Sequential Monte Carlo for Detecting and Deblending Objects in Astronomical Images

Tim White
Department of Statistics
University of Michigan
Ann Arbor, MI
twhit@umich.edu

Jeffrey Regier
Department of Statistics
University of Michigan
Ann Arbor, MI
regier@umich.edu

Abstract

Many of the objects imaged by the forthcoming generation of astronomical surveys will overlap visually. These objects are known as blends. Distinguishing and characterizing blended light sources is a challenging task, as there is inherent ambiguity in the type, position, and properties of each source. We propose SMC-Deblender, a novel approach to probabilistic astronomical cataloging based on sequential Monte Carlo (SMC). Given an image, SMC-Deblender evaluates catalogs with various source counts by partitioning the SMC particles into blocks. With this technique, we demonstrate that SMC can be a viable alternative to existing deblending methods based on Markov chain Monte Carlo and variational inference. In experiments with ambiguous synthetic images of crowded starfields, SMC-Deblender accurately detects and deblends sources, a task which proves infeasible for Source Extractor, a widely used non-probabilistic cataloging program.

1 Introduction

The properties of stars and galaxies are of great scientific importance [1, 2]. Astronomical cataloging, which is the task of inferring these properties from astronomical images, is difficult for images that contain many visually overlapping sources, known as blends [3]. A major challenge of deblending sources in such an image is that the number of sources is uncertain; one must infer whether the intensities of a set of pixels were produced by a single bright source or several dimmer sources. Consequently, the properties of the overlapping sources are also ambiguous. Blending is poised to become an even more prevalent issue in the near future; for example, it is estimated that 62% of galaxies detected by the forthcoming Legacy Survey of Space and Time (LSST) will be blended [4].

Probabilistic cataloging addresses the challenges posed by blends [5, 6, 7, 8]. Whereas traditional astronomical cataloging programs supply single-catalog point estimates that fail to incorporate the uncertainty inherent in ambiguous images [9, 10], probabilistic methods infer a posterior distribution over all possible catalogs, assigning a higher posterior probability to catalogs that are more likely to have yielded a particular image. Point estimates derived from this posterior distribution are endowed with an estimate of uncertainty that reflects the ambiguity of the original image.

One approach to probabilistic cataloging is to sample catalogs from the posterior distribution using Markov chain Monte Carlo (MCMC). However, constructing a viable MCMC sampler for this task is challenging. The number of light sources in a particular image is unknown and each potential source has an associated set of properties, so any MCMC-based cataloging algorithm must propagate catalogs across dimensions using a method like reversible jump MCMC [5] or diffusive nested sampling [6]. Transdimensional MCMC is notoriously difficult to implement and may be inefficient if the dimension-jumping proposals are poorly designed [11].

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

Recently, Liu et al. proposed an alternative method of probabilistic cataloging that utilizes modern techniques in variational inference (VI) [8]. Instead of sampling catalogs from the exact posterior distribution, they posit an approximate form for the posterior and optimize the parameters of this variational approximation. This approach avoids transdimensional sampling and is orders of magnitude faster than MCMC thanks to recent advances in stochastic optimization and amortization. However, VI involves the demanding task of nonconvex optimization, and even if this optimization is successful, the quality of the resulting posterior approximation may be difficult to ascertain.

We propose SMC-Deblender, a novel probabilistic cataloging technique based on sequential Monte Carlo (SMC) [12, 13] rather than MCMC or VI. SMC-Deblender is an efficient method for sampling catalogs from the posterior distribution. It avoids transdimensional sampling by partitioning the SMC particles into dimension-specific blocks, a strategy that is well-suited for the modern age of GPU computing. SMC-Deblender accurately deblends light sources in highly ambiguous synthetic images, a task that is demonstrated to be infeasible for Source Extractor, one of the most widely used non-probabilistic cataloging programs.

2 Statistical Model

We consider a generative model for small images depicting dense clusters of stars. It may be instructive to think of each of these small images as an ambiguous section of a larger image. Let x be an image with a height of H pixels and a width of W pixels; in our experiments, we set H and W to 15. Let $s \sim \mathcal{F}_1$ denote the number of stars in the image, where $\mathcal{F}_1$ is a generic distribution. The number of stars in an image is commonly modeled using a Poisson distribution. Here, however, we set $\mathcal{F}_1$ to be Uniform$\{0, 1, 2, ..., D\}$ for some maximum number of stars D; this choice ensures that we generate images with a variety of star densities in our experiments.

Each star in the image has a location and a flux (i.e., brightness). Given s, we sample locations $u_1, u_2, ..., u_s \overset{\text{iid}}{\sim} \mathcal{F}_2$ and fluxes $f_1, f_2, ..., f_s \overset{\text{iid}}{\sim} \mathcal{F}_3$. In our experiments, $\mathcal{F}_2$ is uniform across the $H \times W$ image, but one could define $\mathcal{F}_2$ differently (e.g., truncated Gaussian). For simplicity, we set $\mathcal{F}_3$ to $\mathcal{N}(\mu, \sigma^2)$, where $\mu = 10 f_{\text{min}}$ and $\sigma = 2 f_{\text{min}}$ for a positive baseline flux f_{min}; however, the fluxes would also be well-modeled by a truncated power-law distribution. If $s = 0$, no locations or fluxes are sampled and the image x is just background noise.

Let $z = \{s, \{u_j\}_{j=1}^s, \{f_j\}_{j=1}^s\}$ denote a catalog, i.e., a collection of latent variables that describes the imaged stars. Given a catalog z, the intensity of the image at pixel (h, w) is $x_{hw} \mid z \sim \mathcal{F}_4$. We set $\mathcal{F}_4$ to Poisson$(\lambda_{hw}(z) + \gamma)$, where γ is the background intensity of the image and $\lambda_{hw}(z) = \sum_{j=1}^s f_j \psi((h, w) - u_j)$. The function $\psi(\cdot)$ is a point spread function (PSF), a deterministic function that describes the appearance of a star at each pixel. We use a bivariate Gaussian density as the PSF.

3 Sequential Monte Carlo

Given an image x, we use sequential Monte Carlo (SMC) to characterize the posterior distribution of possible catalogs $p(z|x)$ by generating a set of weighted samples, or particles, from $p(z|x)$ [12, 13]. SMC constructs a weighted particle approximation of an intractable target distribution by sampling particles from a tractable initial proposal, propagating them through a sequence of intermediate distributions, and updating their weights accordingly. At each step along this sequence, the particles are resampled to prevent the weight of any one particle from dominating the others.

We propose a convenient SMC-based method for comparing catalogs that contain different numbers of light sources. By allocating an equal number of particles to each candidate source count s and preserving this block structure during resampling and propagation, we can use the weighted particles returned in the final iteration to assess the posterior probabilities of catalogs across values of s. This approach delays transdimensional inference until the final iteration and thus avoids transdimensional sampling altogether, a favorable property that is not shared by alternatives such as reversible jump MCMC [14]. We formalize this novel procedure, which we call SMC-Deblender, in Algorithm 1.

SMC-Deblender uses likelihood tempering to impose a sequential structure that links the prior $p(z)$ to the intractable posterior $p(z|x) \propto p(z)p(x|z)$. The intermediate targets $p(z)p(x|z)^{\tau_t}$ are constructed according to a temperature schedule $0 = \tau_0 < \cdots < \tau_T = 1$, which can be prespecified or determined adaptively. In iteration t of SMC, we employ the adaptive scheme proposed in [12]

Algorithm 1 Likelihood-tempered sequential Monte Carlo with block resampling (SMC-Deblender)

Input: Image x; likelihood $p(x|z)$; prior $p(z)$; method to construct invariant kernel $M_\tau(\cdot|\cdot)$ for $\tau \in [0,1]$; number of blocks B; number of particles per block N; minimum effective sample size $\mathrm{ESS}_{\min}$.

$$\left.\begin{array}{l}
\text{Step } t \leftarrow 0. \text{ Temperature } \tau_t \leftarrow 0. \\
\text{Particles } z_{bk}^{(t)} \overset{\text{iid}}{\sim} p(z) \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\} \quad (\forall b,\ s_{b1}^{(t)} = \cdots = s_{bN}^{(t)}). \\
\text{Unnormalized weights } w_{bk}^{(t)} \leftarrow 1 \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\}. \\
\text{Inter-block normalized weights } W_{bk}^{(t)} \leftarrow \frac{1}{NB} \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\}. \\
\text{Intra-block normalized weights } \widetilde{W}_{bk}^{(t)} \leftarrow \frac{1}{N} \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\}. \\
\text{Intra-block effective sample size } \mathrm{ESS}_b^{(t)} \leftarrow \big(\sum_k (\widetilde{W}_{bk}^{(t)})^2\big)^{-1} \text{ for } b \in \{1, \dots, B\}.
\end{array}\right\} \text{INITIALIZE}$$

while $\tau_t < 1$ **do**

$t \leftarrow t + 1$. Update $\tau_t \leftarrow \tau_{t-1} + \delta$, where $\delta \in [0, 1-\tau_{t-1}]$ (see section 3 of text).

$$\left.\begin{array}{l}
\textbf{for} \text{ block } b \in \{1, \dots, B\} \textbf{ do} \\
\quad \textbf{if } \mathrm{ESS}_b^{(t-1)} < \mathrm{ESS}_{\min} \textbf{ then} \\
\quad\quad \text{Resample } \{z_{bk}^{(t-1)}\}_{k=1}^N \text{ using } \{\widetilde{W}_{bk}^{(t-1)}\}_{k=1}^N. \\
\quad\quad \text{Reset } W_{bk}^{(t-1)} \leftarrow \frac{1}{N} \sum_l W_{bl}^{(t-1)} \text{ and } \widetilde{W}_{bk}^{(t-1)} \leftarrow \frac{1}{N} \text{ for } k \in \{1, \dots, N\}.
\end{array}\right\} \text{RESAMPLE}$$

$$\left.\text{Sample } z_{bk}^{(t)} \sim M_{\tau_{t-1}}(\cdot|z_{bk}^{(t-1)}) \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\} \quad (\forall b, k,\ s_{bk}^{(t)} = s_{bk}^{(t-1)}).\right\} \text{PROPAGATE}$$

$$\left.\begin{array}{l}
\text{Compute } w_{bk}^{(t)} \leftarrow W_{bk}^{(t-1)} p(x|z_{bk}^{(t)})^{\tau_t - \tau_{t-1}} \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\}. \\
\text{Compute } W_{bk}^{(t)} \leftarrow w_{bk}^{(t)} / (\sum_b \sum_l w_{bl}^{(t)}) \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\}. \\
\text{Compute } \widetilde{W}_{bk}^{(t)} \leftarrow w_{bk}^{(t)} / (\sum_l w_{bl}^{(t)}) \text{ for } b \in \{1, \dots, B\}, k \in \{1, \dots, N\}. \\
\text{Compute } \mathrm{ESS}_b^{(t)} \leftarrow \big(\sum_k (\widetilde{W}_{bk}^{(t)})^2\big)^{-1} \text{ for } b \in \{1, \dots, B\}.
\end{array}\right\} \text{UPDATE WEIGHTS}$$

Output: Weighted particle set $\{\{W_{bk}^{(t)}, z_{bk}^{(t)}\}_{k=1}^N\}_{b=1}^B$.

to obtain a tempering change $\delta_b \in [0, 1-\tau_{t-1}]$ for each particle block, and we take the minimum of these values as the global δ.

The number of blocks B is a design choice. It should be as large as the maximum anticipated number of light sources in an image (plus one, to allow s=0), but not much larger — without parallelization, the runtime of the algorithm increases with B for a fixed number of particles per block N. Resampling is performed based on the intra-block normalized weights. We use adaptive stratified resampling to reduce the variability introduced by resampling; that is, we resample the particles in block b only if their effective sample size ESS_b falls below a threshold $\mathrm{ESS}_{\min}$, which we set to $N/2$.

In the propagation step, the (possibly) resampled particles are passed to a Markov kernel $M_{\tau_{t-1}}$ that is invariant under the previous target. Following [12], we define $M_{\tau_{t-1}}$ to be a Metropolis-Hastings (MH) kernel comprising a truncated Gaussian random walk for the locations and a Gaussian random walk for the fluxes, with s unchanged for each particle. After propagation, the particle weights are recalculated by applying an incremental update to the previous inter-block normalized weights. When the temperature reaches one, the algorithm returns a set of catalogs sampled from the posterior distribution along with their corresponding inter-block weights.

4 Experiments

We evaluate SMC-Deblender using 1,000 synthetic images generated from the model in section 2, each containing up to 10 light sources. To reflect that this upper bound is unknown in practice, we use 13 blocks of 500 particles, thus allowing SMC-Deblender to detect as many as 12 light sources in an image. We compare the performance of our method to SEP, a Python interface to Source Extractor, which is a popular tool for analyzing astronomical images [9, 15]. The experiments were implemented in PyTorch and run on one NVIDIA GeForce RTX 2080 Ti GPU.[1]

Figure 1 displays five example images from our data set, as well as the reconstructions produced by SMC-Deblender and SEP. Figure 2 reports the classification accuracy, calibration, and mean absolute error of the two methods' estimates of the true source count s across starfields of various densities.

SMC-Deblender yields correct point estimates of the true source count in 765 of the 1,000 images (76.5%), with a mean absolute error of 0.267. It requires 60 seconds per image, on average. Our

[1] Code is available at `https://github.com/timwhite0/smcdeblender`.

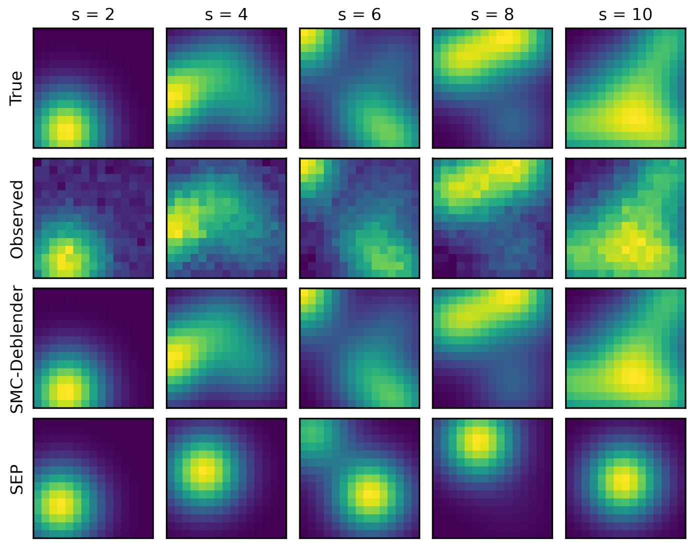

Figure 1: Reconstructions of five synthetic images with different source counts. SMC-Deblender's reconstructions are based on the SMC particle with the highest weight. SEP's reconstructions are obtained by plugging its point estimates into the likelihood function from section 2.

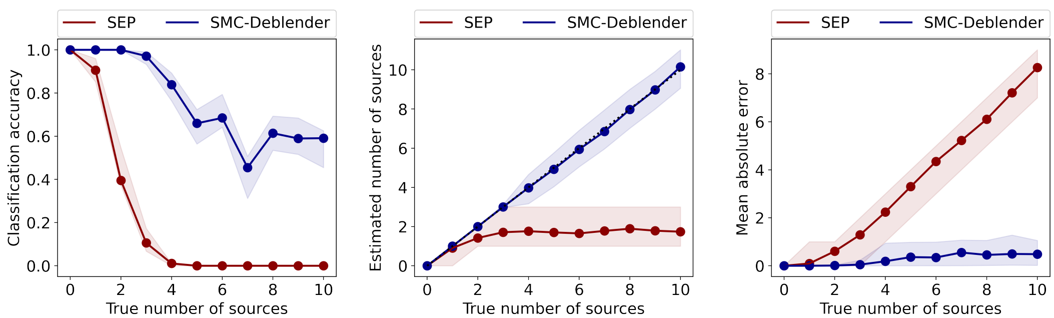

Figure 2: Accuracy, calibration, and mean absolute error of estimated source counts. Error bands indicate middle 90% of bootstrap estimates in left panel and middle 90% of images in center and right panels.

method is also well-calibrated, and it attaches greater uncertainty to its point estimates in more crowded images. In contrast, SEP struggles to separate blends of more than two stars despite extensive tuning, often identifying one or two bright sources and failing to characterize nuanced blends of fainter sources. SEP identifies the correct s in only 209 of the 1,000 images (20.9%), with a mean absolute error of 3.542.

5 Discussion

We have demonstrated the feasibility of sequential Monte Carlo as a probabilistic cataloging method. Our proposed algorithm, SMC-Deblender, provides accurate estimated source counts and calibrated uncertainty assessments in highly ambiguous images. SMC-Deblender outperforms the popular non-probabilistic cataloging tool SEP by a considerable margin, and its design suggests that it has advantages over MCMC and VI. A direct comparison between SMC-Deblender, MCMC-based methods [5, 6, 7], and VI-based methods [8] would be one immediate extension of our work.

Another avenue of further investigation is scalability. The block structure of SMC-Deblender may offer opportunities for parallelization. We also theorize that customization of the propagation kernel — e.g, using delayed-acceptance [16] or normalizing flows [17] — could allow our method to scale more easily to larger images. Finally, it is of interest to adapt our method to other image analysis tasks in astronomy, such as the detection and characterization of gravitational lensing.

Acknowledgements

This manuscript is based upon work supported by the National Science Foundation (OAC-2209720).

References

[1] Timothy M.C. Abbott, Michel Aguena, Alex Alarcon, et al. Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2):023520, 2022.

[2] Gregory M. Green, Edward Schlafly, Catherine Zucker, Joshua S. Speagle, and Douglas Finkbeiner. A 3D dust map based on Gaia, Pan-STARRS 1, and 2MASS. *The Astrophysical Journal*, 887(1):93, 2019.

[3] Peter Melchior, Rémy Joseph, Javier Sanchez, Niall MacCrann, and Daniel Gruen. The challenge of blending in large sky surveys. *Nature Reviews Physics*, 3(10):712–718, 2021.

[4] Javier Sanchez, Ismael Mendoza, David P. Kirkby, and Patricia R. Burchat. Effects of overlapping sources on cosmic shear estimation: Statistical sensitivity and pixel-noise bias. *Journal of Cosmology and Astroparticle Physics*, 2021(07):043, 2021.

[5] Richard M. Feder, Stephen K.N. Portillo, Tansu Daylan, and Douglas Finkbeiner. Multiband probabilistic cataloging: a joint fitting approach to point-source detection and deblending. *The Astronomical Journal*, 159(4):163, 2020.

[6] Brendon J. Brewer, Daniel Foreman-Mackey, and David W. Hogg. Probabilistic catalogs for crowded stellar fields. *The Astronomical Journal*, 146(1):7, 2013.

[7] James J. Buchanan, Michael D. Schneider, Kerianne Pruett, and Robert E. Armstrong. Markov chain Monte Carlo for Bayesian parametric galaxy modeling in LSST, 2023. arXiv:2309.10321.

[8] Runjing Liu, Jon D. McAuliffe, Jeffrey Regier, and the LSST Dark Energy Science Collaboration. Variational inference for deblending crowded starfields. *Journal of Machine Learning Research*, 24(179):1–36, 2023.

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

[10] Robert H. Lupton, Zeljko Ivezic, James E. Gunn, et al. SDSS image processing II: The *Photo* pipelines. Technical report, Princeton University, 2005. Available at `https://www.astro.princeton.edu/~rhl/photo-lite.pdf`.

[11] Yanan Fan and Scott A. Sisson. Reversible jump MCMC. In *Handbook of Markov Chain Monte Carlo*, pages 67–92. Chapman and Hall/CRC, 2011.

[12] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer, New York, 2020.

[13] Christian A. Naesseth, Fredrik Lindsten, and Thomas B. Schön. Elements of Sequential Monte Carlo. *Foundations and Trends in Machine Learning*, 12(3):307–392, 2019.

[14] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[15] Kyle Barbary. SEP: Source Extractor as a library. *Journal of Open Source Software*, 1(6):58, 2016.

[16] Joshua J. Bon, Anthony Lee, and Christopher Drovandi. Accelerating sequential Monte Carlo with surrogate likelihoods. *Statistics and Computing*, 31(5):1–26, 2021.

[17] Alex Matthews, Michael Arbel, Danilo Jimenez Rezende, and Arnaud Doucet. Continual repeated annealed flow transport Monte Carlo. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 15196–15219. PMLR, 2022.