼2 meets the level of accuracy we may reasonably expect from the combination of r'/i'/J magnitudes, and represents a significant improvement over the widely used practice of assuming a fixed extinction value in the absence of a spectroscopic data. Moreover, we are confident that including additional photometry at longer wavelengths (e.g. 2MASS H/K) can further strengthen our handle on the extinction in future work.

6 FUTURE EXTENSIONS

We envisage extending the method presented in this work on two fronts: (i) more comprehensive modelling of the individual stars, and (ii) modelling the global properties of the cluster.

First, in the previous section we found indications that our results would profit from a more detailed model of T Tauri stars, which should include the effects of continuum veiling due to accretion shocks, as well as the presence of dust in the circumstellar environment. We envisage employing a grid of radiation transfer models for this purpose, such as the widely used models by Robitaille et al. (2006, 2007). We note that a maximum-likelihood fitting tool already exists for these models[3]. At present the tool only links observations to 'best-fit' parameters however, and it does not reveal the full posterior. In turn, a more detailed model invites the inclusion of photometry across a wider wavelength range. We note the possibility to include U- and g-band magnitudes from the UVEX and VPHAS galactic plane surveys (Groot et al. 2009), deep JHK-magnitudes from the UKIDSS surveys (Lucas et al. 2008) and infrared photometry from space-based telescopes.

Second, in §4 we explained that our findings on the global properties of NGC 2264 must be interpreted with caution, because we did not incorporate the cluster properties as part of our probabilistic model. This would be desirable, because there are pertinent open issues in the current literature which require a careful treatment of the parameter uncertainties (these questions include the ages of clusters, their age spreads, and the dependency of accretion rates on stellar masses, e.g. see Clarke & Pringle 2006; Jeffries et al. 2011). These questions may be addressed by adding the relevant cluster parameters to the top of the hierarchical model in Fig. 3.

It is worth emphasising that the goal of understanding clusters can be regarded as the problem of finding a hierarchical model which best explains the observations. For this reason, we encourage others to adopt graphical Bayesian models as a generic framework to link observations to theory.

7 CONCLUSIONS

We showed how the theory of graphical Bayesian networks can be used to define a probabilistic model which allows extinction, age, mass and accretion rate to be inferred from IPHAS r'/Hα/i' and 2MASS J-band photometry without the need for spectroscopy. The model combined the evolutionary model due to Siess et al. (2000) and the simulated photometry for Hα emission-line stars due to Barentsen et al. (2011) to compute probabilistic posterior distributions.

Compared to more popular plane-fitting or maximum-likelihood techniques, the advantages of our approach are that (i) we dealt with the degeneracy between stellar mass and extinction by considering the full probability distribution of solutions and (ii) we obtained meaningful expectation values and uncertainties by marginalising over nuisance parameters such as the distance and disc truncation radius.

We used the Python/PyMC library to compute the model using a Markov Chain Monte Carlo (MCMC) algorithm, which was found to take only a small amount of programming effort (Appendix A). We then applied the method to 587 low-mass members of the NGC 2264 star-forming region and found a good agreement between our results and the position of stars in colour/magnitude diagrams, as well as literature spectroscopy. We performed an initial inspection of the sample and found that:

(i) NGC 2264 shows a median age of 3.0 Myr, albeit with a large apparent dispersion between 1.8 and 4.5 Myr (25 and 75% quartiles);

(ii) 115 objects (20±2%) show fractional Hα luminosities above the chromospheric saturation limit ($\log L_{H\alpha}/L_* > -3.3$; Barrado y Navascués & Martín 2003) and are therefore very likely to be CTTS objects which are accreting gas from a circumstellar disc;

(iii) for these CTTS objects, we estimated mass accretion rates on the basis of Hα luminosities and found them to range between $10^{-10.7}$ and $10^{-6.4}\,\mathrm{M}_\odot$/yr with a median of $10^{-8.4}\,\mathrm{M}_\odot$/yr.

The results were shown to be consistent with existing spectroscopic studies in the literature. Our method achieved these results with great efficiency by depending only on photometry, and provides a significant step forward from previous photometric methods because our probabilistic approach ensures that nuisance parameters, such as extinction and distance, are fully included in the analysis with a clear picture on any degeneracies.

In future work, we envisage extending the method to include more physics. We note the possibility to utilise a grid of radiation transfer models which include the effects of continuum veiling and material in the circumstellar environment. In turn, our method would benefit from the inclusion of additional photometric bands across a wider wavelength range.

Graphical Bayesian models provide a generic framework for estimating parameters from sparse data. We expect that the approach will become increasingly important as a tool for the effective utilisation of large surveys, in particular once distance estimates from Gaia can be included. We re-

[3] http://www.astro.wisc.edu/protostars

mind the reader that our source code is made available online[4] and encourage others to reuse or improve the code.

ACKNOWLEDGEMENTS

We thank Nick Wright, Eduardo Martin, Eli Bressert, Christian Knigge, Roberto Raddi, Renko and Thomas Barclay for constructive comments which helped to improve the paper. We thank David W. Hogg for getting us to adopt probabilistic inference through his blog and talks. We thank the referee, Frédéric J. Pont, for insightful corrections which strengthened the paper.

Armagh Observatory is funded by major grants from the Department of Culture, Arts and Leisure (DCAL) for Northern Ireland and the UK Science and Technology Facilities Council (STFC). The IPHAS survey is carried out at the Isaac Newton Telescope (INT), operated by the Isaac Newton Group in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias. IPHAS data were processed by the Cambridge Astronomical Survey Unit at the Institute of Astronomy in Cambridge. 2MASS is a joint project of the University of Massachusetts & IPAC/Caltech, funded by NASA/NSF.

Our work made use of TopCat (Taylor 2005), Aladin, PostgreSQL, Python (PyMC, PyFITS, APLpy modules) and Amazon EC2. Mosaics were generated using the Montage software maintained by NASA/IPAC.

This paper has been typeset from a TEX/ LATEX file prepared by the author.

REFERENCES

Bailer-Jones, C. A. L. 2009, IAU Symposium, 254, 475
Bailer-Jones, C. A. L. 2011, MNRAS , 411, 435
Barentsen, G., Vink, J. S., Drew, J. E., et al. 2011, MNRAS , 415, 103
Barentsen, G., Arlt, R., & Frohlich, H.-E. 2011, WGN, Journal of the International Meteor Organization, 39, 126
Barrado y Navascués, D., & Martín, E. L. 2003, AJ , 126, 2997
Baxter, E. J., Covey, K. R., Muench, A. A., et al. 2009, AJ , 138, 963
Bertout, C. 1989, ARA&A , 27, 351
Calvet, N., & Gullbring, E. 1998, ApJ , 509, 802
Chib, S., & Greenberg, E. 1995, American Statistical Journal, 49, 327
Clarke, C. J., & Pringle, J. E. 2006, MNRAS , 370, L10
Corradi, R. L. M., Rodríguez-Flores, E. R., Mampaso, A., et al. 2008, A&A , 480, 409
Dahm, S. E., & Simon, T. 2005, AJ , 129, 829
Dahm, S. E., Simon, T., Proszkow, E. M., & Patten, B. M. 2007, AJ , 134, 999
Dahm, S. E. 2008, Handbook of Star Forming Regions, Volume I, 966
De Marchi, G., Panagia, N., & Romaniello, M. 2010, ApJ , 715, 1
Drew, J. E., et al. 2005, MNRAS , 362, 753 guilar, A., Wang, J., & Garmire, G. P. 2009, ApJ , 699, 1454
Drew, J. E., Greimel, R., Irwin, M. J., & Sale, S. E. 2008, MNRAS , 386, 1761
Espaillat, C., Ingleby, L., Hernández, J., et al. 2012, ApJ , 747, 103
Evans, N. J., et al. 2009, ApJS , 181, 321
Fang, M., van Boekel, R., Wang, W., Carmona, A., Sicilia-Aguilar, A., & Henning, T. 2009, A&A , 504, 461
Fedele, D., van den Ancker, M. E., Henning, T., Jayawardhana, R., & Oliveira, J. M. 2010, A&A , 510, A72
Fischer, W., Edwards, S., Hillenbrand, L., & Kwan, J. 2011, ApJ , 730, 73
Flaccomio, E., Micela, G., & Sciortino, S. 2006, A&A , 455, 903
Ford, E. B. 2005, AJ , 129, 1706
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2012, arXiv:1202.3665
Gennaro, M., Prada Moroni, P. G., & Tognelli, E. 2012, MNRAS , 420, 986
González-Solares, E. A., et al. 2008, MNRAS , 388, 89
Gregory, P. C. 2005, Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with 'Mathematica' Support. Edited by P. C. Gregory. ISBN 0 521 84150 X. Cambridge University Press, Cambridge, UK, 2005.,
Groot, P. J., Verbeek, K., Greimel, R., et al. 2009, MNRAS , 399, 323
Gullbring, E., Hartmann, L., Briceno, C., & Calvet, N. 1998, ApJ , 492, 323
Gullbring, E., Calvet, N., Muzerolle, J., & Hartmann, L. 2000, ApJ , 544, 927
Haisch, K. E., Jr., Lada, E. A., & Lada, C. J. 2001, ApJL , 553, L153
Hartigan, P., Edwards, S., & Ghandour, L. 1995, ApJ , 452, 736
Hartmann, L. 2008, Accretion processes in star formation. 2nd Edition. Cambridge University Press, Cambridge, UK.
Herczeg, G. J., & Hillenbrand, L. A. 2008, ApJ , 681, 594
Irwin, M., & Lewis, J. 2001, New Astron. , 45, 105
Jeffries, R. D., Littlefair, S. P., Naylor, T., & Mayne, N. J. 2011, MNRAS , 418, 1948
Jørgensen, B. R., & Lindegren, L. 2005, A&A , 436, 127
Kenyon, S. J., & Hartmann, L. 1995, ApJS , 101, 117
Kipping, D. M., Bakos, G. Á., Buchhave, L. A., Nesvorny, D., & Schmitt, A. 2012, arXiv:1201.0752
Kroupa, P. 2001, MNRAS , 322, 231
Lucas, P. W., Hoare, M. G., Longmore, A., et al. 2008, MNRAS , 391, 136
MacKay, D., Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003
Martin, E. L. 1997, A&A , 321, 492
Meyer, M. R., Calvet, N., & Hillenbrand, L. A. 1997, AJ , 114, 288
Mohanty, S., Jayawardhana, R., & Basri, G. 2005, ApJ , 626, 498
Muzerolle, J., Hillenbrand, L., Calvet, N., Briceño, C., & Hartmann, L. 2003, ApJ , 592, 266
Najita, J. R., Strom, S. E., & Muzerolle, J. 2007, MNRAS , 378, 369
Natta, A., Testi, L., Muzerolle, J., Randich, S., Comerón,

[4] https://github.com/barentsen

F., & Persi, P. 2004, A&A , 424, 603
Natta, A., Testi, L., & Randich, S. 2006, A&A , 452, 245
Owen, J. E., Ercolano, B., & Clarke, C. J. 2011, MNRAS , 412, 13
Park, B.-G., Sung, H., Bessell, M. S., & Kang, Y. H. 2000, AJ , 120, 894
Patil, A., Huard D., Fonnesbeck C. J. 2010, Journal of Statistical Software, 35, 4, pp.1–81
Pont, F., & Eyer, L. 2004, MNRAS , 351, 487
Rebull, L. M., Makidon, R. B., Strom, S. E., et al. 2002, AJ , 123, 1528
Reipurth, B., Pettersson, B., Armond, T., Bally, J., & Vaz, L. P. R. 2004, AJ , 127, 1117
Rizzuto, A. C., Ireland, M. J., & Robertson, J. G. 2011, MNRAS , 416, 3108
Robitaille, T. P., Whitney, B. A., Indebetouw, R., Wood, K., & Denzmore, P. 2006, ApJS , 167, 256
Robitaille, T. P., Whitney, B. A., Indebetouw, R., & Wood, K. 2007, ApJS , 169, 328
Russell, S. J. & Norvig, P., 2009, Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River, NJ
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ , 500, 525
Sicilia-Aguilar, A., Henning, T., & Hartmann, L. W. 2010, ApJ , 710, 597
Siess, L., Dufour, E., Forestini, M. 2000, A&A , 358, 593
Skrutskie, M. F., et al. 2006, AJ , 131, 1163
Spezzi, L., de Marchi, G., Panagia, N., Sicilia-Aguilar, A., & Ercolano, B. 2012, MNRAS , 421, 78
Sung, H., Bessell, M. S., & Lee, S.-W. 1997, AJ , 114, 2644
Sung, H., Bessell, M. S., & Chun, M.-Y. 2004, AJ , 128, 1684
Sung, H., Bessell, M. S., Chun, M.-Y., Karimov, R., & Ibrahimov, M. 2008, AJ , 135, 441
Sung, H., Stauffer, J. R., & Bessell, M. S. 2009, AJ , 138, 1116
Taylor, M. B. 2005, Astronomical Data Analysis Software and Systems XIV, 347, 29
Trotta, R. 2008, Contemporary Physics, 49, 71
Walker, M. F. 1956, ApJS , 2, 365
Williams, J. P., & Cieza, L. A. 2011, ARA&A , 49, 67

APPENDIX A: PYTHON SOURCE CODE FOR THE INFERENCE MODEL

```python
import numpy as np
from scipy.interpolate.rbf import Rbf
import pyfits
import pymc

""" Interpolation functions for intrinsic magnitudes """
siess = pyfits.getdata("siess_isochrones.fits", 1)  # Siess et al. (2000)
# Interpolation is performed using linear Radial Basis Functions
siess_Mr = Rbf(siess.field("logMass"), siess.field("logAge"),
               siess.field("Mr_iphas"), function="linear")
siess_Mi = Rbf(siess.field("logMass"), siess.field("logAge"),
               siess.field("Mi_iphas"), function="linear")
siess_Mj = Rbf(siess.field("logMass"), siess.field("logAge"),
               siess.field("Mj"), function="linear")
siess_logR = Rbf(siess.field("logMass"), siess.field("logAge"),
                 siess.field("logRadius"), function="linear")

""" Functions for magnitude offsets due to emission & exactinction """
sim = pyfits.getdata("simulated_iphas_colours_barentsen2011.fits", 1)  # PaperI
# Functions for r’/Ha/i’ offsets as a function of colour, extinction and EW
r_offset = Rbf(sim.field("ri_unred"), sim.field("av"), sim.field("logew"),
               sim.field("d_r"), function="linear")
ha_offset = Rbf(sim.field("ri_unred"), sim.field("av"), sim.field("logew"),
                sim.field("d_ha"), function="linear")
i_offset = Rbf(sim.field("ri_unred"), sim.field("av"), sim.field("logew"),
               sim.field("d_i"), function="linear")
# Intrinsic (r’-Ha) colour as a function of intrinsic (r’-i’)
intrinsic = (sim.field("av") == 0) & (sim.field("logew") == -1)
rminHa_intrinsic = Rbf(sim.field("ri_unred")[intrinsic],
                       sim.field("rha")[intrinsic], function="linear")


def make_model(observed_sed, e_observed_sed):
    """ This function returns all prior and likelihood objects """

    # Prior: mass (Kroupa 2001)
    @pymc.stochastic()
    def logM(value=np.array([np.log10(0.5)]), a=np.log10(0.1), b=np.log10(7)):

        def logp(value, a, b):
            if value > b or value < a:
                return -np.Inf  # Stay within the model limits (a,b).
            else:
                mass = 10 ** value
                if mass < 0.5:
                    return np.log(mass ** -1.3)  # Kroupa (2001)
                else:
                    return np.log(0.5 * mass ** -2.3)  # Kroupa (2001)

        def random(a, b):
            val = (b - a) * np.random.rand() + a
            return np.array([val])

    # Prior: age (uniform in the logarithm)
    logT = pymc.Uniform("logT", np.array([5]), np.array([8]))

    # Prior: accretion rate (uniform in the logarithm)
    logMacc = pymc.Uniform("logMacc", np.array([-15]), np.array([-2]))

    # Prior: disc truncation radius (Rin = 5 +\- 2 R, Gullbring et al. 1998)
    Rin = pymc.TruncatedNormal("Rin", mu=np.array([5.0]), tau=2.0 ** -2,
                               a=1.01, b=9e99)

    # Prior: distance (d = 760 +\- 5 pc, Sung 1997)
    d = pymc.TruncatedNormal("d", mu=np.array([760.0]), tau=5.0 ** -2,
                             a=700, b=9e99)
    # Prior: extinction (logA0 = -0.27 +/- 0.46, Rebull et al. 2002)
    logA0 = pymc.Normal("logA0", mu=np.array([-0.27]), tau=0.46 ** -2)

    # Likelihood: intrinsic SED
    @pymc.deterministic()
    def SED_intrinsic(logM=logM, logT=logT):
        r = siess_Mr(logM, logT)  # IPHAS r' as a function of (mass, age)
        i = siess_Mi(logM, logT)  # IPHAS i
        j = siess_Mj(logM, logT)  # 2MASS J
        ha = r - rminHa_intrinsic(r - i)  # IPHAS H-alpha
        return np.array([r[0], ha[0], i[0], j[0]])

    # Likelihood: H-alpha excess luminosity
    @pymc.deterministic()
    def logLacc(logM=logM, logT=logT, logMacc=logMacc, Rin=Rin):
        logR = siess_logR(logM, logT)  # Radius as a function of (mass, age)
        return 7.496 + logM + logMacc - logR + np.log10(1 - (1 / Rin))
    logLha = pymc.Normal("logLha", mu=(0.64 * logLacc - 2.12), tau=0.43 ** -2)

    # Likelihood: H-alpha equivalent width (EW).
    @pymc.deterministic()
    def logEW(logLha=logLha, SED_intrinsic=SED_intrinsic):
        Lha = 10 ** logLha  # Excess luminosity
        Lha_con = 0.316 * 10 ** (-0.4 * (SED_intrinsic[1] + 0.03))  # Continuum
        ew = -95.0 * Lha / Lha_con  # Equivalent width.
        return np.log10(-ew)

    # Likelihood: apparent SED
    @pymc.deterministic()
    def SED_apparent(d=d, logA0=logA0, SED_intr=SED_intrinsic, logEW=logEW):
        dismod = 5.0 * np.log10(d) - 5.0  # Distance modulus.
        A0 = 10.0 ** logA0  # Extinction parameter
        ri_intr = np.array([SED_intr[0] - SED_intr[2]])  # Intrinsic (r'-i')
        # Correct the intrinsic magnitudes for extinction and H-alpha emission:
        r = SED_intr[0] + dismod + r_offset(ri_intr, A0, logEW)
        ha = SED_intr[1] + dismod + ha_offset(ri_intr, A0, logEW)
        i = SED_intr[2] + dismod + i_offset(ri_intr, A0, logEW)
        j = SED_intr[3] + dismod + 0.276 * A0
        return np.array([r[0], ha[0], i[0], j[0]])

    # Likelihood: observed SED
    @pymc.stochastic(observed=True)
    def SED_observed(value=observed_sed, SED_apparent=SED_apparent):
        e_calib = np.array([0.1, 0.1, 0.1, 0.1])  # Absolute uncertainty term
        D2 = sum((observed_sed - SED_apparent) ** 2 /
                 (e_observed_sed ** 2 + e_calib ** 2))
        logp = -D2 / 2.0
        return logp

    return locals()  # Return all model components defined above


if __name__ == "__main__":
    """ Example code which demonstrates how to obtain the posterior mass """
    # Input: the observed magnitudes and 1-sigma uncertainties
    sed_observed = np.array([19.41, 18.14, 17.56, 15.44])  # r, Ha, i, J
    e_sed_observed = np.array([0.03, 0.03, 0.02, 0.06])  # e_r, e_Ha, e_i, e_J
    # Initialize the model.
    mymodel = make_model(sed_observed, e_sed_observed)
    M = pymc.MCMC(mymodel)
    # Demo: run the MCMC sampler and print the expectation value for log(Mass)
    M.sample(50000)
    samples_logM = M.trace("logM")[:]
    print "logM = %.2f +/-%.2f" % (np.mean(samples_logM), np.std(samples_logM))
```