CUDAHM: GPU-Accelerated Bayesian Inference for Single-Plate Hierarchical Models

János M. Szalai-Gindl*
Department of Physics of Complex Systems,
Eötvös Loránd University
and
Department of Applied Mathematics & Statistics,
The Johns Hopkins University
and
Tamás Budavári
Department of Applied Mathematics & Statistics,
The Johns Hopkins University
and
Thomas J. Loredo
Cornell Center for Astrophysics & Planetary Science,
Cornell University
and
Brandon C. Kelly
Department of Physics,
Broida Hall, University of California
and
István Csabai
Department of Physics of Complex Systems,
Eötvös Loránd University
and
László Dobos
Department of Physics of Complex Systems,
Eötvös Loránd University

January 18, 2017

Abstract

• • •*[Omit bracketed text if the abstract really needs to be < 200 words; check recent issues about this.]*• • •
We describe CUDAHM, a C++ framework for hierarchical Bayesian inference with single-plate parametric graphical models that uses graphics processing units (GPUs) to accelerate computations, enabling deployment on very large datasets. CUDAHM exploits conditional independence between instances of a plate, which enables massively parallel exploration of the replication parameter space using the single instruction, multiple data (SIMD) architecture of GPUs. It provides support for constructing Metropolis-within-Gibbs samplers that iterate between GPU-acclerated robust adaptive Metropolis (RAM) sampling of plate-level parameters conditional on upper-level parameter values, and Metropolis-Hastings sampling of upper-level parameters on the host central processing unit (CPU) conditional on the GPU results. The GPU computations are implemented using the Compute Unified Device Architecture (CUDA). CUDAHM is motivated by measurement error problems in astronomy, where density estimation and linear and nonlinear regression problems must be addressed for populations of thousands to millions of objects whose features are measured with possibly complex uncertainties. *[We briefly describe an example of regression with measurement error: inferring the distribution of properties of dust throughout a star-forming region from noisy observations of the infrared light in multiple passbands (filter colors). This is a problem where the likelihood functions for the plate-level parameters are complex; CUDAHM duplicates an earlier analysis with a speedup* $\sim \times X$*. We also use CUDAHM for luminosity function estimation, a widespread problem in astronomy that requires density deconvolution (demixing) using noisy data subject to thinning or truncation.]* We demonstrate accurate GPU-accelerated parametric conditional density deconvolution for simulated populations of 10^6 objects in about two hours using a single NVIDIA Tesla K40c GPU.

Keywords: Hierarchical Bayesian models, Metropolis-within-Gibbs sampling, parallel computing, astrostatistics, graphical processing units (GPUs)

[*] *This work was supported by the Hungarian Scientific Research Fund via grant OTKA NN 114560. Budaváři, Kelly, and Loredo gratefully acknowledge the NSF-funded Statistical and Applied Mathematical Sciences Institute (SAMSI) for support for visits to SAMSI, where this project originated. Loredo's effort was additionally supported by NSF grant AST-1312903.*

1 Introduction

Bayesian inference with graphical models has rapidly grown in popularity and sophistication since the emergence of Markov chain Monte Carlo (MCMC) algorithms for Bayesian computation nearly three decades ago. The work we report here focuses on models with classic, simple graphical structures—directed acyclic graphs (DAGs) with a single plate, i.e., a single level of replication of random variables at the lower level of a hierarchical model. Our work aims to extend the range of application of Bayesian graphical modeling in the direction of increased dataset size, rather than in the direction of increased graphical complexity.

We are motivated by measurement error problems in astronomy: density estimation with measurement error (density deconvolution, or demixing, often of a *conditional* density), and linear and nonlinear regression with measurement errors in both predictors and response. Hierarchical Bayesian modeling is well-suited to such problems, but is relatively new in astronomy (see Loredo 2013 for a recent survey). Although some recent astrostatistical research develops models with rich graphical structure, most astronomers are unfamiliar with hierarchical modeling, and models with simple graphical structure can provide new capability in many areas of astronomy, including basic two-level hierarchical models.[1] But dataset size can be an obstacle to use of such models. Large-scale, automated surveys are providing astronomers with increasingly large datasets for demographic studies of cosmic populations (e.g., categories of stars, galaxies, and planets). Current and emerging surveys are providing measurements for populations with sizes ranging from tens of thousands to 10^8 or even larger. For datasets of these scales, exploration or integration over the latent variables specifying imprecisely measured characteristics of objects in a population can be prohibitive, even for simple models with univariate member characteristics. Yet as population size grows, it becomes increasingly important to account for uncertainty in such latent variables. For example, it is well known that regression and density estimators that ignore measurement error are typically inconsistent, with the ratio of bias to reported pre-

[1] We follow the convention of naming hierarchical DAGs by the number of levels with uncertain nodes, e.g., the number of open nodes in Fig. 2.1, which depicts a model with three levels of random variables, but with the data variables (bottom level) known, i.e., to be conditioned on.

cision growing with sample size (Carroll et al. 2006). Single-plate hierarchical models can account for measurement error in many astronomically interesting scenarios, provided the implementation enables efficient computation for relevant dataset sizes.

In the following section (Sec. 2), we describe the architecture of the CUDAHM framework, which is motivated by a common computational structure underlying example hierarchical models arising in measurement error problems in astronomy. • • •*[Perhaps add a brief section (§ 3) treating Brandon's nonlinear regression with measurement error problem; perhaps he'd be willing to write it; a very brief summary of it is in § 2 and may suffice, with Brandon's example presented in the followup A&C paper.]*• • • In Sec. 4, we describe a common astronomical data analysis problem: inferring the luminosity distribution of a class of objects from distance and flux observations of a sample subject to selection effects and measurement error. Such problems may be modeled using latent, thinned marked point processes observed with measurement error; we show that the resulting likelihood function can be cast in a form mirroring the computational structure of single-plate graphical models, enabling implementation with CUDAHM. We present tests of such an implementation, using simulated data, in Sec. 5. Sec. 6 provides a summary and plans for future work. In the supplementary material, Sec. 7 presents an approximation technique for the computation of an integral appearing in the luminosity function example.

2 CUDAHM motivation and design

2.1 Motivating problem structures

Suppose we observe N members of a large population, with the observed members indexed by $i = 1$ to N. Each object (member) has a property or properties ψ_i (a vector in multiple-property cases); we are interested in estimating the collection of properties, $\{\psi_i\}$, or their distribution, but cannot measure every component of ψ_i with high precision. Instead, for each object, we have observed data, D_i, that provide information about ψ_i. In the following, we use bold symbols to refer to quantities collectively, e.g., $\boldsymbol{\psi} \equiv \{\psi_i\}$, and $\boldsymbol{D} \equiv \{D_i\}$.

We consider problems where the nature of the observations motivates models that specify a joint sampling distribution for the data, conditional on the member properties, that

factors into a product of conditionally independent *member sampling distributions*,[2]

$$p(\{D_i\} \,|\, \{\psi_i\}) = \prod_{i=1}^{N} p(D_i \,|\, \psi_i). \tag{2.1}$$

We model the member properties, $\boldsymbol{\psi}$, as IID draws from a *population probability density function* (PPDF), $f(\psi_i; \theta)$, with uncertain parameters, θ. Goals of inference may include estimation of the PPDF (i.e., estimation of θ), or estimation of the member properties, $\boldsymbol{\psi}$.

Fig. 2.1 shows the DAG for this type of model, both explicitly and using plate notation. Following standard conventions, open nodes indicate uncertain random variables that are targets of inference, and shaded nodes indicated observed quantities, i.e., random variables that are uncertain a priori, but that become known after observation, and thus may be conditioned on. If we denote the prior PDF for the population distribution parameters by $\pi(\theta)$, this DAG indicates that the joint PDF for all random quantities in this model may be written,

$$p(\theta, \{\psi_i\}, \{D_i\}) = \pi(\theta) \prod_{i=1}^{N} f(\psi_i; \theta)\, p(D_i \,|\, \psi_i) \tag{2.2}$$

$$\propto \pi(\theta) \prod_{i=1}^{N} f(\psi_i; \theta)\, \ell_i(\psi_i), \tag{2.3}$$

where we have defined the *member likelihood functions*,

$$\ell_i(\psi_i) \propto p(D_i|\psi_i).$$

Note that, as likelihood functions (vs. sampling distributions), these functions need only be specified up to proportionality. In particular, any dependence on D_i that does not influence the dependence on ψ_i can be ignored.

The DAG describes a generative model for all of the random variables, including the data. However, when the task is inference of parameters conditional on observed data (versus prediction of unobserved data), the use of member likelihood functions can be a significant simplification. In typical astronomical applications, estimation of ψ_i from D_i

[2]For the sake of simplicity, we denote random variables and their values with the same symbol. Also, $p(\bullet)$ will be used to denote the probability of an event or the probability density function, depending on the type of the argument.

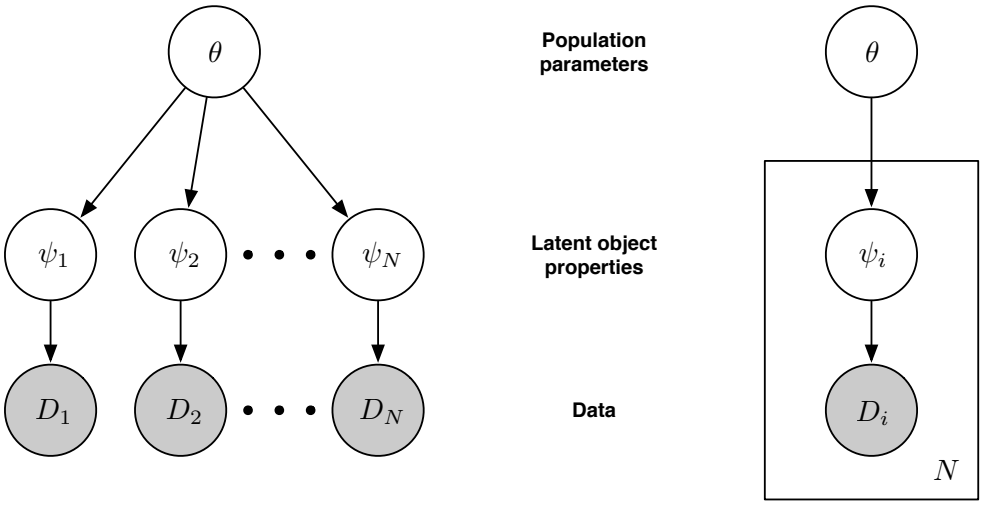


Figure 2.1: Directed acyclic graph (DAG) for a 2-level hierarchical Bayesian model. *Left*: DAG explicitly showing replicated conditionally independent subgraphs. *Right*: DAG depicting replicated elements with a plate.

is often a nontrivial inference problem in itself. For example, when ψ_i denotes the apparent brightness of a star and D_i denotes image data, inference may involve fitting the complicated point spread function of an imaging instrument to Poisson distributed photon counts in dozens or hundreds of pixels (often marginalizing over an uncertain background or instrument calibration component). The resulting likelihood function for ψ_i (or marginal likelihood function, when there are nuisance parameters) will often be relatively easy to summarize as a function of ψ_i; e.g., it may be well approximated by a Gaussian or multivariate Gaussian function (perhaps after a transformation). On the other hand, the sampling distribution for the data may be quite complicated. In many circumstances, it may not even be well-defined. Weather or spacecraft conditions may affect the precision, accuracy, and even the quantity of data for a member observation; the repeated sampling distribution may be hard or even impossible to define objectively, while the likelihood function may be well defined. Most astronomical surveys produce member estimates with heteroscedastic uncertainties, in the sense of producing member likelihood functions with widths that vary from object to object. It may be difficult or impossible to accurately describe the repeated sampling properties of the heteroscedastic uncertainties. But for inference based on *given*

observations, only the actually available member likelihood functions matter. Implementing inference in a manner that requires specifying only the member likelihood functions, rather than the sampling distributions, is a better fit to the nature of astronomical survey catalog data summaries than an implementation requiring unique specification of the lowest level sampling distributions.

A widely-used approach for posterior sampling in the context of two-level hierarchical models is the *Metropolis-within-Gibbs* (MWG) algorithm, where the θ population parameters and the ψ_i member properties are sampled in separate, alternating steps. First, the member properties are sampled by holding the population parameters fixed, then the population parameters are sampled by holding the member properties fixed. These steps may each be implemented with Metropolis or Metropolis-Hastings algorithms; their sequential combination amounts to Gibbs sampling on the joint space. Explicitly, the steps are:

$$\psi_i \sim p(\psi_i \,|\, \theta, D_i), \quad \forall i \in 1:N; \tag{2.4}$$

$$\theta \sim p(\theta \,|\, \boldsymbol{\psi}, M). \tag{2.5}$$

The departure point for CUDAHM is recognition that, since the ψ_i properties are conditionally independent in (2.4), they may be sampled in parallel, making this part of the MwG algorithm suitable for a massively parallel implementation using GPUs. We describe such an implementation further below.

Since ψ_i may be a vector, the ψ_i node in the DAG may admit a factorization leading to further structure within the plate in Fig. 2.1. Fig. 2.2 shows DAGs for several other single-plate modeling scenarios for which inference may be implemented using MWG with massively parallel sampling of member properties.

The DAG in the left panel depicts a frequently arising structure in astronomy, where the object properties ψ_i consist of intrinsic *characteristics* χ_i that, if known, can predict *observables*, $\mathcal{O}_i$, i.e., quantities that can predict observed data. An important example is inference of *number-size distributions* (also known as number counts or $\log N$–$\log S$ distributions). Here the object characteristics are distance, r_i, and luminosity, L_i (amount of energy emitted per unit time). The observable is flux (rate of energy flow per unit area normal to the line of sight, per unit time, at the telescope), F_i, related to the characteristics via the inverse-square law, $F_i = L_i/(4\pi r_i^2)$ (or its cosmological generalization).

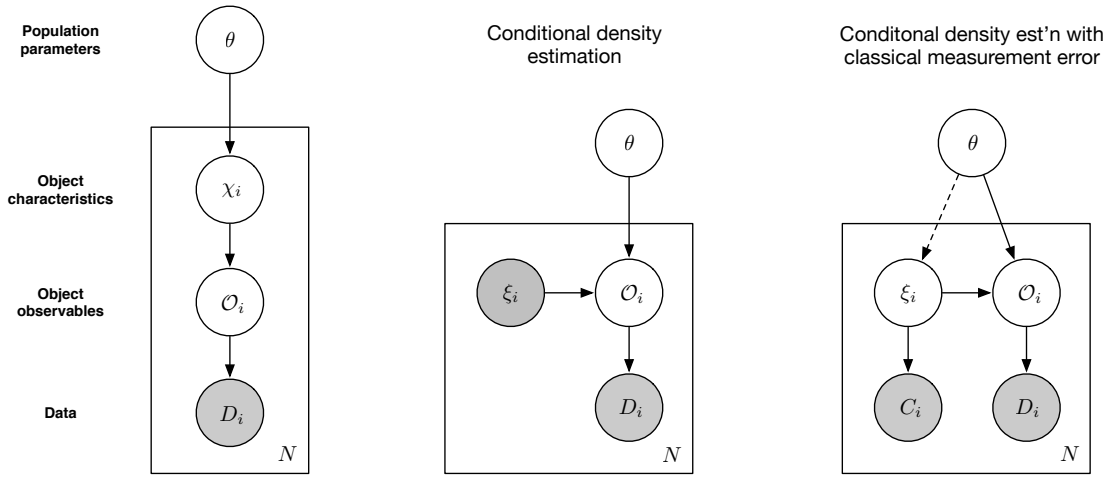


Figure 2.2: Example single-plate DAGs that may be implemented in CUDAHM. *Left*: DAG for a 3-level hierarchical model corresponding to demographic inference for objects with latent characteristics χ_i, related to latent observables $\mathcal{O}_i$. *Center*: DAG for conditional density estimation, expressed via a latent observable $\mathcal{O}_i$, and a precisely measured predictor (covariate), ξ_i. *Right*: DAG for conditional density estimation with classical measurement error, with a latent predictor, ξ_i, measured indirectly via data C_i. The predictor may have an a priori known prior distribution, or it may be parameterized (with parameters included in θ, in which case the dashed edge would be present).

The DAG in the middle panel depicts conditional density estimation, where the properties ψ_i are comprised of precisely measurable predictors (covariates), ξ_i, that, together with the population parameters θ, specify the PDF for observables, $\mathcal{O}_i$; the data provide likelihood functions for the $\mathcal{O}_i$. An important example is inference of a *luminosity function*, which describes the population distribution for the luminosities of a class of sources (say, a stellar or galaxy type). If the PDF for luminosity is $f(L;\theta)$, and the distances to objects may be precisely measured (say, via spectroscopic redshift data), then by a simple change of variables the PDF for the flux observable for a source at distance d is $4\pi d^2 f(4\pi d^2 F;\theta)$ (in Euclidean space). We treat a more complicated version of this problem below, where the object sample is subject to flux-dependent selection effects.

As a final example, the DAG in the right panel depicts conditional density estimation with measurement error (i.e., uncertainty in the predictors), with a classical measurement error structure (data distributions conditional on latent predictor values). A wide variety of astronomical data analysis problems have this structure. Kelly et al. (2012) describes a noteworthy example studying how the spectrum of infrared emission from heated interstellar dust depends on properties of the dust grains; this is one of the specific problems motivating CUDAHM. Earlier studies, based on maximum likelihood estimates of dust properties (ignoring measurement error), found a surprising negative correlation between dust temperature and a spectral index parameter indicating how the dust properties tilt the infrared spectrum away from a black body spectrum. Accounting for measurement error *reversed the sign* of the inferred correlation. Kelly et al. (2012) analyzed measurements from $\sim 10^4$ dust regions; CUDAHM dramatically accelerates the calculations and makes such studies feasible with 10 to 100 times larger samples.

2.2 CUDAHM architecture

To sample according to Eq. 2.4, we use the robust adaptive Metropolis (RAM) algorithm devised by Vihola (2012). It works by adaptively refining a Metropolis algorithm proposal distribution during the sampling process until a target mean acceptance rate α_* is reached. CUDAHM currently uses a multivariate normal distribution as the proposal q, and sets the target mean acceptance probability to a default value of $\alpha_* = 0.4$. Adaptation involves

using new samples to adjust the proposal covariance matrix in a manner that decays with time along the Markov chain so as to guarantee correct asymptotic sampling. Specifically, adjustments enter with a decaying weight, $\eta_n = n^{-2/3}$, where n is the iteration number along the Markov chain.

Following Vihola (2012), let S_1 be the identity matrix and X_1 some point in the space to be sampled for which the target density $\pi(X_1) > 0$. Each RAM iteration cycles through the following steps:

1. Compute $Y_n = X_{n-1} + S_{n-1}U_n$, where $U_n \sim q$ is an independent random vector.

2. With probability $\alpha_n \equiv \min\{1, \pi(Y_n)/\pi(X_{n-1})\}$ the step is accepted, and $X_n = Y_n$; otherwise the step is rejected and $X_n = X_{n-1}$.

3. Compute the lower-diagonal matrix S_n with positive diagonal elements satisfying the equation

$$S_n S_n^T = S_{n-1}\left(I + \eta_n(\alpha_n - \alpha_*)\frac{U_n U_n^T}{\| U_n \|^2}\right)S_{n-1}^T \tag{2.6}$$

   where I is an identity matrix. The solution for S_n is unique as it is the Cholesky factor of the right hand side.

•••*[Provide some further details about the CUDAHM architecture, e.g., how users must structure their code, classes/structures available for passing information to/from the GPU, how random number generation is handled, etc..]*•••

3 Example: Dust properties in a star-forming region

•••*[Add material here for Brandon's dust SED example if desired; we should probably try to save this for the A&C version.]*•••

4 Example: Luminosity function estimation

As a simple but useful demonstration of CUDAHM, we here consider luminosity function estimation, a parametric conditional density estimation problem arising across many areas of astronomy.

• • •*[Introductory content here was moved from Janos's intro.]*• • •

Fundamental observables for localized astronomical sources include position (both direction on the celestial sphere, and distance, d, in some chosen coordinate system), and apparent brightness, quantified in terms of flux, F, the electromagnetic power incident on a surface normal to the line-of-sight. Astronomers use these observables to study demographic properties of many classes of sources, including stars and galaxies of various types, minor planets (such as asteroids), and explosive transients (gamma-ray bursts, supernovae). For concreteness, here we focus on observations of nearby galaxies, for which distance may be measured using spectroscopy to find the *redshift*, z, of spectral lines (i.e., the fractional shift in wavelength). Due to the cosmological expansion, for relatively nearby galaxies, the distance is proportional to redshift,

$$d = \frac{cz}{H_0},$$

where c denotes the speed of light, and H_0 is Hubble's constant, describing the current expansion rate of the universe, with $H_0 \approx 70$ km s^{-1} Mpc^{-1} (with Mpc denoting megaparsecs). H_0 is measured with a precision of several percent, and spectroscopic redshifts for nearby galaxies can be measured to sub-percent precision. For simplicity, here we consider distances to be precisely measured, via redshifts. Often, astronomers use redshift as a proxy for distance.

A fundamental intrinsic characteristic of a source is its *luminosity* (emitted power), L, a measure of its intrinsic (vs. apparent) brightness. For nearby galaxies (i.e., for distances where space is very nearly Euclidean), the inverse-square law relates L to the observables F and d:

$$F = \frac{L}{4\pi r^2}.$$

The *luminosity function*, $\phi(L, d)$, describes the distribution of intrinsic brightness for a population at a specified distance (or redshift). It may be defined as the intensity function for a point process, i.e., as specifying the expected number of galaxies per unit volume at distance d, per unit luminosity interval. If we denote the spatial number density of galaxies at distance d by $n(d)$, then $n(d) = \int dL\, \phi(L, d)$. The *luminosity distribution* for galaxies at distance d (a PDF) is then

$$f(L, d) = \frac{\phi(L, d)}{n(d)}.$$

Note that $\phi(L, d)$ and $f(L, d)$ specify *conditional* distributions, i.e., distributions for L at a given d.[3]

The galaxy luminosity function carries valuable information about the formation and evolution of galaxies, therefore it is an important target of inquiry in astronomy. CUDAHM can address parametric luminosity function inference; we denote the luminosity function parameters by θ. For simplicity we here focus on a homogeneous population, so that the distance dependence may be ignored, but it is important in many applications. The following development is straightforward to generalize to account for distance dependence.

4.1 Parametric luminosity function models

For most cosmic populations, including galaxies, the luminosity function falls very steeply with increasing luminosity. The canonical starting point for parametric modeling of luminosities is the *Schecter function*,

$$\phi(L; \theta) = \frac{A}{L_*} \left(\frac{L}{L_*} \right)^{\beta} e^{-L/L_*},$$

where the parameters $\theta = (L_*, \beta, A)$ comprise a luminosity scale, L_*, a power law index, β, and an amplitude, A.[4] The form of the Schecter function would seem to imply a luminosity distribution that is a gamma distribution (with shape parameter $\alpha = \beta - 1$). However, the observed samples of many populations follow Eq. 4.1 with β in the interval $(-2, -1)$, in which case the integral of Eq. 4.1 over L is infinite, and the luminosity distribution is formally improper (with α outside of the allowed range for the gamma distribution). Low-luminosity sources are unobservable (due to noise and background, discussed below), so in practice the *observable* luminosity function is truncated at low luminosities, and the impropriety is often ignored. But the actual luminosity function must fall less steeply or turn over or be cut off at low luminosities.

•••*[Perhaps move some of the following details to an appendix?]*•••

[3]Authors vary on the definition of the luminosity function, some defining it as done here, and others using "luminosity function" to denote what we here call the luminosity distribution.

[4]There are varying conventions for parameterizing the amplitude of the Schecter function. In this parameterization, A has units of space density. In similar parameterizations, A is often denoted ϕ_*, although it neither has the units of ϕ, nor is it equal to $\phi(L_*)$, as the symbol might misleadingly suggest.

For some populations, an increase in the power law index is in fact observed at low luminosities. For example, the stellar initial mass function (related to the stellar luminosity function, and fit with similar models) has a low-mass (low-luminosity) index that increases by ≈ 1 (REF). Motivated by such observations, and to keep the luminosity distribution proper, we here adopt a "break-by-one" (BB1) generalization of the Schecter function, with $\phi \propto L^{\beta+1}$ at low luminosities, and thus integrable for $\beta > -2$. Specifically, the BB1 model has a luminosity distribution with three parameters: a mid-luminosity power law index, β, and two parameters defining the mid-luminosity range, (l, u), with $l < u$ and u playing the role of L_* in the Schecter function, and the power law index smoothly breaking to $\beta + 1$ as L decreases below l. The BB1 luminosity PDF has following functional form:

$$f(L;\theta) = \frac{C(\beta,u,l)}{u}\left(1 - e^{-L/l}\right)\left(\frac{L}{u}\right)^{\beta} e^{-L/u}, \tag{4.1}$$

where the normalization constant $C(\theta)$ is

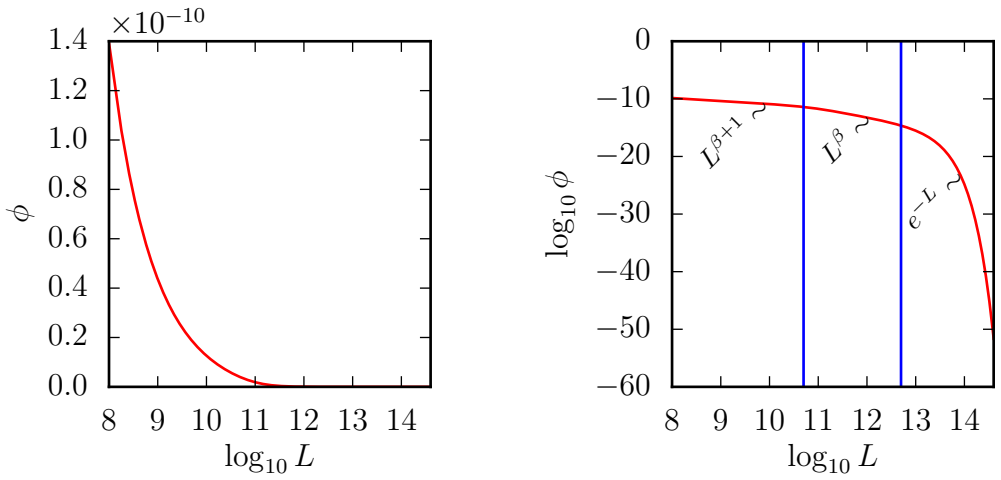
$$C(\beta,u,l) = \begin{cases} \dfrac{1}{\Gamma(\beta+1)\cdot\left(1 - \frac{1}{\left(1+\frac{u}{l}\right)^{\beta+1}}\right)} & \text{if } \beta > -2 \text{ and } \beta \neq -1; \\ \dfrac{1}{\log\left(1+\frac{u}{l}\right)} & \text{if } \beta = -1. \end{cases} \tag{4.2}$$

Note that as $l \to 0$, the BB1 distribution becomes a gamma distribution (if $\beta > -1$). We designed the BB1 distribution to have smooth power law break behavior, yet also have an analytical normalization constant; it is proper for $\beta > -2$. It can also be sampled from using a modification of a widely-used algorithm used to sample from the gamma distribution (REF). These properties make it useful for simulation experiments.

•••*[Specify param values for the fig in the next paragraph, and relabel the cutoff region in the fig with* $e^{L/u}$*.]*•••

We define a BB1 luminosity function by multiplying the BB1 luminosity distribution by the galaxy spatial number density, $n(d)$, which is simply a constant, n, for a homogeneous population. Fig. 4.1 shows an example BB1 luminosity function, with $\beta =?$, and $(l, u) =?, ?$ (with arbitrary units); it is plotted both with linear axes, and in log-log space, where the varying power law behavior is evident. The local power law index corresponds to the slope, $G(L)$, in log-log space, defined by

$$G(L) \equiv \frac{\mathrm{d}\log f}{\mathrm{d}\log L} = \frac{L}{f}\frac{\mathrm{d}f}{\mathrm{d}L} = g(L) + \beta - \frac{L}{u}, \tag{4.3}$$



(a) (b)

Figure 4.1: Typical shape of the probability distribution of galaxy luminosities on a log-linear scale (Panel a) and on log-log scale (Panel b). On Panel b the blue verticals show the lower limit l and upper limit u as defined in Eq. 4.1. The scaling of the "break-by-one" power law is different in the three regions defined by the limits. Parameters or the distribution are listed in Tab. 5.1.

with

$$g(L) = \frac{(L/l)e^{-L/l}}{(1 - e^{-L/l})}. \tag{4.4}$$

Evidently, $g(L) \to 0$ for $L \gg l$ and $g(L) \to 1$ for $L \ll l$. Thus the logarithmic slope, $G(L)$, corresponds to an exponential cutoff at large L, and at small L, a slope of $\beta + 1$. When $u \gg l$, so there is a range where $L \gg l$ but $L \ll u$, the logarithmic slope is $\approx \beta$ in that range.

Finally, the BB1 cumulative luminosity distribution function is

$$F(L;\theta) = C(\theta)\left[\Gamma(\beta+1) - \gamma\left(\beta+1, L/u\right) - \frac{\Gamma(\beta+1) - \gamma\left(\beta+1, L\cdot\left(\frac{1}{u}+\frac{1}{l}\right)\right)}{\left(1+\frac{u}{l}\right)^{\beta+1}}\right], \tag{4.5}$$

where $\Gamma(\cdot)$ and $\gamma(\cdot,\cdot)$ represent the gamma function and the upper incomplete gamma function, respectively.

4.2 Selection effects and measurement error

Luminosity functions are estimated using galaxy samples containing flux and distance measurements. Flux measurements are affected by photon noise that is often dominated by Poisson fluctuations in the photon counting rate. The measurement error thus approximately scales with the square root of the flux, and is fractionally greater at low flux than at high flux. At low fluxes, fluctuations from astrophyisical and instrumental backgrounds can produce false source detections. To prevent this, surveys adopt detection criteria to strongly mitigate against false detections. A simple, representative criterion is to accept sources only if the estimated flux is ν times the flux uncertainty, with $\nu \approx 5$ so that the probability for false detection is low even for large catalogs. Detection criteria introduce *selection effects* into catalogs. Most obviously, faint sources (low luminosity sources, or distant high luminosity sources) are excluded; the observable luminosity function is a thinned version of the actual luminosity function. More subtly, measurement error distorts the shape of the observable distribution, a phenomenon well known in the density deconvolution literature, and also recognized in the astronomical literature, where it is sometimes called *Eddington bias*, in reference to early discussions of the distortion by Eddington and Jeffreys (REFS). They noted that an object with a measured flux of $\hat{F}$ is more likely to be an object with a true flux $F < \hat{F}$ than one with $F > \hat{F}$, because the former are more numerous than the latter in most astronomical settings. Selection effects can exacerbate the distortion in the vicinity of a flux threshold, with measurement error and the falling flux distribution conspiring to scatter more below-threshold sources into the observed sample than above-threshold sources out of it, a phenomenon dubbed *Malmquist bias* (Binney & Merrifield 1998) (although the term is used inconsistently). Hierarchical modeling can automatically account for such thinning and distortion, in a manner that adapts to the shape of the luminosity function; this is a major motivation for its increasing popularity in astrostatistics.

4.3 Hierarchical model for luminosity function inference

4.4 Simulation setup

In case of real measurements the distance of the objects is known from redshift observations. In our simplified case we assume these distance measurements to be without error. To compile the simulated data set, we assume a spherically symmetric, homogeneous distribution of galaxies and generate the random distances r_i accordingly. We use the inverse-square law

$$F_i = \frac{L_i}{4\pi r_i^2}. \tag{4.6}$$

The observational noise E of the flux F is modelled as Gaussian with zero mean and a standard deviation of

$$\sigma(F) = \sqrt{\sigma_0^2 + (0.01F)^2}. \tag{4.7}$$

The flux limit of the telescope is the constant T and C will denote the event when an object is detected, i.e. the noisy flux is above the threshold: $D > T$. When generating the random sample, the luminosity is limited between $10 \leq \log_{10} L \leq 14$ which implies the distance limit

$$r_{\max} = \sqrt{\frac{L_{\max}}{4\pi T}}. \tag{4.8}$$

As it was discussed above in Sec. 1, the spatial distribution of galaxies is considered homogeneous, hence the density function becomes

$$\delta(r) = \begin{cases} \dfrac{3r^2}{r_{\max}^3} & \text{if } 0 \leq r \leq r_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

and the cumulative distribution function:

$$\Delta(r) = \begin{cases} 0 & \text{if } r < 0 \\ \dfrac{r^3}{r_{\max}^3} & \text{if } 0 \leq r \leq r_{\max} \\ 1 & \text{if } r \geq r_{\max} \end{cases} \tag{4.10}$$

4.5 Choice of prior

Now we consider the choice of prior for the population parameters. For β, we adopt a prior that is flat with respect to the angle in log-log space. This choice has the virtue of not

putting a lot of prior probability on the steep slope range, which a flat prior on β would do. Denoting the angle by φ, if we adopt a prior PDF of $h(\varphi)$ on φ, the prior on $\beta = \tan\varphi$ is

$$p(\beta) = \frac{h(\varphi)}{1+\beta^2}. \tag{4.11}$$

For a flat φ prior between two cut-offs φ_L and φ_U,

$$p(\beta) = \frac{1}{\varphi_U - \varphi_L} \cdot \frac{1}{1+\beta^2}. \tag{4.12}$$

This is a truncated Cauchy distribution. The Break-By-1 distribution requires $\beta > -2$, corresponding to $\varphi_L = -1.107$. If we require Eq. 4.1 to be decreasing, the upper limit becomes $\beta < 0$, corresponding to $\varphi_U = 0$. Thus, the prior on β is

$$p(\beta) = \frac{0.903}{1+\beta^2} \qquad \text{for} -2 < \beta < 0. \tag{4.13}$$

For the upper scale we use a log-flat prior, a conventional choice for a scale parameter that must be positive, even though this will be improper on both sides, we can ignore the impropriety and the normalizing constant since the likelihood function will make the posterior proper. A prior flat in $\log u$ corresponds to $p(u) \propto \frac{1}{u}$. The lower scale l must be below the upper scale u, which we can ensure by using a prior factored as $p(l,u) = p(u)p(l\,|\,u)$ and taking $p(l\,|\,u)$ to vanish for $l \geq u$. A log-flat prior also seems appealing for l but the data (i.e. luminosity measurements) do not probe the distribution down to zero due to the flux limit of the telescope, we could not have a proper prior without introduction a lower cut-off. Instead, we simply use a flat prior on l. Hence, the overall prior will be

$$p(\beta, l, u) \propto \frac{l}{u \cdot (1+\beta^2)} \qquad \text{for} -2 < \beta < 0, l < u. \tag{4.14}$$

4.6 Working with the Flux Limit

In order to implement the hierarchical Bayesian model outlined above, we have to calculate the probability distribution of functions of the characteristics and the population parameters, cf. Eq. **??** and 2.5. In our particular case the characteristics are the noiseless fluxes F_i (or real intrinsic luminosities L_i, as distances are known exactly). The population parameters are β, u and l. The measurements in our case are the noisy fluxes D_i and the C selection effect is the event when $D > T$, i.e. the observed flux surpasses the limit T of

the telescope. Modeling assumptions M involves selection effect, known distances and the standard deviation of observational noise. Eq. **??** becomes

$$p(F_i|M,\theta,D_i) = p(F_i|C,\sigma(F_i),r_i,\theta,D_i) \tag{4.15}$$

$$= \frac{p(C|D_i,F_i,\sigma(F_i),r_i,\theta)\cdot p(D_i|F_i,r_i,\sigma(F_i),\theta)\cdot p(F_i|r_i,\sigma(F_i),\theta)\cdot p(r_i,\sigma(F_i),\theta)}{p(C,\sigma(F_i),r_i,\theta,D_i)} \tag{4.16}$$

$$\propto \underbrace{p(C|D_i,F_i,\sigma(F_i),r_i,\theta)}_{=1}\cdot p(D_i|F_i,r_i,\sigma(F_i),\theta)\cdot p(F_i|r_i,\sigma(F_i),\theta) \tag{4.17}$$

$$= p(D_i|F_i,\sigma(F_i))\cdot p(F_i|r_i,\theta) \tag{4.18}$$

The term with the brace equals to 1 since all galaxies must surpass the flux limit in order to be in the sample. At the same time,

$$p(F_i|r_i,\theta) = \frac{d}{dF_i}\Phi(F_i 4\pi r_i^2;\theta) = \phi(F_i 4\pi r_i^2;\theta)\cdot 4\pi r_i^2. \tag{4.19}$$

To take the selection effect of the flux limit into account, we have to calculate the probability of measuring a galaxy with a given luminosity and distance above the flux limit. By assuming Gaussian noise, as we already did in Sec. **??**, the sought probability becomes

$$p(C|L,r) = p(C|F) = \Pr(D > T|F) \tag{4.20}$$

$$= \Pr(F + E > T) \tag{4.21}$$

$$= \Pr(\underbrace{F+E}_{\sim\mathcal{N}(F,\sigma(F))} > T) \tag{4.22}$$

$$= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{F-T}{\sqrt{2}\sigma(F)}\right)\right) \tag{4.23}$$

$$=: \zeta(F;T,\sigma_0) \tag{4.24}$$

We now calculate the probability of *any* galaxy being observed above the flux limit. Here we make the assumptions that L, r and the population parameters θ are independent.

From Eq. 4.24, if follows that

$$\mu(\theta) := p(C|\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(C, L, r|\theta)\,\mathrm{d}L\,\mathrm{d}r \tag{4.25}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{p(C, L, r, \theta)}{p(\theta)}\,\mathrm{d}L\,\mathrm{d}r \tag{4.26}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{p(C|L, r, \theta)\cdot p(L|r, \theta)\cdot p(r|\theta)\cdot p(\theta)}{p(\theta)}\,\mathrm{d}L\,\mathrm{d}r \tag{4.27}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(C|L, r)\cdot p(L|\theta)\cdot p(r)\,\mathrm{d}L\,\mathrm{d}r \tag{4.28}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \zeta\left(\frac{L}{4\pi r^2}, T, \sigma_0\right)\cdot \phi(L|\theta)\cdot \delta(r)\,\mathrm{d}L\,\mathrm{d}r. \tag{4.29}$$

The probability of a galaxy with true luminosity at a given distance surpassing the flux limit can be calculated as

$$\phi^T(L; r, \theta) := p(L|C, r, \theta) = \frac{p(L, C, r, \theta)}{p(C, r, \theta)} \tag{4.30}$$

$$= \frac{p(C|L, r, \theta)\cdot p(L|r, \theta)\cdot p(r, \theta)}{p(C|r, \theta)\cdot p(r, \theta)} \tag{4.31}$$

$$= \frac{p(C|L, r)\cdot p(L|\theta)}{p(C|\theta)} \tag{4.32}$$

$$= \frac{\zeta\left(\frac{L}{4\pi r^2}; T, \sigma_0\right)\cdot \phi(L|\theta)}{\mu(\theta)}. \tag{4.33}$$

Using Eq. 4.19 and (the deduction of) Eq. 4.33, we obtain

$$p(F_i|C, r_i, \theta) = \frac{\zeta(F_i; T, \sigma_0)\cdot \phi(4\pi r_i^2 F_i; \theta)\cdot 4\pi r_i^2}{\mu(\theta)} = \phi^T(F_i; r_i, \theta)\cdot 4\pi r_i^2. \tag{4.34}$$

Eq. 4.34 is the flux density function with the selection effect taken into account. In Fig. 4.2 we plot the logarithm of $\phi(L; \theta)$ and $\phi^T(L; r, \theta)$ to illustrate the difference introduced by the selection effect.

Until now, F_i was used to denote the flux of objects whether they passed the selection limit or not, i.e. for random variables sampled from Eq. 4.19. To distinguish F_i from random variables sampled from Eq. 4.34, we introduce the set of $\boldsymbol{F'} = \{F_i'\}$, where $F_i' \sim \phi^T(F_i'; r_i, \theta)\cdot 4\pi r_i^2$. Just like F_i, F_i' are also iid. The condition in $p(\theta \mid \boldsymbol{F'}) = p(\theta|M, \boldsymbol{F})$, therefore, means that the given sample is flux limited. We now show that the probability distribution of the population parameters, given the characteristics of *all* objects and the

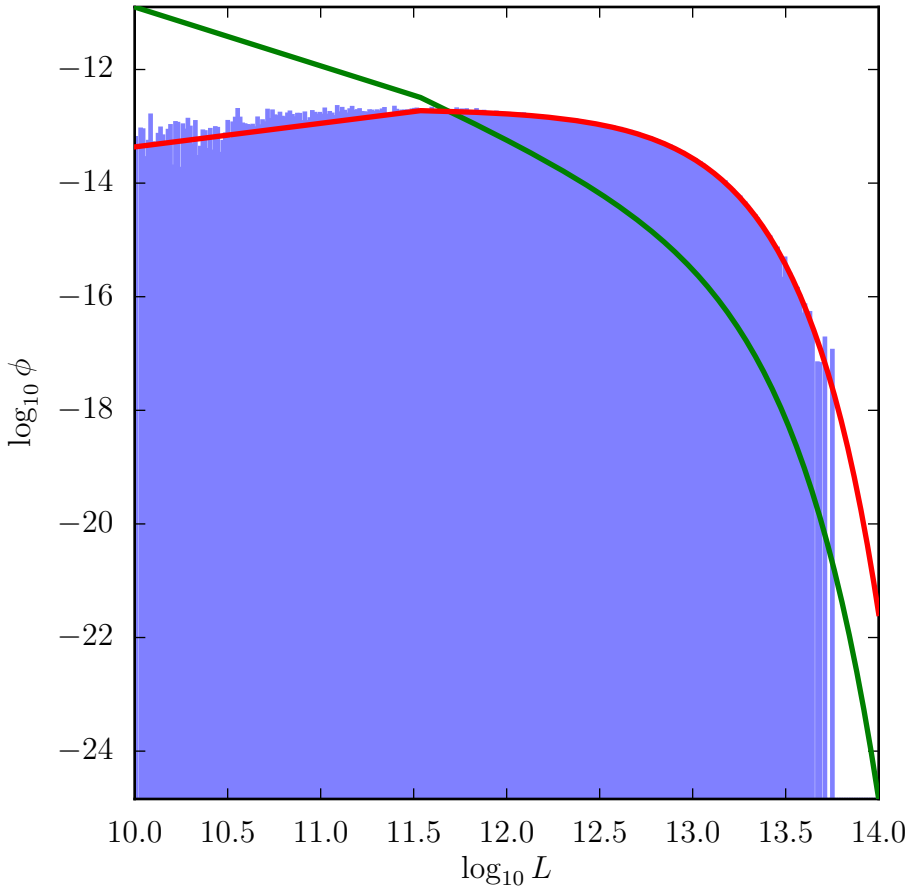


Figure 4.2: The luminosity function $\phi(L;\theta)$ (green curve) and the luminosity distribution of galaxies affected by the Malmquist bias $\phi^T(L;r,\theta)$ (red curve). We generated 100,000 galaxies with noisy flux (blue histogram). In case of the random sample, the limiting flux was set to $T = 5.0$ which gave a corresponding distance limit of $r_{\max} = 1.12 \times 10^6$.

selection criteria, is proportional to the following product of independent probabilities:

$$
\begin{aligned}
p(\theta|M, \boldsymbol{F}) &= p(\theta|\boldsymbol{F'}) && (4.35)\\
&= \frac{p(\theta, \boldsymbol{F'})}{p(\boldsymbol{F'})} && (4.36)\\
&= p(\theta) \cdot \prod_{i=1}^{N} \frac{p(F_i'|\theta)}{p(F_i')} && (4.37)\\
&\propto p(\theta) \cdot \prod_{i=1}^{N} p(F_i'|\theta) && (4.38)\\
&= p(\theta) \cdot \prod_{i=1}^{N} \phi^T(F_i'; r_i, \theta) \cdot 4\pi r_i^2 && (4.39)\\
&= p(\theta) \cdot \prod_{i=1}^{N} \frac{\zeta(F_i'; T, \sigma_0) \cdot \phi(4\pi r_i^2 F_i'; \theta) \cdot 4\pi r_i^2}{\mu(\theta)} && (4.40)\\
&\propto p(\theta) \cdot \prod_{i=1}^{N} \frac{\phi(4\pi r_i^2 F_i; \theta)}{\mu(\theta)} && (4.41)
\end{aligned}
$$

where $p(\theta)$ is the prior for the population parameters (see Eq. 4.14).

4.7 Generating Random Luminosities, Distances and Fluxes

To generate a random sample of galaxies with noisy fluxes, we perform the following steps:

1. Sample the luminosity function with a fixed set of population parameters to get the real luminosities L_i
2. Sample the distance distribution to get the distances r_i
3. Calculate the real fluxes F_i from L_i and r_i
4. Generate the noisy fluxes D_i from F_i.

Fig. 4.3 illustrates the dependency graph of the parameters and random variables.

The luminosity function Eq. 4.1 is sampled using the rejection sampling method at fixed population parameters to generate the real luminosities L_i. Then we apply the inverse transform method to sample the distances r_i from Eq. 4.10. The real flux F_i is calculated from L_i and r_i and a Gaussian random noise is added. At last, the selection effect is taken

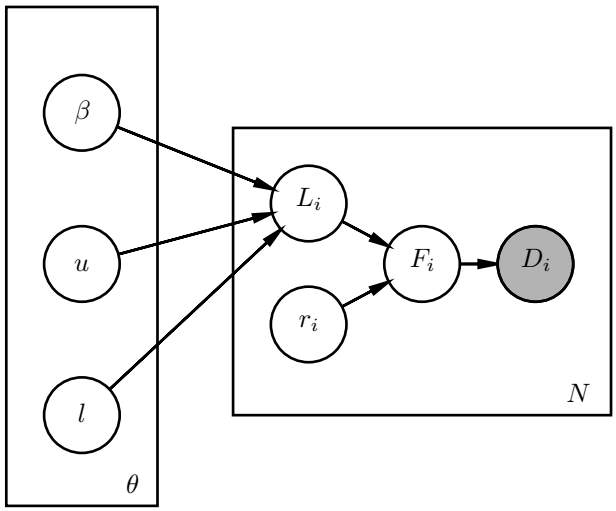


Figure 4.3: The hierarchical model to generate random samples of galaxies. Luminosity depends on the population parameters $\theta = (\beta, l, u)$. The real flux is computed directly from the real luminosity and the distance. Observational statistical error only affects the flux measurement and the noise depends on the real flux only. The selection effect $p(D_i < T) = 0$ is not indicated on the graph.

into account by rejecting samples which would not pass the flux limit. To visualize the Malmquist bias, we plot the real luminosities as a function of distance in Fig. 4.4. The blue curve represents the flux limit T expressed in luminosity units. Red dots appearing *below* the luminosity limit are galaxies which got into sample because the selection function was applied to the noisy fluxes.

5 Application to Simulated Data

We apply the hierarchical Bayesian model to estimate the model parameters of a simulated random sample of 100,000 galaxies. We compare the results of the Bayesian analysis to maximum likelihood and evaluate the performance of the GPU-based implementation. The true values of the population parameters are listed in the first row of Tab. 5.1. The value of σ_0 in Eq. 4.7 is chosen to be $\sigma_0 = 1$. The flux limit is $T = 5.0$ and the distance limit is $r_{\max} \gtrsim 1.12 \times 10^6$.

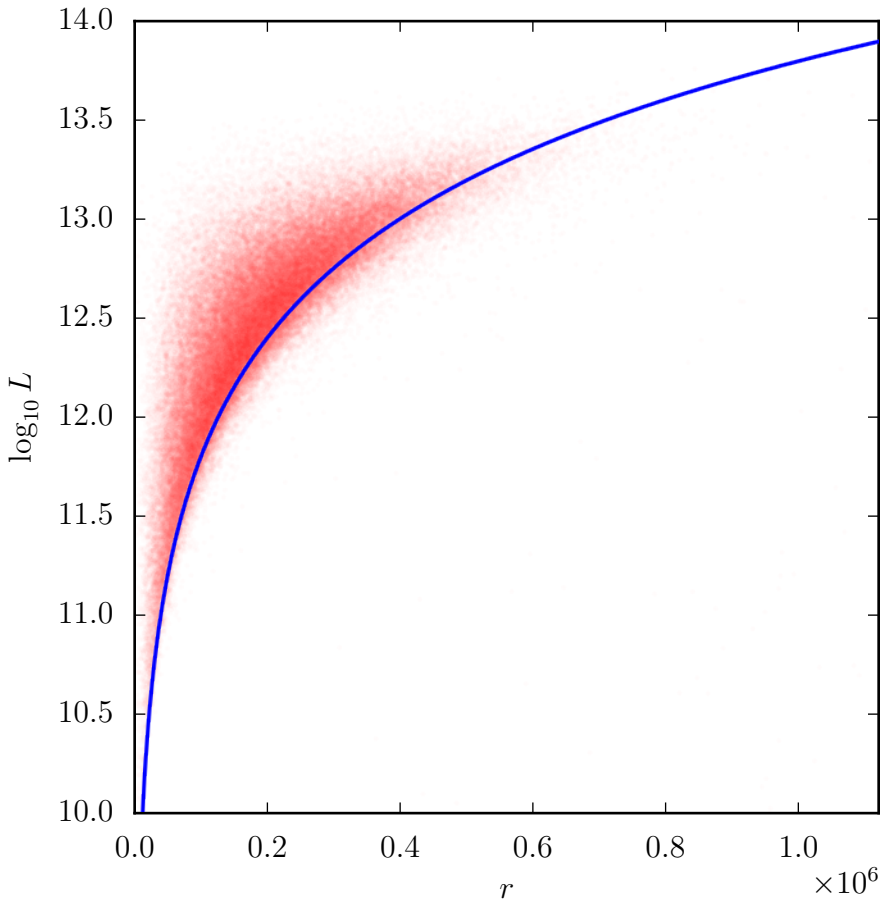


Figure 4.4: A random sample of luminosities generated using our model, plotted as a function of distance. Red dots mark individual object, whereas the blue curve represent the luminosity equivalent of the flux limit $T = 5.0$. Since the flux limit is applied to the noisy fluxes, a significant fraction of the objects is scattered below the luminosity limit. Any parameter estimation method has to account for this effect in order to successfully approximate the population parameters.

	β	l	u
θ_{true}	-1.5	5.0×10^{10}	5.0×10^{12}
θ_{MCMC}	-1.5037	5.0302×10^{10}	5.0000×10^{12}
$\sigma_{\theta,\text{MCMC}}$	0.0059	3.3548×10^{9}	3.1806×10^{10}
θ_{MLE}	-1.5564	7.3222×10^{10}	5.7207×10^{12}
$\theta_{\text{MLE, no noise}}$	-1.5009	4.9341×10^{10}	4.9819×10^{12}
$\lvert\theta_{\text{true}} - \theta_{\text{MLE}}\rvert$	$> 9.5525 \cdot \sigma_{\beta,\text{MCMC}}$	$> 6.9221 \cdot \sigma_{l,\text{MCMC}}$	$> 22.6591 \cdot \sigma_{u,\text{MCMC}}$

Table 5.1: Summary of parameter estimation results. The first row of the table indicates the true values which the simulated data was generated with. The second row shows the mean of the distributions coming of the Bayesian model, whereas the third row contains the standard deviation of them. For reference, we indicate the outcome of the ML estimatior run on the simulated data with and without noise in rows 4 and 5, respectively. To compare the Bayesian model to ML, the last row of the table shows the difference between the ML estimator and the true values in terms of the standard deviation of the posterior from the Bayesian model.

We executed 1.5M burn-in and the same number of live MCMC steps to sample the probability distribution of the population parameters. The length of the burn-in sequence was chosen by visual inspection of the autocorrelation plots. To reduce autocorrelations in the Markov chain, θ samples were thinned and only every 150^{th} sample was kept, hence the final number of samples was 10,000. Fig. 5.1 shows the traces, histograms and autocorrelation plots of the population parameters whereas Tab. 5.1 lists the result in a numerical format.

5.1 Performance tests

We used NVIDIA Tesla K40c cards for the performance tests. First, we executed tests without imposing a flux limit, which is a much simpler case as it does not contain the time-consuming numerical integration of Eq. 4.29. Next, we executed the with the flux limit turned on. Fig. 5.2 shows the elapsed time as a function of iteration number (non-

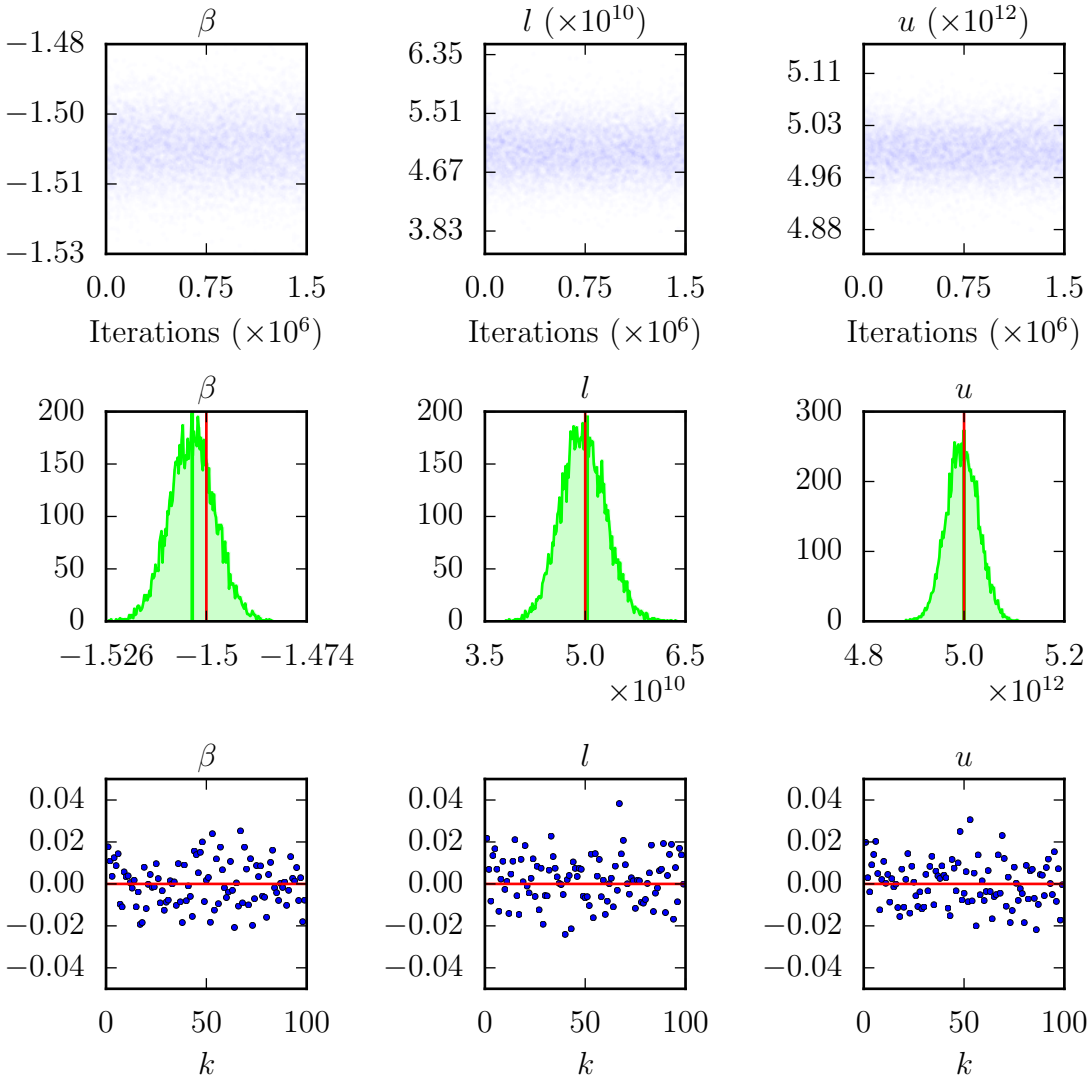


Figure 5.1: Trace (upper row), histogram (middle row) and autocorrelation (lower row) plots of the three population parameters. The red vertical line overplotted the histograms show the true value of the parameter. With the exception of β, Bayesian modelling can recover model parameters with excellent accuracy.

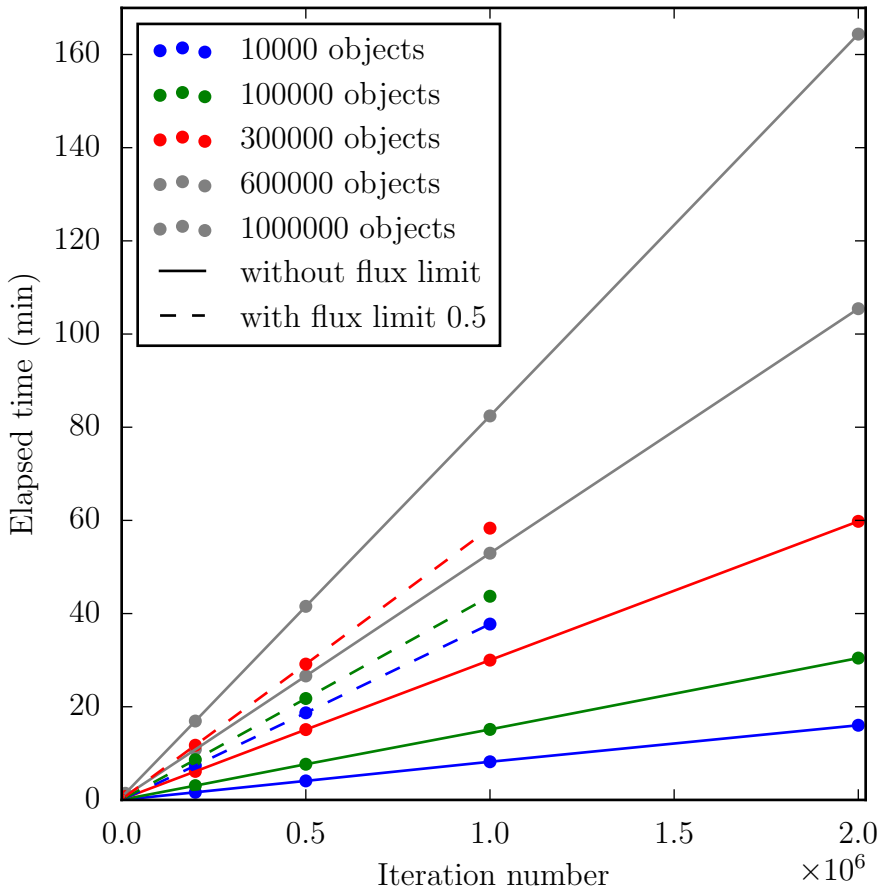


Figure 5.2: Runtime of the GPU code as a function of iterations for different numbers of objects, with (solid lines) and without (dashed lines) considering a flux limit of $T = 0.5$. Taking the flux limit into account results in an approximately fourfold increase of runtime.

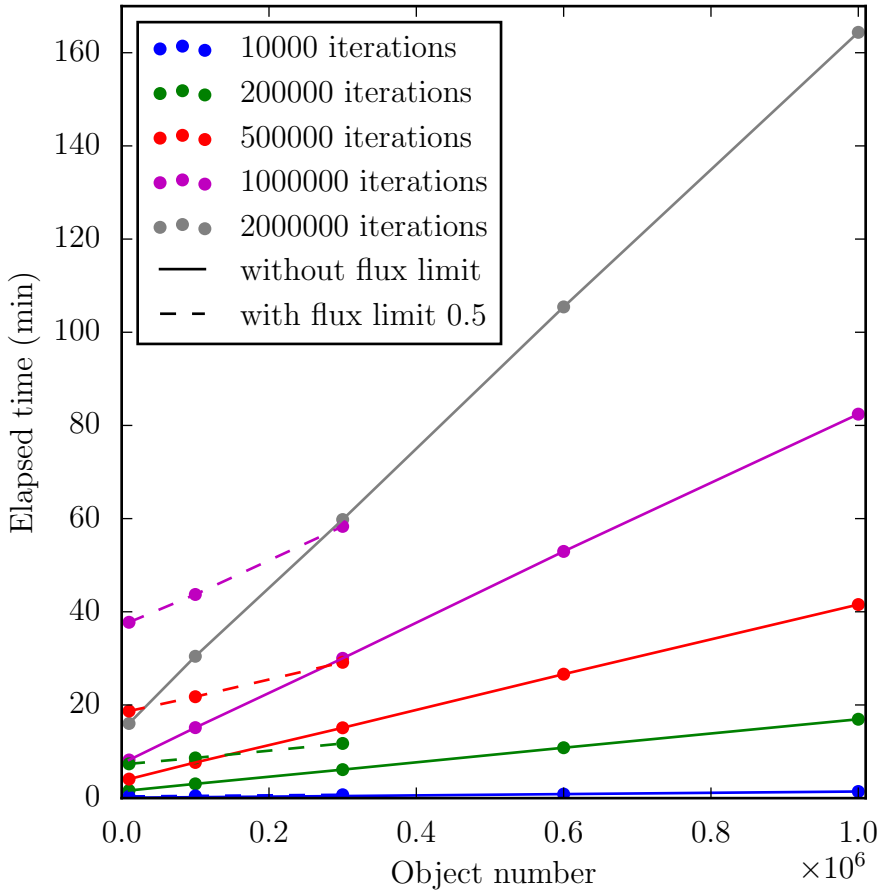


Figure 5.3: Runtime of the GPU code as a function of the number of objects, after a given number of iterations, with (solid lines) and without (dashed lines) considering a flux limit of $T = 0.5$. Taking the flux limit into account results in an approximately fourfold increase of runtime. On the other hand, the scaling of runtime with the number of objects is linear. This is due to the fact that fluxes are statistical independence, hence their distribution can be sampled in parallel on the GPU.

thinned) for different number of objects. The functions are trivially linear as iteration steps always take a fixed amount of time. More informative is Fig. 5.3 where we plot the elapsed time as a function of the number of objects for various numbers of iteration steps. The linear scaling of computation time with the number of objects is due to the fact that the characteristics are independent. As it is visible from the figures, turning on the flux limit clearly decreases the performance. Real galaxy catalogs contain objects on the order of 10^8, two magnitudes more than our simulated data set. Extrapolating from out performance numbers, estimating the parameters of the luminosity function with 2×10^5 Markov steps would take about 2000 minutes, a bit less then one and a half days, which makes applying our method to real data feasible.

6 Summary

We have presented a general hierarchical Bayesian method to estimate the probability distribution of population-level parameters and characteristics. We have applied this framework for the luminosity distribution function of galaxies. The method not only estimates the parameters of the luminosity function, but is also capable of recalibrating noisy observations. Since the luminosities of galaxies are independent, the problem is massively parallelizable on the GPU. Our implementation shows linear scaling with both, the number of Markov chain iterations and the number of objects, which makes it applicable to real data sets of size 10^8. We have made simplifications to the model that need to be mended when the method is used for real astronomical data, namely, correct cosmological distances have to be used. Less trivial extensions to the model are folding in the spatial correlation of galaxies (large scale structure of the universe) and the distance–luminosity correlations (evolution of galaxy brightness).

SUPPLEMENTARY MATERIAL

7 Computing $\mu(\theta)$

We want to compute the following integral numerically:

$$\mu(\theta) = \int_0^\infty \mathrm{d}r \int_0^\infty \mathrm{d}L \; \zeta\left(\frac{L}{4\pi r^2}, T, \sigma_0\right) \cdot \phi(L|\theta) \cdot \delta(r), \tag{7.1}$$

where

$$\zeta(F; T, \sigma_0) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{F-T}{\sqrt{2}\cdot\sqrt{\sigma_0^2 + (0.01F)^2}}\right)\right). \tag{7.2}$$

It turns out that direct numerical integration of Eq. 7.1 using the `cubature` package (Johnson n.d.) is unstable due to the very steep nature of the error function part. We plot the argument of the error function $A = \frac{F-T}{\sqrt{2\cdot(\sigma_0^2+(0.01F)^2)}}$ in Fig. 7.1 as a function of F for reference. The value of A approaches $\frac{1}{\sqrt{2}\cdot 0.01} \approx 70.710678$ in the $F \to \infty$ limit. For any value $c < \frac{1}{\sqrt{2}\cdot 0.01}$, the inverse error function yields $X = \mathrm{erf}^{-1}\, c = \frac{T+\sqrt{2}c\cdot\sqrt{(1-2(0.01)^2c^2)\sigma_0^2+(0.01)^2T^2}}{1-2(0.01)^2c^2}$. In luminosity units, X corresponds to $L = 4\pi r_{\mathrm{max}}^2 X$. We can approximate the error function by 1 if its argument is sufficiently large. For our purposes $c > 6$ is a good choice (the equivalent of a 6σ measurement error), for which $\mathrm{erf}(6.0) \approx 0.99999999999999784803$. Hence, we split the integral in Eq. 7.1 into two intervals: $F < X$ and $X \leq F$, corresponding to $L < 4\pi r^2 X$ and $L \leq 4\pi r^2 X$. To avoid the variable r in the limit of the integration, we will use the constant $4\pi r_{\mathrm{max}}^2 X$ as the threshold. Eq. 7.1 becomes

$$\begin{aligned}
\mu(\theta) &= \int_0^{r_{max}} \int_0^{\infty} \zeta\left(\frac{L}{4\pi r^2}; T, \sigma_0\right) \cdot \phi(L;\theta) \cdot \delta(r)\, \mathrm{d}L\, \mathrm{d}r && (7.3)\\
&\approx \int_0^{r_{max}} \int_0^{4\pi r_{\mathrm{max}}^2 X} \zeta\left(\frac{L}{4\pi r^2}; T, \sigma_0\right) \cdot \phi(L;\theta) \cdot \delta(r)\, \mathrm{d}L\, \mathrm{d}r && (7.4)\\
&\quad + \int_0^{r_{max}} \delta(r) \cdot \int_{4\pi r_{\mathrm{max}}^2 X}^{\infty} \phi(L;\theta)\, \mathrm{d}L\, \mathrm{d}r && (7.5)\\
&= \int_0^{r_{max}} \int_0^{4\pi r_{\mathrm{max}}^2 X} \zeta\left(\frac{L}{4\pi r^2}; T, \sigma_0\right) \cdot \phi(L;\theta) \cdot \delta(r)\, \mathrm{d}L\, \mathrm{d}r && (7.6)\\
&\quad + 1 - \Phi(4\pi r_{\mathrm{max}}^2 X; \theta), && (7.7)
\end{aligned}$$

where we made the approximation $\zeta(F; T, \sigma_0) \approx 1$ for $X \leq F$. The numerical integral of the first term now converges and we can integrate the second term by parts analytically. One can recognize Eq. 4.5, the cumulative probability function of luminosity in the result, which can be computed directly.

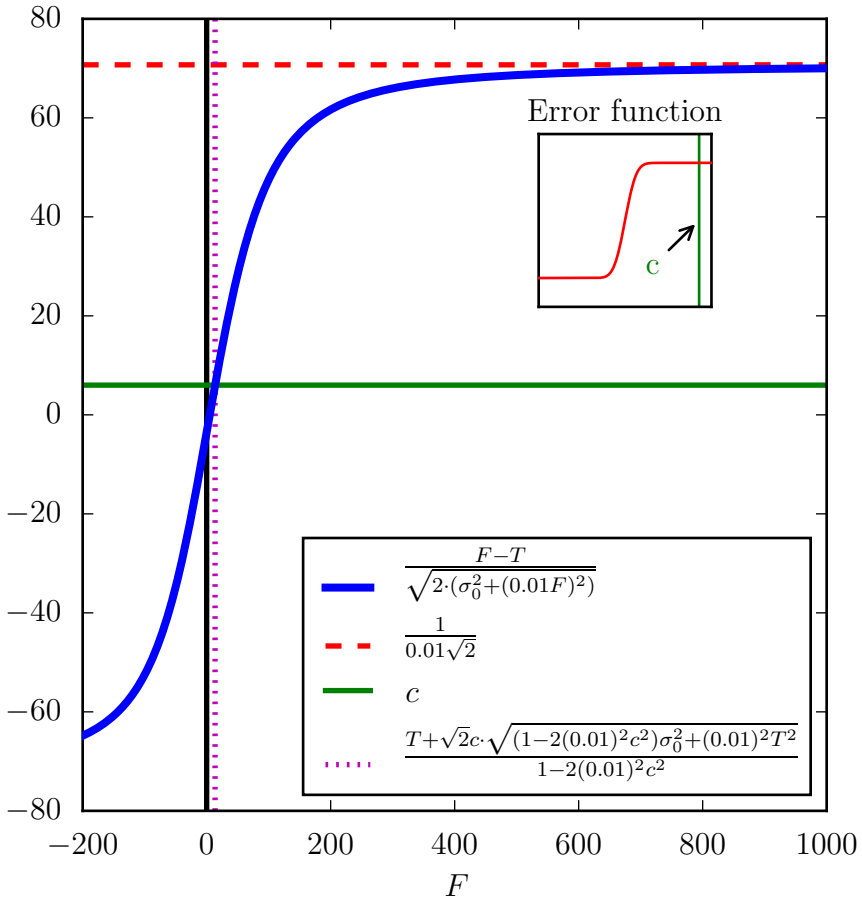


Figure 7.1: The argument of the error function in Eq. 7.2 (blue curve). See text for discussion.

References

Binney, J. & Merrifield, M. (1998), *Galactic Astronomy*.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006), *Measurement error in nonlinear models*, Vol. 105 of *Monographs on Statistics and Applied Probability*, second edn, Chapman & Hall/CRC, Boca Raton, FL. A modern perspective.
URL: *http://dx.doi.org/10.1201/9781420010138*

Johnson, S. G. (n.d.), 'Cubature: C package for adaptive multivariate integration of vector-valued integrands over hypercubes', http://ab-initio.mit.edu/wiki/index.php/Cubature.

Kelly, B. C., Shetty, R., Stutz, A. M., Kauffmann, J., Goodman, A. A. & Launhardt, R. (2012), 'Dust Spectral Energy Distributions in the Era of Herschel and Planck: A Hierarchical Bayesian-fitting Technique', *The Astrophysical Journal* **752**, 55.

Loredo, T. J. (2013), Bayesian Astrostatistics: A Backward Look to the Future, *in* J. M. Hilbe, ed., 'Astrostatistical Challenges for the New Astronomy', number 1 *in* 'Springer Series in Astrostatistics', Springer New York, pp. 15–40.

Vihola, M. (2012), 'Robust adaptive metropolis algorithm with coerced acceptance rate', *Statistics and Computing* **22**(5), 997–1008.